UNIVERSITÀ DELLA CALABRIA

Dipartimento di ELETTRONICA,
INFORMATICA E SISTEMISTICA

UNIVERSITÀ DELLA CALABRIA

Dipartimento di Elettronica,
Informatica e Sistemistica

Dottorato di Ricerca in
Ingegneria dei Sistemi e Informatica
XX ciclo

Tesi di Dottorato

Resource Reservation Protocol and Predictive Algorithms for QoS support in Wireless Environments

Peppino Fazio

UNIVERSITÀ DELLA CALABRIA

Dipartimento di Elettronica,
Informatica e Sistemistica

Dottorato di Ricerca in
Ingegneria dei Sistemi e Informatica
XX ciclo

Tesi di Dottorato

Resource Reservation Protocol and Predictive Algorithms for QoS support in Wireless Environments

Peppino Fazio

Coordinatore
Prof. Domenico Talia

Tutor
Prof. Salvatore Marano

To my father Tommaso

Amplissimis verbis gratias ago tibi,
omnium rerum quae mihi dedisti.
Beneficium reddere non potui,
spero me te dignum esse.

S.P.D. Peppino

Index

Chapter 1 - The wireless communication systems, wireless LANs and link modeling

Chapter 2 - The Quality of Service (QoS) in wireless environments and resources reservation

Chapter 3 - The rate adaptation in wireless networks

Chapter 4 - A new direction-based mobility prediction algorithm

Chapter 5 - Simulation tool and performance evaluation

Introduction

This **Philosophiæ Doctoral** (PhD) thesis faces some problems about the shared resources allocation in wireless environments, through a new algorithm that takes into account both users mobility and wireless channel degradation level; in addition, a prediction reservation policy for passive resources is also presented, in order to guarantee a certain Quality of Service (QoS) level to mobile services. An analytical analysis of the wireless channel model has also been carried out.

Wireless technologies represent a rapidly emerging area of growth and importance for providing ubiquitous access to the network for all of the community. Students, faculty and staff increasingly want un-tethered network access from general-purpose classrooms, meeting rooms, auditoriums, and even the hallways of campus buildings. There is interest in creating mobile computing labs utilizing laptop computers equipped with wireless cards. Recently, industry has made significant progress in resolving some constraints to the widespread adoption of wireless technologies. Some of the constraints have included disparate standards, low bandwidth, and high infrastructure and service cost. Wireless is being adopted for many new applications: to connect computers, to allow remote monitoring and data acquisition, to provide access control and security, and to provide a solution for environments where wires may not be the best solution. The most important peculiarity of wireless communications is represented by their capability of ensure the requested services to moving hosts in the coverage domain.

A telecommunications network can be considered as a structure that is able to establish, after a user service request, a connection for the communication with another user or a multicast group. The main functionalities of a network are: the choice of the appropriate level of bandwidth that must be assigned to the requesting user, the transmission over the channel the effective desired information, the administration of the resources in order to allow the correctness of the transmissions. Supplying an appropriate service by a network needs the cooperation of different entities of different network nodes, connected through physical communication means. In every kind of telecommunication, each entity coordinates itself with the other one by the *communication protocols*, generally based on the *layer* concept: the functionalities set is subdivided into a stack of overlapped layers, where each of them offers a subset of the total needed functionalities, in order to allow the communication with other systems. Among the all reference models, the ISO-OSI (International Standard Organization – Open System Interconnection) is the most famous, based on a 7-layers stack.

There are two different categories of LAN (Local Area Network): the wired one (wired-LAN) and the wireless one (wireless-LAN or WLAN) that uses radio waves as transmitting media. WLANs minimize the need of wired connections, combining the high grade of connectivity with the mobility of users, which can access to shared resources without the need of finding a physical point where the device must be connected; in addition, network administrators can expand system dimensions without install or moving cables. Other advantages of WLANs are the quick installation, high scalability and very bounded costs.

As all the telecommunication systems, there are also some disadvantages as the employment of radio waves for information transmissions (that leads to the presence of a high amount of electromagnetic phenomena, intrinsic in waves propagation), the physical obstacles (that are not perfectly penetrable by radio waves), the limited bandwidth availability and the frequency reuse.

For the Internet real-time traffic the ReSerVation Protocol (RSVP) is used as a resource reservation protocol; the main aim of the RSVP is the exchange, among the system nodes, of the *reservation state*, i.e. the reservation state of the resource request of the single connection. Users' has a heavy impact on the QoS parameters of the real-time applications. The existing architectures for the management of real-time services on a network with fixed hosts (no users' mobility) are not adequate for mobility support, in fact some additional functionalities are needed to the RSVP, in order to face such kind of problems. For example, the effects of users nobility can be reduced by "pre-reserving" the resources on the locations (cells) that a mobile host will visit during its active connection (as it will be explained, this is called "*passive reservation*" policy). When a mobile host moves among different coverage areas with an active session, the packet delivery delay may vary (the congestion level on a new path with a new set of involved routers can be different if compared with the previous path); if the new coverage node, where the mobile host has moved, is overloaded, the available bandwidth of the new location may be scarce or not able to satisfy the service request: in this case the mobile host must adapt its QoS request. In addition, during a hand-off event, the mobile host may have some temporary service degradations (in some cases the call may be also dropped). In order to avoid these problems, a certain admission control scheme must be executed in the routers and in the Access Points (APs, like the coverage nodes) that belong to the path from the source to the locations that the mobile host may visit during the whole connection duration. A flow is admitted into the network only if all the admission controls reply with a positive answer.

All the operations and funcionalities above are allowed if the Mobile RSVP (MRSVP) protocol is employed, that has been formulated as an extension of the RSVP for the hosts mobility support expecially in the (ISPNs – Integrated Services Packet Networks): two kinds of reservations are used, the *active* reservation (between the sender and the current host coverage cell) and the *passive* one (between the sender and all the future cells specified in the user mobility profile). In this way, the right amount of bandwidth can be "in-advance" reserved needed by the flow in order to avoid service degradations during hand-off events.

Another extension of the RSVP is the Dynamic RSVP (DRSVP), able to offer dynamic QoS in a network with variable bandwidth, where some wireless links are present (their BER performances vary during an active connection) with mobile intermediate nodes (network topology varies, modifying the quality of a generic path between two nodes); in this scenario, the DRSVP limits the assigned amount of bandwidth, by executing an acccurate analysis of the network *bottlenecks*: when a request is admitted into the system, reserving high amounts of bandwidth in some nodes has none-sense if it cannot be used because of the presence of some bottlenecks; if the "over-reservation" is avoided, then there will be a resource availability gain.

As earlier introduced, the advantages of the wireless networks (portability, high connectivity and low costs) are accompanied by some undesired propagation effects: in the WLAN environments the *path-loss* is the outstanding effect, expecially in the free-space communications, where the Line Of Sight (LOS) communication is affected by the distance between transmitter and receiver; in the urban or rural environments the presence of obstacles and buildings smooths the transmitted signal while creating the such called "shadow zones", that block the signal propagation. There are different propagation effects more evident in urban environments (evanescence, trasmission, refraction and shadowing), but the most important is the *reflection* of the electromagnetic beam, that generates additional paths over which the reflected signal propagates itself; this is called *multipath-fading*. Under the multipath-fading effects, the electromagnetic beam is affected by a delayed diffusion, that is to say the reflected signal components arrive to destination at different time instants, creating a shattered version of the original transmitted signal. While in the satellite or micro-wave point-to-point communications the multipath effects are negligible, in the mobile telephony environments many obstacles and object are present, so the destination station will receive many copies of the original signal, which arrives with different delays.

The *fading* term refers to the temporal variation of the received signal power, caused by changes in the transmission mean or in the different paths. In a static environment (low or null grade of node mobility) the fading is supplied by the variations in atmospheric conditions (an example can be the presence of rain); in a mobile scenario, where the receiving/transmitting antenna position changes during an active session, complex transmission phenomena can be observed: if the same digital signal is transmitted twice at two different and well separated time instants, the receiver will observe two different signals, although the wireless media is the same for the two transmissions; it is due to the continuous aleatory evolutions of the wireless channel physical characteristics (the wireless channel is said to have an aleatory and time-variant impulse response). This is verified with the transmission of a short-duration signal (ideally an impulse) through a wireless multipath channel: the received signal will be an impulses train, with different shapes.

So, the only way to describe a wireless link behaviour is represented by a statistical model, which can take into account the time-variant nature of the link; our research has been based on the Markov stochastic process, able to describe, with a certain grade of precision, the trend of the wireless channel during time, in terms of transmission errors and degradation). The most popular fading channels are: the *Rayleigh* one (heavily used when there are many indirect paths from the transmitter to the receiver, without a dominant component) and the *Rice* one (when a direct and dominant path is present in the communication). The Rice and Rayleigh models are often used in the outdoor and indoor environments respectively. In order to relate the stochastic process with the channel model the Signal-to-Noise Ratio (SNR) must be partitioned in a finite set of intervals, associating each one of them to the states of the Markov Chain (in this way the physical characteristics and the time-variant behaviour of the channel are taken into account). In particular, in this thesis the channel model study has been focused on the IEEE802.11 standards, under a *slow fading* condition (the wireless link evolution can be considered "slow" if compared with the packet transmission time).

So, from the discussion above, it is clear that in the wireless communications there is not the "ideality" concept when considering the link between transmitter and receiver and the wastage of bandwidth due to the inevitable presence of positive values of Bit Error Rate (BER) must be taken into account: as in the ISPNs, when a host specifies a minimum level of QoS in its service request, the network must ensure the required level and, if a wireless link is involved, a major amount of bandwidth must be assigned, so the received *effective bandwidth* will respect the desired minimum QoS threshold.

QoS guarantees are the main aims of the modern telecommunication systems: adequate Call Admission Control (CAC) and Rate-Adaptation schemes must be employed, because they ensures to the intermediate and/or coverage nodes that the overload effect probability will be null or below a fixed threshold, so the minimum QoS levels will be always respected. A CAC policy is generally designed for the management of the maximum number of users that can be admitted into the system but, in the same time, it must ensure high system utilization with low congestion levels. A new service request is admitted into the system only if it passes the CAC over the the whole set of involved cells.

The rate-adaptation policies dynamically change the delivered flow-rates, respecting the requested bounds (like the *packet-delay* and *delay-jitter*) and some principles, as the *fairness*, *minimum overhead* and *high system utilization*. In this work a cellular coverage area has been considered, where the geographic area is subdivided in a certain number of regions, called *cells*. Each cell is covered by an AP that must ensure to mobile hosts a global wireless access to the system. In these environments, bandwidth reallocations are necessary when a *hand-in* (a new flow is admitted in the current cell or it arrives from the previous coverage area) or *hand-out* (a flow departs from the current cell) event occurs; in the first case, if the cell is high utilized, a degradation operation is necessary, in order to reduce the bandwidth level assigned to the current users; in the second case, the released bandwidth is redistributed with an upgrade operation. In the considered applications each flow can work with different bandwidth levels and the upgrade/degrade operations consist on increasing or decreasing the assigned rate with constant step. Different algorithms have been proposed in the literature in order to face these problems: some of them are based on the QoS parameters statistics, like the Degradation Ratio (DR, the ratio between the period of time when the host receives a degraded service and the cell residence time) and the Upgrade/Degrade Frequency (UDF, the frequency of switch between full service and degraded service of an admitted call), derived from an accurate analysis of the system evolution through Markov chains; other algorithms try to maximize the system utilization by defining and fixing some *Thresholds*, offering different admission grades, based on the current congestion level and on the arriving calls priorities. Other types of reallocation algorithms are based on the dynamics of some parameters (utilization, number of admitted flows, etc.), taking into account the evolution of the system and the traffic behaviour: in this way the overestimation of bandwidth requests is avoided, as well as the low system utilization. In order to make dynamic evaluations of the system status the *time windows* are used (that is to say certain time periods when the system parameters are sampled and observed). In this work, a "utility-oriented" algorithm has been used in order to manage the different service requests, taking into account both wireless channel conditions and users satisfaction level (through the employment of utility functions): the employed scheme is called "utility-oriented" because it takes into account the level of satisfaction obtained for the received bandwidth; utility functions are able to describe the QoS users requests and how a user can be satisfied by the obtained service. The concept of utility function has been introduced, as an indicator of the user satisfaction level. If the perceived utility must be maximized, then there must be a way to describe how a user is satisfied when the received bandwidth level varies. Utility functions are useful to solve such kind of problems and it has been shown that there are many works in literature that describe the best trends of utility functions, appropriate for the specific application (tolerant, intolerant, etc.).

After a description of some important concepts of utility functions, a new bandwidth allocation protocol has been introduced, with the aim of having a new scheme that can be applicable in the ISPNs systems. Since different applications can be introduced in an ISPN system, the proposed idea takes care of considering the specific utility function, as well as the wireless channel modelling. In this way some important goals can be reached: fairness among users belonging to the same class; high system utilization and QoS guarantees. A complexity analysis has been given for the UB CAC and BAG schemes and the importance of a dynamic bandwidth allocation scheme has been demonstrated, through an addicted campaign of simulations. The obtained results have shown that the introduction of a dynamic scheme for bandwidth management increases system performance, in terms of utilization and number of admitted flows. In addition, it has been shown that the introduction of a channel model is mandatory if channel degradations must be taken into account when dimensioning a wireless system or while serving MDP requests.

In addition, the algorithm must ensure the following criteria:

a) *QoS requirements*: the *utility-outage* (the received utility falls below a lower bound) event must be managed in an adequate manner, ensuring that the outage probability does not exceed a fixed threshold;

b) *Fairness*: for each user in the system an index must be defined (for example, referring it to the received utility) and, in the long term, the algorithm must ensure that there will not be high difference between different indexes of different users;

c) *High system utilization*: the offered utility is the criterion to measure the average utilization of the available bandwidth.

After the definition of the general behaviour of a rate-adaptation algorithm and the relative CAC module, they must be integrated with the ISPNs and their predictive service classes; in particular, in this work the Mobility Independent Predictive (MIP - users that request this kind of services want to avoid mobility effects) and Mobility Dependent Predictive (MDP - users that request this kind of services may suffer the effects of mobility, like degradations or connection droppings). The MRSVP guarantees service continuity a MIP user by making passive reservations over the APs that will serve the user during its connection; this kind of management may cause a resource wastage, because of the amount of passive and unused bandwidth that will switch into active resource only when a MIP user makes a hand-in; the obtained low system utilization is more evident when the MDP flows cannot have the access to the passive bandwidth.

The undesired resource wastage due to the passive reservations of MIP users cannot be avoided but it can be efficiently reduced if the following consideration is taken into account: the passive reservations may not be made over all the cells in the system, because some of them will never be reached by a mobile host with a "relative slow" mobility behaviour; if the reservation scheme takes into account the average *Cell Stay Time* (CST – the time spent in a cell by a user) then the number of possible (and more probable) future visited cells can be obtained, having the knowledge of the average *Call Holding Time* (CHT – the duration of a connection); in a 1-dimensional (1D) environment the CST analysis is able to full characterize the passive prediction policy. In a 2D environment, considering only the CST will not give all the prediction information that are necessary: this time the number of future visited cells must be accompanied by the identification of every single cell where, with high probability, the user will move. It is evident that an additional statistical treatment is necessary in order to consider the directional behaviour of mobile hosts, under a certain mobility model; in the proposed thesis the Random WayPoint Mobility Model (RWPMM) has been considered, as well as the Smooth Random Mobility Model (SRMM); the latter makes hosts movements smoother and more realistic than the various proposed mobility models in the literature. In particular, directional information must be added to the knowledge of the CST in order to make possible the passive reservation prediction in a 2D environment.

Many simulation campaigns have been carried out: first of all the performances of the proposed CAC and rate-adaptation algorithm have been analysed; different "monitor" campaigns have been carried out in order to evaluate the directional behaviour of users in the 2D system as well as the CST distribution, then the passive reservation policy has been introduced; the enhancements have been appreciated, through the analysis of the obtained curves. The passive reservation is able to guarantee service continuity to MIP users, but it makes the system low utilized because of the large amount of passive and unused bandwidth. In order to avoid the disadvantage, two ideas have been applied:

a) make MDP users able to reuse the passive bandwidth, with a certain dropping probability if the passive reservation must be switched into an active one, after a MIP hand-in event;

b) make MIP users able to multiplex their passive reservations: the CST knowledge is very important in order to obtain an estimation of the hand-in event for each user and for each probably visited cell; in this way it is possible to multiplex the pre-reserved bandwidth until the user makes its own hand-in in the considered cell.

The importance of mobility prediction for wireless systems (such as WLANs) has been outlined, introducing some schemes for passive reservations enhancements. Two mobility models have been considered, RWPMM and SRMM, but the proposed schemes are completely uncorrelated with the employed mobility model. The introduction of the proposed schemes has started with the analysis of a simplified 1D scenario, in order to give some knowledge about the CST evaluation and analysis, in terms of statistical distribution. After the quantitative analysis of CST distribution, an extension has been introduced, by considering the complete 2D space and the directional behaviour of mobile hosts. So after a circular reservation, a directional one has been proposed. The obtained CST distributions have been shown, in terms of mean and standard deviation parameters. In addition, a polynomial regression has been shown to be a good way to obtain a CST evaluation by simply introducing system and mobility parameters. The static scheme and the dynamic one have been proposed and formally described by pseudo-code. Obviously, the performance of the proposed schemes must be evaluated; this will be made in the next chapter.

In this PhD thesis the enhancements of the introduction of the proposed reservation scheme and CAC are shown, with their performances analysis. A gain in terms of QoS and system utilization has been obtained by the integration of the rate-adaptation scheme with a passive-reservation algorithm. If the passive bandwidth multiplexing is also activated, the system reaches a utilization level that is near to the saturation state.

The PhD thesis is concluded with the performance evaluation of the proposed ideas. Initially the 1D CST-based prediction scheme has been investigated and good results have been obtained regarding the MIP QoS guarantees: system utilization can considerably increase if passive reservations are made in an adequate manner and service continuity is always guaranteed for MIP service requests. However the simulated 1D scenario is too simply if compared with real mobile environments, so it has been extended with a 2D clustered simulation scenario, where users can move according to the Random WayPoint or Smooth Random Mobility Models. The same CST-based prediction scheme of the 1D case has been employed in the 2D scenario but, although the prediction error is negligible, too many resources are wasted, because of the high number of C_r cells which are interested by passive reservations; so additional and directional information has been introduced in the prediction algorithms in order to make them more selective. Optimal results have been obtained for some combinations of input parameters.

Chapter 1 – The wireless communication systems, wireless LANs and link modeling

1.1 Cellular systems and history

The need of information availability everytime and everywhere has heavy influenced the rapid growth of wireless communications. The idea of a "mobile information" society is taking the windward and, by now, it is usual to see people communicating with the hand-phones or mobile devices. The growth of wireless communications has began in the end of 80s and it is still going on: it is predicted that in the first 10 years of the 21st century it will continue, until the number of subscribers reaches a value of 1200 millions.

Cellular telephony is the economically most important form of wireless communications: It is characterized by the following properties:

- the information flow is bi-directional. A user can transmit and receive information at the same time;
- the user can be anywhere within a (nationwider or international) network. neither he nor the calling party need to know the user's location: it is the network that has to take the mobility of the user into account;
- a call can originate from either the network, or the user. In other words, a cellular customer can be called, or initiate a call;
- a call is intended only for a single user; other users of the network should not be able to listen in;
- the location of a user can change significantly during a call (high mobility of the users).

Since each user wants to transmit or receive different information, the number of active users in a network is limited. The available bandwidth must be shared between the different users; this is an important difference from broadcast systems, where the number of users (receivers) is unlimited, since they all receive the same information.

In order to increase the number of users, the *cellular principle* is used: the area served by a network provider is divided into a number of subareas, called *cells*. Within each cell, different users have to share the available bandwidth: let us consider in the following the case that each user occupies a different carrier frequency. Even users in neighboring cells have to use different frequencies, in order to keep co-channel interference low. However, for cells that are sufficiently far apart, the same frequencies can be used, because the signals get weaker with increasing distance from their transmitter. Thus, within one country, there can be hundreds or thousands of cells that are using the same frequencies.

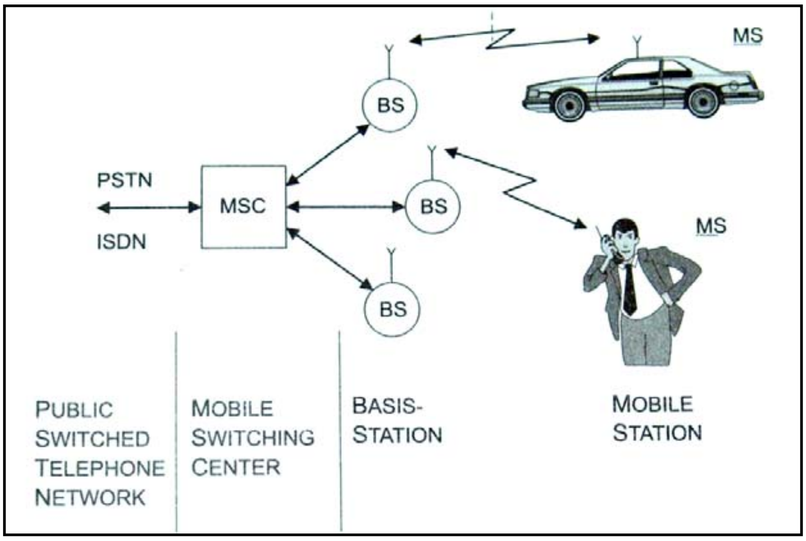


Figure 1.1. Principle of cellular systems.

Figure 1.1 shows a block diagram of a cellular system. A mobile user is communicating with a Base Station (BS) that has a good radio connection with that user. The BSs are connected to a mobile switching center, which is connected to the public telephone system.

Another important aspect of cellular telephony is the unlimited mobility. The user can be anywhere within the coverage area of the network (i.e., is not limited to a specific cell), in order to be able to communicate. Also, he can move from one cell to the other during one call. The cellular network interfaces with the Public Switched Telephone Network (PSTN), as well as with other wireless systems. Now a brief overview is given, in order to appreciate the fundamental steps that led to the actual condition of wireless communications.

These types of systems were born in the 1950s by AT&T's Bell Labs in U.S.A., but they have not got particular interest from users because of their limited number of admissible calls; the engineers worked hard in order to make possible a better utilization of the bandwidth spectrum. The 1970s saw a reviewed interest in cellular communications and in scientific research these years saw the formulation of models for some quantities that are very important to determine the performance of analog telephone systems [1]. From 1971 Bell introduced the *Advanced Mobile Phone Service* (AMPS), the first cellular network, standardized in the U.S.A. in 1982; the AMPS were the most used radio-telephony communication system in the north of America. In the 80s different cellular networks were built in the world and each nation adopted a specific technology for the analogical telephony. For example, England, Italy and Spain chosen the American system, under the name of *Total Access Cellular System* (TACS); Scandinavian countries and France chosen the *Nordic Mobile Telephone* (NMT) [2], while Deutschland chosen the C-Net standard. The networks based on these standards are considered as the First Generation (1G) systems. The analog systems paved the way for the wireless revolution. During the 1980s, they grew at a frenetic pace, and reached market penetrations up to 10% in Europe, though their impact was somewhat less in the U.S.A. In the beginning of the 1980s, the phones were "portable", but definitely not handheld. In most languages, they were just called "carphones", because the battery and transmitter were stored in the trunk of the car and were too heavy to be carried around. But at the end of the 1980s, handheld phones with good speech quality and quite acceptable battery lifetime abounded. The quality had become so good that in some markets, digital phones had difficulties to establish themeselves.

In 1982 the *European Postal and Telecommunication Conference* (CEPT) gives to the *Groupe Speciale Mobile* (GSM) the task to find a newer European telecommunication system; then the *Global System for Mobile communications* (GSM) was born, which represents the first digital system of the Second Generation (2G) ones. The GSM system was developed thoroughout the 1980s and deployment started in early 1990s. Due to additional features, better speech quality and the possibility for secure communications, GSM-based services overtook analog services, typically within 2 years of their introduction. In the U.S.A., the change to digital systems was somewhat slower, but by the end of the 1990s, also this country was overwhelmingly digital.

In the first half of 90s the *World Administrative Radio Conference* (WARC) starts to investigate the possible use of the frequencies near to 2GHz, giving to researchers the opportunity to begin the first studies of the *Universal Mobile Telephony System* (UMTS).

The analogical cellular systems represent a mobile telephony system that offers a total geographic coverage through the use of a radio cells network. These systems use the Frequency Modulation (FM) for the voice transmission and the Frequency Shift Keying (FSK) for the signalling messages. The used radio access is the Frequency Division Multiple Access (FDMA): a frequency channel is assigned to a user for the whole duration of the conversation. The main characteristics and disadvantages are:

- bad hand-off management: a call is dropped if the coverage area is changed during an active call; the user must make again the request;
- too much high dimensions for coverage areas: a single Base Station (BS) covers an area with a radius of 20 or 30 kilometers;
- in order to start a conversation with a mobile user, the calling station must exactly know the right position of the person that must be reached.

The increasing trend of the demand for such kind of services has developed, in the second half of 70s, the development of new radiomobile systems, called cellular systems.

Standard	Frequency band (MHz)	Channel width (kHz)	Number of channels	Nation
AMPS	824-849/869-894	30	832	United States
TACS	890-915/935-960	25	1000	Europe
ETACS	872-905/917-950	25	1240	Great Britain
NMT 450	453-457.5/463-467.5	25	180	Europe
C-450	450-455.74/460-465.74	10	573	Deutschland Portugal
RTMS	450-455/460-465.	25	200	Italy
Radiocom 2000	192.5-199.5/200.5-207.5 215-233.5/207.5-215.5 165.2-168.4/169.8-173 414.8-418/424.8-428	12,5	560 640 256 256	France
JTACS/NTACS	915-925/860-870 898-901/843-846 918.5-922/863.5-867	25/12.5 25/12.5 12.5	400/800 120/240 280	Japan

Table 1.2. Main characteristics of analogical cellular systems.

They born with the aim of making possible a better utilization of the available frequency spectrum with a higher number of channels (about 200 or 1000 if the carrier frequency is 450MHz or 900MHz respectively) with many enhancements on the hand-off procedures. In addition some localization functions were added, in order to know the position of the called station.

Table 1.2 resumes the characteristics of different analogical communication systems. The main American analogical cellular system is the AMPS, with the FM and FSK modulation for voice and signalling. C-450 has been used in Deutschland and Portugal, Radiocom 2000 in France. The hand-over strategies were based on the received power levels of the BSs that are around the mobile host (only the C-450 hand-over strategy is based on the measurements of the propagation delay).

For 2G cellular systems, not only signalling messages are sent in digital form, but also the voice is digitally encoded. The introduction of the A/D conversions has made possible the employment of more robust end flexible access methods, like the *Time Division Multiple Access* (TDMA) or the *Code Division Multiple Access* (CDMA). The main advantages of digital cellular systems, against the analogical ones, are:

- better integration with existing wired digital networks;
- higher number of supported services and high flexibility for mixed data/voice transmissions;
- increased channel capacity through the employment of better voice encoders;
- reduced transmission power, with a related increasing of the batteries duration;
- higher privacy and security with the integration of cryptographic algorithms.

As earlier discussed, in the 1982 the GSM workgroup founded the first digital system: each user can move in every nation where the GSM is installed; it offers automatic roaming functions, also in the international countries. A GSM system has a maximum number of 200 full-duplex channels for each cell: a channel consists of a downlink and an uplink frequencies. A time slot in a channel is assigned to each active station; the dedicated control channel is used for location-updates, registrations and calls configuration. In particular, each BS maintains a database with the list of users that are actually covered by the BS. The shared control channel is subdivided into three logical channels: the first one is dedicated to the paging functions, the seconds one is the random access channel, based on the slotted-ALOHA; the third sub-channel is the guaranteed access channel, through wich the assigned channel is communicated to the mobile host.

3G radiomobile systems are derived from the emerging services requests with high data-rates and a better spectral-efficiency than the older cellular systems [3]. 2G systems were essentially pure voice transmission systems (though some simple data services, like the Short Message Service - SMS - were included as well). The new systems were to provide data transmission at rates comparable with the ill-fated Integrated Services Digital Network (ISDN), and even up to 2Mbps at speeds of up to 500Km/h.

In Europe, 3G systems are identified with the acronym UMTS. The main goals of the UMTS are:

- full coverage and mobility support for the 144Kbps services;
- limited coverage and mobility support for the 3Mbps services;
- higher spectral efficiency than the actual systems;
- high flexibility for new services support.

Many regional standardization institutes has already made their choices about the air interface of UMTS systems. Choosing a particular tecnology in spite of another one depends on technological, commercial and political reasons. In Europe, due to the enormous success of the GSM system, the UMTS is growing up as a natural evolution of the previous system, allowing a smooth switch from GSM to UMTS. The preliminar works of the European Telecommunications Standards Institute (ETSI) has started in the end of 1996; in the beginning 5 different ideas were proposed by 5 different workgroups; later they have been unified. In 1998 a final proposal has been formulated with a double operating modality: in the *Frequency Division Duplex* (FDD) scheme, the UMTS uses the radio access protocol that was originarly proposed by the W-CDMA group; in the *Time Division Duplex* (TDD) scheme the idea of the TD-CDMA group is used. For sake of brevity the treatment of the main known multiple access methods is avoided, but a complete treatment can be found in [5], [15].

1.2 Wireless Local Area Networks (WLANs): an overview

Like personal computers in the 1980s and the Internet in the 1990s, wireless Local Area Networks (wireless LANs - WLANs) are proving to be the next major evolution of technology for businesses. And just as businesses were forced to adopt and provide necessary security for the preceding technologies to keep up with their users, wireless LANs present similar productivity-boosting opportunities while introducing new security concerns. However, the benefits far outweigh the risks when appropriate actions are taken to minimize those risks. The adoption of personal computers in the 1980s led to the creation of local-area networks that laid the initial roads to allow communication to flow like automobiles through a city. A decade later the Internet created the highways that efficiently connect each locality to the other.

Today, wireless LANs introduce the concept of complete mobility provided by air travel; communication is no longer limited to the infrastructure of wires. This provides new opportunities and challenges. Wireless LANs offer a quick and effective extension of a wired network or standard LAN. By simply installing access points to the wired network, personal computers and laptops equipped with wireless LAN cards can connect with the wired network at broadband speeds from up to 275 meters from the access point.

Over the last few years, most deployments of Wireless LANs have been on the 802.11b standard that operates over the unregulated 2.4 GHz frequency spectrum. The 802.11b standard offers connectivity of up to 11 Mbps (fast enough to handle large e-mail attachments and run bandwidth-intensive applications like video conferencing). While the 802.11b standard now dominates the wireless LAN market, other variations of the 802.11 standard [4], such as 802.11a and 802.11g are being developed to handle increased speeds [6]. Wireless LAN vendors have also committed to support a variety of standards ([14], [16]) and different standards have been proposed by the IEEE, like HiperLAN 1 [7], HiperLAN 2 [8], [10] and Bluetooth [9].

The main job of the MAC protocol is to regulate the usage of the medium and this is done through a channel access mechanism. A channel access mechanism is a way to divide the main resource between nodes, the radio channel, by regulating the use of it. It tells each node when it can transmit and when it is expected to receive data. The channel access mechanism is the core of the MAC protocol.

The most used MAC mechanisms in wireless environments are TDMA, CSMA and POLLING; they are the 3 main classes of channel access mechanisms for radio systems [12].

Wireless LAN technology allows untethered workers to connect to the corporate network from a conference room, the cafeteria or a bench outside the building at speeds 200-times faster than a dial-up modem. Businesses are quickly deploying new networks without the costs and time of wiring offices and workstations. Installation can be accomplished in days rather than weeks by simply attaching wireless access points to wired high-speed networks.

The benefits of deploying wireless LANs can be summarized as the following:

- Mobility: boost productivity with the convenience of wirelessly connecting to the network from any point within range of an access point;
- Rapid & flexible deployment: quickly extend a wired network with the ease of attaching an access point to a high-speed connection;
- Application agnostic: as an extension of the wired network, wireless LANs work with all existing applications;
- Attractive price: deploying a wireless LAN can be cheaper than a wired LAN;

1.2.1 Anatomy of a wireless LAN

A radio network is a collection of nodes communicating together through radio devices, using radio waves to carry the information exchanged. It is sometime called a radio Ethernet, by analogy of the wired technology. Most radio devices are a card (ISA, PCMCIA) to plug in a PC (or workstation), and interact directly with the standard networking stack on it.

A radio device is composed of two main parts (figure 1.3). The first is the radio modem. This is the part transmitting (modulating) the data onto the frequency and receiving other transmissions. It is composed of antenna(s), amplificators, frequency synthesisers and filters. The modem main characteristics are the frequency band, the signalling rate, the modulation and the transmitted power.

The second part of the radio device is the MAC controller, responsible to run the MAC protocol. This is implemented mainly in an ASIC and/or a microcontroller on the card, but some functionalities of the MAC may be as well in the driver on the PC.

The card also includes some memory for the MAC controller to store incoming and outgoing packets (buffers) and other data (configuration, statistics). Most of the time the few most time critical parts are handled in the radio modem ASIC (the baseband), the bulk of the MAC in a microcontroller and only some management functionality in the driver. But, the different manufacturers place the boundary between the different functionalities differently (cost/performance tradeoff), and some have implemented driver only MACs for lower cost. The main characteristics of the MAC are the packet format (size, headers), the channel access mechanisms and the network management features. The amount of on-board memory is also important, because the MAC may need a significant number of buffers to compensate the PC and interface latencies.

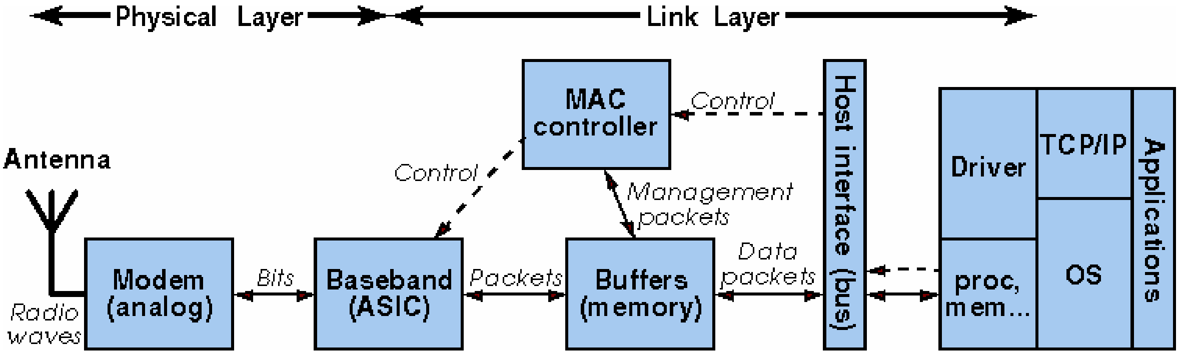


Figure 1.3. Functional diagram of a wireless device.

The card interface to the PC through one of its buses (ISA, PCI, Pcmcia...) or communication ports (serial, parallel, USB or Ethernet). This interface allows the software (mostly the driver) to communicate with the MAC controller and most of the time directly to the on board memory (the software writes packets to a specific location of it, then the controller reads them and sends them). The main characteristic of the interface is mainly the speed (i/o, shared memory or DMA) and the ability to process requests in parallel.

1.3 Fading, interferences and transmission errors

Fading defines all the temporal variations of the signal attenuation due to its propagation in a real environment like an office or a house. The radio signal interacts in various ways with the environment, so it varies a lot with the environment configuration. Moving a few centimetres can make a big different in signal quality. Moreover, the environment is not static: humans are moving, things are moving and the nodes may be moving themselves. All these small movements may produce important variations in time in the attenuation of the signal. For example the propagation between two nodes may alternate from poor to good on a packet basis. People usually describe the pattern of attenuation with a Rayleigh fading model (case where there is no line of sight) or a Ricean model (line of sight plus additional paths). The main consequence is that transmission errors on the channel tend to be clustered, following a Gaussian distribution. Fading causes transmissions errors that need to be overcomed by the system. Of course, recovering from these errors will add overhead. The greater the range the greater will be the impact of the fading and the system will degrade with higher range until it looses communication. The most efficient technique to overcome the effect of fading is antenna diversity (a set of directional antenna, in spite of an omnidirectional one). The deployment of unlicensed systems is totally uncoordinated. So, other radio systems operating in the area create interferences. This includes other WLANs, cordless phones (900MHz and now 2.4GHz) and other communication systems. The 2.4GHz band is also the frequency where water molecules resonate, so it is used for microwave oven. Domestic microwave oven generates a limited amount of interferences, the various regulations limit the power of the radiation they can leak to less than 1W; they emit periodic short bursts and pollute only a limited portion of the 2.4GHz band. Commercial microwave ovens (for example a huge dryer in a paper factory) generate much more interferences. The result of interferences is that packets collide with interference signal and can be received corrupted. If the SNR between the packet and the interferer is high enough, the receiver can "capture" the packet, otherwise it is corrupted. Most Wireless LANs cope very well with interferers, in fact usually much better than cordless phones, but interferences do reduce performance.

The most obvious way to overcome transmission errors is to use Forward Error Correction (FEC), which provides additional redundant information that is opportunely added to transmitted data and correctly decoded at the receiver side. It goes further than Cyclic Redundancy Check (CRC) which just detects errors; FEC adds in every transmission some additional redundancy bits. Depending on the number of bits added and the FEC code used (the strength of the code), this allows repairing a certain number of errors in the transmission. FEC has been used with success in many systems and the Turbo Codes are probably the most efficient one: they are very close to the Shannon limit in a Gaussian channel. In other words, if the error follow a Gaussian distribution (and the parameters are known), there is a turbo code nearly optimal giving the highest throughput in this channel.

Unfortunately for us, errors on a radio channel (for WLAN) follow a fading model and they are clustered. This means that most of the time the signal is strong, so the packet is error free, but when the signal is weak the packet contains lots of error. Interferences have roughly the same effect as fading, either the packet is collision free so intact, or when a collision occurs most of the packet is corrupted. To correct all those errors in corrupted packets, it would require a very strong FEC code. Unfortunately, this code would add lots of redundancy bits, so lots of overhead. A normal FEC code would add less overhead, but be useless with the correct packets and inefficient with the highly corrupted packets. So, for WLANs, using FEC tends to be ineffective against fading and interferers and no WLAN implements FEC. A much better solution is to use retransmissions (just retransmit the original packet in case of errors - some form of packet scheduling and retransmission has been proven to be nearly optimal in Rayleigh fading channels). This is usually implemented at the MAC level. However, in a few cases, FEC might be needed in WLANs. Some receivers, either due to poor implementation or specific design, generate random (Gaussian) errors and they might benefit from FEC.

Radio waves reflect or diffract on obstacles and are attenuated differently by different materials. This is exactly like light, which goes through glass, is reflected by mirrors and stop by most obstacles, except that much more materials are transparent or reflector to radio than to light. In a real environment like an office or a house, there is a lot of surface reflecting radio (walls, ceilings, metal), being semi-transparent to radio (walls, ceilings, humans) or opaque to radio (metal). This gives trouble estimating the range of the system. This also means that the signal received at a node may come from different directions (depending on reflections on the environment) with different strength (depending on attenuations), and the receiver sees only the combinations of all these reflections. This phenomenon is called multipath. Most of the time, multipath is good, because the addition of all the reflections of the signal increase its strength. The main effect of multipath is that range is very difficult to evaluate and the receiver experiences some kind of fading. The main problem of multipath is that it creates delay spread, as illustrated in figure 1.4. Depending on the number of reflections and the propagation speed in different signals, they do not arrive exactly at the same time at the receiver. It is like the "echo" you may hear in the mountains, the signal going directly will be faster than one reflecting twice on the walls. Of course, as radio propagates at the speed of light, those differences are very small (below the microsecond). But, when the bitrate of the system increases, those time differences becomes significant with regards to the symbol time, to the point of creating destructive interferences (the current symbol will be corrupted by the echo of the previous symbols). Bit rate lower than 1Mb/s are relatively immune to delay spread problems (the symbol time is 1μs and higher), but as the bit rate increases above 1Mb/s, the effect of delay spread increases. It is considered that systems faster than 5Mb/s should have some technique to overcome delay spread [11].

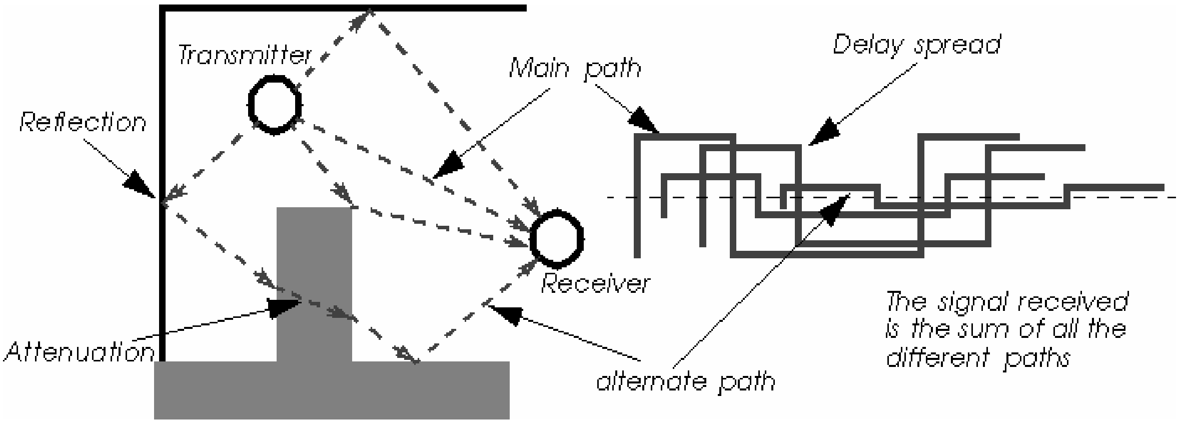


Figure 1.4. Multipath and delay spread.

1.3.1 Characterization of fading multipath channels

If we transmit an extremly short pulse, ideally an impulse, over a time-varying multipath channel, the received signal might appear as a train of pulses, as shown in figure 1.5. Hence, one characteristic of a multipath medium is the time spread introduced in the signal that is transmitted through the channel.

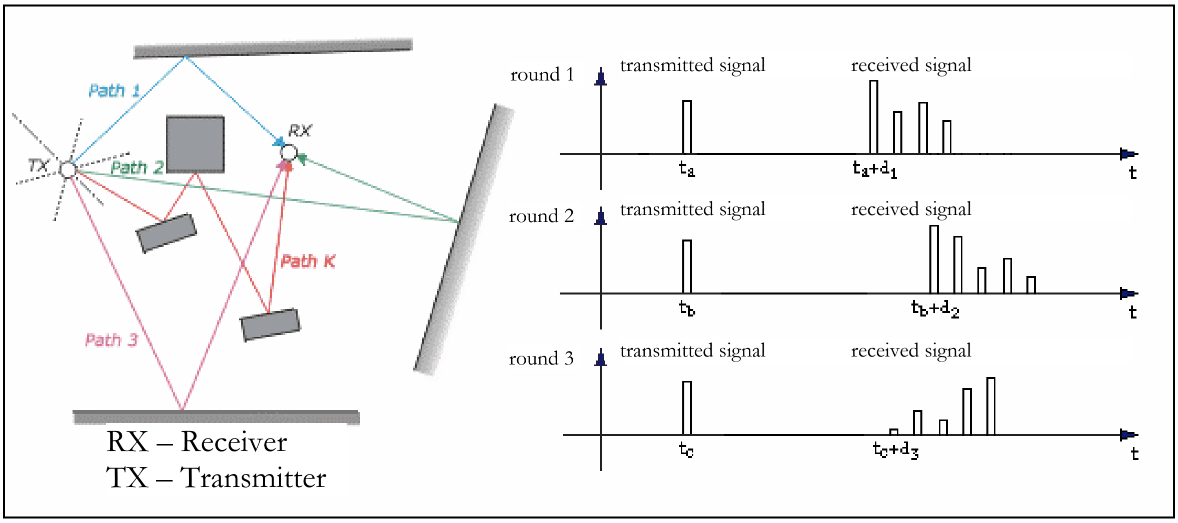


Figure 1.5. Example of the response of a multipath channel to a very narrow pulse.

A second characteristic is due to the time variations in the structure of the medium. As a result of such time variations, the nature of the multipath varies with time. That is, if we repeat the pulse-sounding experiment over and over, we shall observe changes in the received pulse train, which will include changes in the sizes of the individual pulses, changes in the relative delays among the pulses, and, quite often, changes in the number of pulses observed in the received pulse train, as shown in figure 1.5. Moreover, the time variations appear to be unpredictable to the user of the channel statistically. Toward this end, let us examine the effects of the channel on a transmitted signal that is represented in general as:

$$s(t) = \mathrm{Re}\left[s_l(t)e^{j2\pi f_c t}\right] \tag{1.1}$$

We assume that there are multiple propagation paths. Associated with each path is a propagation delay and an attenuation factor. Both the propagation delays and the attenuation factors are time-variant as a result of changes in the structure of the medium. Thus, the received band-pass signal may be expressed in the form:

$$x(t) = \sum_n \alpha_n(t) s[t - \tau_n(t)], \tag{1.2}$$

where $\alpha_n(t)$ is the attenuation factor for the signal received on the n-th path and $\tau_n(t)$ is the propagation delay for the n-th path. Substituting for *s(t)* from eq. (1.1) into eq. (1.2) yields the result:

$$x(t) = \mathrm{Re}\left(\left\{\sum_n \alpha_n(t) e^{-j2\pi f_c \tau_n(t)} s_l\left[t - \tau_n(t)\right]\right\} e^{j2\pi f_c t}\right), \tag{1.3}$$

it is apparent from eq. (1.3) that the equivalent low-pass received signal is:

$$r_l(t) = \sum_n \alpha_n(t) e^{-j2\pi f_c \tau_n(t)} s_l\left[t - \tau_n(t)\right]. \tag{1.4}$$

Now let us consider the transmission of an unmodulated carrier at frequency f_c. Then $s_l(t)=1$ for all t, and, hence, the received signal for the case of discrete multipath, given by eq. (1.4), reduces to ($\theta_n(t)=2\pi f_c \tau_n(t)$):

$$r_l(t) = \sum_n \alpha_n(t) e^{-j\theta_n(t)}. \tag{1.5}$$

Thus, the received signal consists of the sum of a number of time-variant vectors (phasors) having amplitude $\alpha_n(t)$ and phase $\theta_n(t)$. Note that large dynamic changes in the medium are required for $\alpha_n(t)$ to change suffciently to cause a significant change in the received signal. On the other hand, $\theta_n(t)$ will change by 2π rad. whenever τ_n changes by $1/f_c$. But $1/f_c$ is a small number and, hence, θ_n can change by 2π rad. with relatively small motions of the medium. We also expect the delays $\tau_n(t)$ associated with the different signal paths to change at different rates and in an unpredictable (random) manner. This implies that the received signal $r_l(t)$ in eq. (1.5) can be modeled as a random process. When there are a large number of paths, the central limit theorem can be applied. That is, $r_l(t)$ may be modeled as a complex-valued Gaussian random preocess, as well as the time-variant impulse response $c(\tau;\ t)$. The multipath propagation model for the channel embodied in the received signal $r_l(t)$, given in eq. (1.5), results in signal fading. The fading phenomenon is primarily a result of the time variations in the phases $\{\theta_n(t)\}$. That is, the randomly time variant phases $\{\theta_n(t)\}$associated with the vectors $\{\alpha_n e^{-j\theta n}\}$ at times result in the vectors adding destructively. When that occurs, the resultant received signal $r_l(t)$ is very small or pratically zero. At other times, the vectors $\{\alpha_n e^{-j\theta n}\}$ add constructively, so that the received signal is large. Thus, the amplitude variations in the received signal, termed signal fading, are due to the time-variant multipath characteristics of the channel.

When the impulse response *c(τ; t)* is modeled as a zero-mean complex-valued Gaussian process, the envelope |*c(τ; t)*| at any instant *t* is Rayleigh-distributed. In this case the channel is said to be a *Rayleigh fading channel.* In the event that there are fixed scatterers or signal reflectors in the medium, in addition to randomly moving scatterers, *c(τ; t)* can no longer be modeled as having zero-mean. In this case, the envelope |*c(τ; t)*| has a Rice distribution and the channel is said to be a *Ricean fading channel.* Another probability distribution function that has been used to model the envelope of fading signals is the *Nakagami-m* distribution [11]. The Rayleigh model is often used when there are many indirect paths from the transmitter to the receiver and there is not a dominant Line Of Sight (LOS) path, while the Ricean model takes into account a sigle and dominant LOS path from the transmitter to the receiver; that is, the Ricean model is used for indoor environments, while the Rayleigh one for outdoor environments.

1.3.2 Channel correlation functions and power spectra

In order to well understand the channel behaviour, some parameters must be defined. Our starting point is the equivalent low-pass impulse response *c(τ; t)*, which is characterized as a complex-valued random process in the *t* variable. We assume that *c(τ; t)* is wide-sense stationary. Then we define the autocorrelation function of *c(τ; t)* as:

$$\phi_c(\tau_1,\tau_2;\Delta t)=\frac{1}{2}E[c^*(\tau_1;t)c(\tau_2;t+\Delta t)]. \tag{1.6}$$

In most radio transmission media, the attenuation and phase shift of the channel associated with path delay τ_1 is uncorrelated with the attenuation and phase shift associated with path delay τ_2. This is usually called uncorrelated scattering. We make the assumption that the scattering at two different delays is uncorrelated and incorporate it into eq. (1.6) to obtain:

$$\frac{1}{2}E[c^*(\tau_1;t)c(\tau_2;t+\Delta t]=\phi_c(\tau_1;\Delta_t)\delta(\tau_1-\tau_2), \tag{1.7}$$

if we let Δ_t=0, the resulting autocorrelation function $\phi_c(\tau;0)\equiv\phi_c(\tau)$ is simply the average power output of the channel as a function of the time delay *τ*. For this reason, $\phi_c(\tau)$ is called the *multipath intensity profile* or the *delay power spectrum* of the channel.

In general, $\phi_c(\tau;\Delta t)$ gives the average power output as a function of the time delay τ and the difference Δt in observation time.

In practice, the function $\phi_c(\tau;\Delta t)$ is measured by transmitting very narrow pulses orm equivalently, a wideband signal and cross-correlating the received signal with a delayed version of itself. Typically, the measured function $\phi_c(\tau)$ may appear as shown in figure 1.6. The range of values of τ over which $\phi_c(\tau)$ is essentially nonzero is called the multipath spread of the channel and is denoted by T_m.

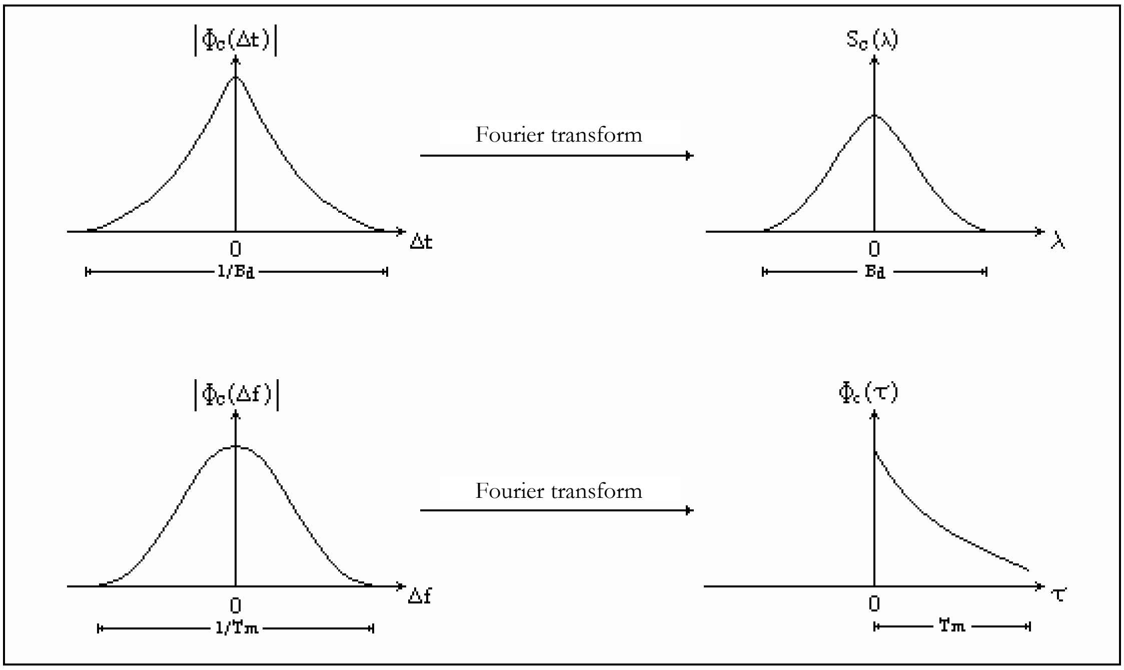


Figure 1.6. Trends and reletions of the defined correlation functions.

A completely analogous characterization of the time-variant multipath channel begins in the frequency domain. By taking the Fourier transform of $c(\tau;t)$, we obtain the time-variant transfer function $C(f;t)$, where f is the frequency variable. If $c(\tau;t)$ is modeled as a complex-valued zero-mean Gaussian random process in the t variable, it follows that $C(f;t)$ also has the same statistics. Under the assumption that the channel is wide-sense stationary, we define the autocorrelation function:

$$\phi_C(f_1, f_2;\Delta t) = \frac{1}{2} E[C^*(f_1;t)C(f_2;t+\Delta t)], \tag{1.8}$$

since C(f;t) is the Fourier transform of $c(\tau;t)$, $\phi_C(f_1,f_2;\Delta t)$ is related to $\phi_c(\tau;\Delta t)$ by the Fourier transform. Thus, $\phi_C(f_1,f_2;\Delta t) = \phi_C(\Delta f;\Delta t)$, where $\Delta f = f_2 - f1$.

Furthermore, the assumption of uncorrelated scattering implies that the autocorrelation function of $C(f;t)$ in frequency is a function of only the frequency difference Δf. Therefore, it is appropriate to call $\phi_C(\Delta f;\Delta t)$ the *spaced-frequency, spaced-time correlation function of the channel.* It can be measured in practice by transmitting a pair of sinusoids separated by Δf and cross-correlating the two separately received signals with a relative delay Δt. If we set Δt=0 then $\phi_C(\Delta f;0)=\phi_C(\Delta f)$ and $\phi_c(\tau;0)$; the relationship is shown in figure 1.6. Since $\phi_C(\Delta f)$ is an autocorrelation function in the frequency variable, it provides us with a measure of the frequency coherence of the channel. As a result of the Fourier transform relationship between $\phi_C(\Delta f)$ and $\phi_c(\tau)$, the reciprocal of the multipath spread is a measure of the *coherence bandwidth of the channel.* That is:

$$(\Delta f)_c \cong \frac{1}{T_m}, \tag{1.9}$$

where $(\Delta f)_c$ denotes the coherence bandwidth. Thus, two sinusoids with frequency separation greater than $(\Delta f)_c$ are affected differently by the channel. When an information-bearing signal is transmitted through the channel, if $(\Delta f)_c$ is small in comparison to the bandwidth of the transmitted signal, the channel is said to be *frequency-selective.* In this case, the signal is severely distorted by he channel. On the other hand, if $(\Delta f)_c$ is large in comparison with the bandwidth of the transmitted signal, the channel is said to be *frequency-nonselective.*

We now focus our attention on the time variations of the channel as measured by the parameter Δt in $\phi_C(\Delta f;\Delta t)$. The time variations in the channel are evidenced as a Doppler broadening and, perhaps, in addition as a Doppler shift of a spectral line. In order to relate the Doppler effects to the time variations of the channel, we define the Fourier transform of $\phi_C(\Delta f;\Delta t)$ with respect to the variable Δt to be the function:

$$S_C(\lambda) = \int_{-\infty}^{+\infty} \phi_C(0;\Delta t)e^{-j2\pi\lambda\Delta t}d\Delta t, \tag{1.10}$$

with Δf set to 0.

As depicted in figure 1.6, the function $S_C(0;\lambda)=S_C(\lambda)$ is a power spectrum that gives the signal intensity as a function of the Doppler frequency λ. Hence, we call $S_C(\lambda)$ the Doppler *power spectrum of the channel.* From eq. (1.10) we observe that if the channel is time-invariant, $\phi_C(\Delta t)$=1 and $S_C(\lambda)$ becomes equal to the delta function $\delta(\lambda)$. Therefore, when there are no time variations in the channel, there is no spectral broadening observed in the transmission of a pure frequency tone. The range of values of λ over wich $S_C(\lambda)$ is essentially nonzero is called the Doppler *Spread B_d of the channel.* Since $S_C(\lambda)$ is related to $\phi_C(\Delta t)$ by the Fourier transform, the reciprocal of B_d is a measure of the coherence time of the channel. That is:

$$(\Delta t)_c \cong \frac{1}{B_d}, \tag{1.11}$$

where $(\Delta t)_c$ denoted the *coherence time.*

Clearly, a slowly changing channel has a large coherence time or, equivalently, a small Doppler spread. Finally, the scattering function of the channel can be defined as:

$$S(\tau;\lambda) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \phi_C(\Delta f;\Delta t) e^{-j2\pi\lambda\Delta t} e^{j2\pi\tau\Delta f} d\Delta t d\Delta f. \tag{1.12}$$

It provides us with a measure of the average power output of the channel as a function of the time delay τ and the Doppler frequency λ.

1.3.3 Effects of signal characteristics on the choice of a channel model

Having discussed the statistical characterization of time-variant multipath channels generally in terms of the correlation functions, we now consider the effect of signal characteristics on the selection of a channel model that is appropriate for the specific signal. Thus, let $s_l(t)$ be the equivalent low-pass signal transmitted over the channel and let $S_l(f)$ denote its frequency content. Then the equivalent low-pass received signal, exclusive of additive noise, may be expressed either in terms of the time domain variables $c(\tau;t)$ and $s_l(t)$ as:

$$r_l(t) = \int_{-\infty}^{+\infty} C(f;t) S_l(f) e^{j2\pi ft} df. \tag{1.13}$$

Suppose we are transmitting digital information over the channel by modulating (either in amplitude, or in phase, or both) the basic pulse $s_l(t)$ at a rate $1/T$, where T is the signaling interval. It is apparent from eq. (1.13) that the time-variant channel characterized by the transfer function $C(f;t)$ distorts the signal $S_l(f)$. If $S_l(f)$ has a bandwidth W greater than the coherence bandwidth $(\Delta f)_c$ of the channel, $S_l(f)$ is subjected to different gains and phase shifts across the band. In such a case, the channel is said to be frequency-selective. Additional distortion is caused by the time variations in $C(f;t)$. This type of distortion is evidenced as a variation in the received signal strength, and has been termed *fading*. It should be emphasized that the frequency selectivity and fading are viewed as two different types of distortion. The former depends on the multipath spread or, equivalently, on the coherence bandwidth of the channel relative to the transmitted signal bandwidth W. The latter depends on the time variations of the channel, which are grossly characterized by the coherence time $(\Delta t)_c$ or, equivalently, by the Doppler spread B_d.

The effect of the channel on the transmitted signal $s_l(t)$ is a function of our choice of signal duration. For example, if we select the signaling interval T to satisfy the condition $T>>T_m$, the channel introduces a negligible amount of intersymbol interference. If the bandwidth of the signal pulse $s_l(t)$ is $W\approx 1/T$, the condition $T>>T_m$ implies that:

$$W << \frac{1}{T_m} \approx \left(\Delta f\right)_c . \tag{1.14}$$

That is, the signal bandwidth W is much smaller than the coherence bandwidth of the channel. Hence, the channel is frequency-nonselective. In other words, all the frequency components in $S_l(f)$ undergo the same attenuation and phase shift in transmission through the channel. But this implies that, within the bandwidth occupied by $S_l(f)$, the time-variant transfer function $C(f;t)$ of the channel is a complex-valued constant in the frequency variable. Since $S_l(f)$ has its frequency content concentrated in the vicinity of $f=0$, $C(f;t)=C(0;t)$. Consequently eq. (1.13) reduces to:

$$r_l(t) = C(0;t)\cdot \int_{-\infty}^{\infty} S_l(f)e^{j2\pi ft}df = C(0;t)\cdot s_l(t). \tag{1.15}$$

Thus, when the signal bandwidth W is much smaller than the coherence bandwidth $(\Delta f)_c$ of the channel, the received signal is simply the transmitted signal multiplied by a complex-valued random process $C(0;t)$, which represents the time-variant characteristics of the channel. In this case, we say that the multipath components in the received signal are not resolvable because $W<<(\Delta f)_c$. The transfer function C(0;t) for a frequency-nonselective channel may be expressed in the form:

$$C(0;t) = \alpha(t)\cdot e^{-j\phi(t)}, \tag{1.16}$$

where $\alpha(t)$ represents the envelope and $\phi(t)$ represents the phase of the equivalent low-pass channel. When $C(0;t)$ is modeled as zero-mean complex-valued Gaussian random process, the envelope $\alpha(t)$ is Rayleigh-distributed for any fixed value of t and $\phi(t)$ is uniformly distributed over the interval $(-\pi,\pi)$. The rapidity of the fading on the frequency-nonselective channel is determined either from the correlation function $\phi_C(t)$ or from the Doppler power spectrum $S_C(\lambda)$. Alternatively, either of the channel parameters $(\Delta t)_c$ or B_d can be used to characterize the rapidity of the fading.

For example, suppose it is possible to select the signal bandwidth W to satisfy the condition $W<<(\Delta f)_c$ and the signaling interval T to satisfy the condition $T<<(\Delta t)_c$. Since T is smaller than the coherence time of the channel, the channel attenuation and phase shift are essentially fixed for the duration of at least one signaling interval. When this condition holds, we call the channel a *slowly fading channel.* Furthermore, when $W\approx 1/T$, the conditions that the channel be frequency-nonselective and slowly fading imply that the product of T_m and B_d must satisfy the condition $T_mB_d<1$. The product TmBd is called the *spread factor* of the channel. If $T_mB_d<1$, the channel is said to be underspread; otherwise, it is overspread.

If the spread factor is known and if the channel is underspread, it is possible to select the signal $s_l(t)$ such that these channels are frequency-nonselective and slowly fading. The slow-fading condition implies that the channelcharacteristics vary sufficiently slowly that they can be measured. Since the multipath components in the received signal are not resolvable when the signal bandwidth W is less than the coherence bandwidth $(\Delta f)_c$ of the channel, the received signal appears to arrive at the receiver via a single fading path. On the other hand, we may choose $W>>(\Delta f)_c$, so that the channel becomes frequency-selective. Under this condition, the multipath components in the received signal are resolvable with a resolution in time delay of $1/W$. Thus, the frequency-selective channel can be modeled as a tapped delay line (transversal) filter with time-variant tap coefficients [11].

1.4 Analysis and modeling of a slow-fading frequency-nonselective channel

The slow-fading channel is the simpliest to study, but it well describes the performance of digital signaling over fading channels, giving some information about the waveforms to be used in order to attenuate the distorsions that are intrinsic in the wireless communications. Before approaching the study of the channel behaviour, some concepts about stochastic processes must be mentioned.

A stochastic process is a family of random variables, all defined over the same samples space Ω and indexed through a t parameter that varies in the index set T:

$$\{X_t(\omega), \omega \in \Omega,\ \ t \in T\}. \tag{1.17}$$

The difference between two stochastic processes is the type of dependence existing among the random variables that compose the processes; this leads to a certain difficulty in the mathematical treatment of the problems. From the definition and from eq. (1.17) a stochastic process can be also considered as a temporal evolution of a non-deterministic system (in our case a wireless link). For our purposes, the real temporal axis (or a discrete subset) can be associated to the set T and the possible channel states to the samples space Ω, so: $X(\omega,t)$ (or $X_t(\omega)$ if t is continuous) will represent the system state at time t; a process is said to be "*discrete values*" if the random variables can assume a finite set of values. A stochastic process is said to be "*markovian*" when, fixed an observation time instant t_k, the process evolution, beginning from t_k, depends only on t_k and not on all the previous time instants:

$$P(X(t_{k+1}) = x_{k+1} \mid X(t_k) = x_k \cap X(t_{k-1}) = x_{k-1} \cap ... \cap X(t_1) = x_1) \overset{\Delta}{=} P(X(t_{k+1}) = x_{k+1} \mid X(t_k) = x_k), \tag{1.18}$$

Eq. (1.18) shows the so called *chain dependence property*. In our work the discrete-time, discrete-values and finite-state Markov processes (also called Markov chains, where $X(t_k)\equiv X_k$) will be employed, in order to analyse the behaviour of a wireless link.

1.4.1 Wireless channel modeling through Finite-State Markov Chain (FSMC)

The study of the Finite-State Markov channel (FSMC) emerges from early works ([17],[18]). They study a two-state Markov channel known as the Gilber-Elliot (GE) channel. In their model, each state corresponds to a specific channel quality which is either noiseless or totally noisy. In general, a Binary Symmetric Channel (BSC) with a given crossover probability can be associated with each state so that the channel quality for each state can be identified. The GE channel is the special case where the crossover probabilities of the BSC's are 0 and 0.5, respectively. In some cases, modeling a radio communication channel as a two-state GE channel is not adequate when the channel puality varies dramatically. A straightforward solution is to form a channel model with more than two states.

Let $S=\{s_0, s_1, s_2, \ldots, s_{K-1}\}$ denote a finite set of states and $\{S_n\}$, n=0,1,2,… be a constant Markov process. Since the constant Markov process has the property of stationary transitions, the transition probability is independent of the time index *n* and can be written:

$$t_{j,k} = \Pr(S_{n+1} = s_k \mid S_n = s_j), \tag{1.18}$$

for all $n=0,1,2,\ldots$ and $j,k\in\{0,1,2,\ldots,K-1\}$.

With this definition, we can define a *KxK* state transition probability matrix $\boldsymbol{T}$ with its elements $t_{j,k}$ as in (1.18). Note that a state transition probability matrix has the property that the sum of the elements on each row is equal to 1, or:

$$\sum_{l=0}^{K-1} t_{k,l} = 1, \qquad \forall k \in \{0,1,2,\ldots,K-1\}. \tag{1.19}$$

Moreover, with the stationary transition property, the probability of state *k* at any permissible time index *n* without any state information at other time indices can also be defined as:

$$p_k = \Pr(S_n = s_k), \ \ k \in \{0,1,2,\ldots,K-1\}. \tag{1.20}$$

A Kx1 steady state probability vector $\boldsymbol{p}$ can be defined with its elements p_k as in eq. (1.20). In many cases, this vector can be served as the set of initial state probabilities. Note that eq. (1.18) and eq. (1.20) must satisfy the equilibrum condition which states that for any given state k, the incoming flow and outgoing flow must be equal. That is:

$$\sum_{j=0}^{K-1} p_j t_{j,k} = p_k, \qquad \forall k \in \{0,1,2,\ldots,K-1\}, \tag{1.21}$$

or simply:

$$\boldsymbol{p}^t \boldsymbol{T} = \boldsymbol{p}^t, \tag{1.22}$$

where $\boldsymbol{p}^t$ is the transpose of $\boldsymbol{p}$.

A complete description of a finite-state Markov channel requires additional information on the channel quality for each state. Define a *Kx1* crossover probability vector $\boldsymbol{e}$ with its elements e_k, $k\in\{0, 1, 2, \ldots, K\text{-}1\}$, being the crossover probability of the binary symmetric channel associated with state k. A FSMC is then uniquely defined by $\boldsymbol{T}$, $\boldsymbol{p}$ and $\boldsymbol{e}$. The overall average error probability $\boldsymbol{e}$ of the FSMC is then:

$$e_m = p^t e = \sum_{k=0}^{K-1} p_k e_k. \tag{1.23}$$

The choices of $\boldsymbol{T}$, $\boldsymbol{p}$ and $\boldsymbol{e}$ may not be arbitrary. Actually, in addition to eq. (1.19) (1.21) and (1.23) there are four obvious constraints imposed on $\boldsymbol{p}$ and $\boldsymbol{e}$ as follows:

$$1)\ \ 0 < p_k \le 1,\ \ \forall k \in \{0,1,2,...,K-1\}; \tag{1.24}$$

$$2)\ \ \sum_{k=0}^{K-1} p_k = 1; \tag{1.25}$$

$$3)\ \ 0 \le e_k \le 0.5,\ \ \forall k \in \{0,1,2,...,K-1\}; \tag{1.26}$$

$$4)\ \ e_i \ne e_j,\ \ se\ \ i \ne j\,\forall i,j \in \{0,1,2,...,K-1\}. \tag{1.27}$$

Most studies on the performance of the GE channel were based on variations of ***T***, ***p*** and ***e*** under the constraints mentioned above. A natural question is how can the performance be optimized with respect to the channel parameters subject to the four constraints given above. An attempt to solve this question motivates the study of [19], on establishing a connection between Rayleigh fading channels and their FSMC models such that the transmission technologies thereby designed can be applied to the channel efficiently. As done in [19], a partitioning of the received SNR into a finite number of intervals leads to a FSMC model. With the partitioning, the elements in ***T***, ***p*** and ***e*** can be obtained and the corresponding FSMC model is established.

When the Channel State Information (CSI) is available, the capacity C^{CSI} is simply the average capacity over all the states [20], or:

$$C^{CSI} = \sum_{k=0}^{K-1} p_k \left[1 - h(e_k)\right], \tag{1.28}$$

where *h(·)* is the binary entropy function defined as:

$$h(e) = e \cdot \log\frac{1}{e} + (1-e)\cdot\log\frac{1}{1-e}\ . \tag{1.29}$$

On the other hand, it is quite difficult to calculate the capacity of the finite-state Markov channel when CSI is not available.

1.4.2 The Rayleigh fading channel and its finite-state model

Unreasonable results can be concluded for the FSMC if the obtained model does not represent a real physical channel. Therefore, the methodology in establishing the relationship between physical channels and their finite-state models is important. In this section a typical radio communication channel, namely Rayleigh fading channel, is presented; this channel produces time-varying received SNR characterizing the channel quality in terms of the average error probability. By partitioning the range of the received SNR into a finite number of intervals, a finite-state model for the Rayleigh fading channel is built. The physical characteristics and the time-varying behaviour of a Rayleigh fading channel are introduced. Here, a Markov chain is used to describe the transition activities between states from one observation time to the next.

The digital cellular radio transmission environment usually consists of a large number of scatterers that result in multiple propagation paths. Associated with each path is a propagation delay and an attenuation factor depending on the obstacles in the path that reflects the elecromagnetic waves. When a Continuous Waveform (CW) is transmitted, the multipath effect results in the fluctuation of the received signal envelope that is Rayleigh distributed. The channel is known as a Rayleigh fading channel. The time variations of the signal level are characterized by the Doppler frequency effect, which is due to the motion of the mobile terminal.

Let A denote the received SNR which is proportional to the square of the signal envelope. The probability density function (pdf) of A is exponential [11] and can be written as:

$$p_A(a) = \frac{1}{\rho} e^{-\frac{a}{\rho}}, \qquad with\, a \geq 0 \quad and \quad \rho = E[A]. (1.30)$$

Following the same treatment in [1], let f_m be the maximum Doppler frequency defined as:

$$f_m = \frac{v}{\lambda} \ , \qquad (1.31)$$

where v is the speed of the vehicle and λ is the wavelength. Then, let N_a be the expected number of times per second the received SNR A passes downward across a givel level a, we have:

$$N_a = \sqrt{\frac{2\pi a}{\rho}} f_m e^{-\frac{a}{\rho}}. \qquad (1.32)$$

As early mentioned, a finite-state state Markov channel can be uniquely defined by the state transition matrix, the initial state probability vector and the error probability vector. It was also noted that any partition of the received SNR into a finite number of intervals forms a finite-state channel model. Let $0=A_0<A_1<A_2<\ldots<A_k=\infty$ ($A_0=0$ if expressed in dB, $A_0=1$ else) be the thresholds of the received signal to noise ratio. The the Rayleigh fading channel is said to be in state s_k, k$=0, 1, 2,..., K-1$, if the received SNR is in the interval $[A_k, A_{k+1})$. Associated with each state, there is a binary symmetric channel with crossover probability e_k. With the assumption of discrete channel structure, modulation and demodulation are considered as an inherent part of

the channel. Given a specific digital modulation scheme, the average error probability is a function of the received SNR. The crossover probability e_k for each state can be related to the received SNR thresholds and the Bit Error Rate (BER) depends on the adopted modulation scheme (BPSK, DQPSK, CCK, etc.). For example, for the BPSK the error probability as a function of the received SNR can be written as:

$$e_m(a)=1-F\left(\sqrt{2a}\right), \tag{1.33}$$

where:

$$F(\alpha)=1-Q(\alpha)=\int_{-\infty}^{\alpha}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx. \tag{1.34}$$

With the pdf of the received SNR as in (1.30), the steady state probability and the crossover probability for each state are:

$$p_k=\int_{A_k}^{A_{k+1}}\frac{1}{\rho}e^{-\frac{a}{\rho}}da=e^{-\frac{A_k}{\rho}}-e^{-\frac{A_{k+1}}{\rho}}, \tag{1.35}$$

$$e_k=\frac{\left[\int_{A_k}^{A_{k+1}}\frac{1}{\rho}e^{-\frac{a}{\rho}}P_e(a)da\right]}{\int_{A_k}^{A_{k+1}}\frac{1}{\rho}e^{-\frac{a}{\rho}}da}=\left[\int_{A_k}^{A_{k+1}}\frac{1}{\rho}e^{-\frac{a}{\rho}}P_e(a)da\right]/\,p_k, \tag{1.36}$$

where $P_e(a)$ represents the BER for the chosen modulation scheme. In the specific cases of BPSK and CCK ([5], [11], [21]) we have:

$$P_{e(BPSK)}(x)=Q\left(\sqrt{2x}\right), \qquad P_{e(CCK)}(x)=12Q\left(2\sqrt{x}\right), \tag{1.37}$$

with:

$$Q(z)=\frac{1}{\sqrt{2\pi}}\int_{z}^{+\infty}e^{-\frac{\lambda^2}{2}}d\lambda \qquad and \qquad F(z)=1-Q(z)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}e^{-\frac{\lambda^2}{2}}d\lambda. \tag{1.38}$$

The following expressions for the cross-over probabilities of the state s_k are obtained:

$$e_{kBPSK}=\frac{\gamma_k-\gamma_{k+1}}{p_k}, \qquad \gamma_k=e^{\frac{-Ak}{\rho}}\left[1-F\left(\sqrt{2A_k}\right)\right]+\sqrt{\frac{\rho}{\rho+1}}F\left(\sqrt{\frac{2Ak(\rho+1)}{\rho}}\right) \tag{1.39}$$

$$e_{kCCK}=12\cdot\frac{\gamma_k-\gamma_{k+1}}{p_k}, \qquad \gamma_k=e^{\frac{-Ak}{\rho}}\left[1-F\left(\sqrt{4A_k}\right)\right]+\sqrt{\frac{2\rho}{2\rho+1}}F\left(\sqrt{\frac{2Ak(2\rho+1)}{\rho}}\right) \tag{1.40}$$

To calculate the transition probabilities $t_{j,k}$ defined in eq. (1.18), we make the following assumptions. We first assume that the Rayleigh fading channel is slow enough that the received SNR remains at a certain level for the time duration of a channel symbol. Furthermore, the channel states associated with consecutive symbols are assumed to be neighbouring states. In other words, each state can have no more than three outgoing and incoming transitions as illustrated in figure 1.7.

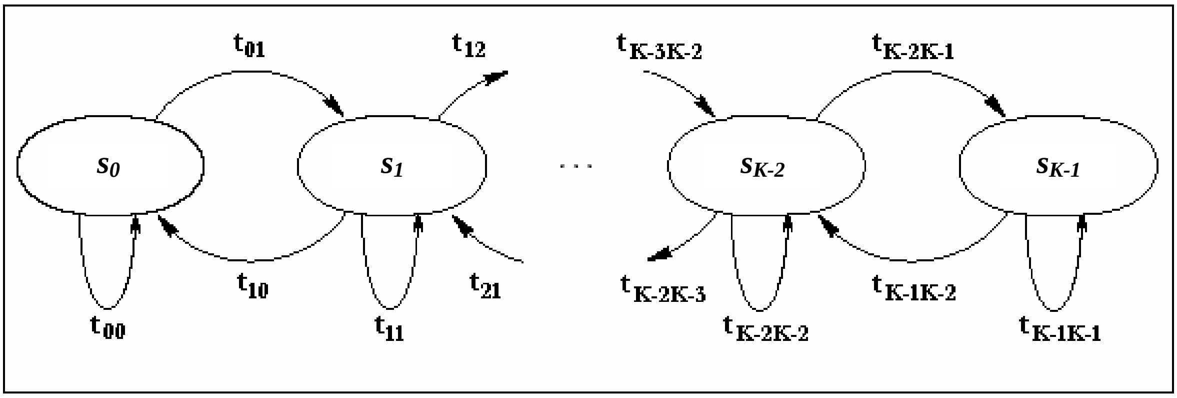


Figure 1.7. K-state noisy channel with Markov transitions modeling a Rayleigh fading channel.

That is:

$$t_{j,k} = 0, \quad \forall |j-k| > 1. \tag{1.41}$$

Now, consider a communication system with a transmission rate of R_t symbols per second. There are, on the average:

$$R_t^{(k)} = R_t \times p_k \tag{1.42}$$

symbols per second transmitted during which the channel is in state s_k. Due to the slow fading assumption, we can conclude that the level crossing rate in eq. (1.32) at A_k and/or A_{k+1} is much smaller than the value $R_t^{(k)}$. The transition probability $t_{k,k+1}$ can then be approximated by the ratio of the expected level crossing at A_{k+1} divided by the average symbols per second the SNR falls in the interval associated with state s_k. Similarly, the transition probability $t_{k,k-1}$ can be approximated by the ratio of the expected level crossing rate at A_k divided by the average symbols per second the SNR falls in the interval associated with state s_k. Specifically, let N_k, $k=1,2,3,...,K-1$, be the expected number of times per second the received SNR passes downward across the threshold A_k. Then, from eq. (1.32) we have:

$$N_k = \sqrt{\frac{2\pi A_k}{\rho}} f_m e^{-\frac{A_k}{\rho}}. \tag{1.43}$$

The Markov transition probabilities are then approximated by:

$$t_{k,k+1} \cong \frac{N_{k+1}}{R_t^{(k)}}, \qquad k = 0,1,2,...,K-2 \tag{1.44}$$

$$t_{k,k-1} \cong \frac{N_k}{R_t^{(k)}}, \qquad k = 1,2,3,\ldots,K-1 \tag{1.45}$$

The other probabilities are evaluated by complement of the previous values:

$$t_{0,0} = 1 - t_{0,1}, \qquad t_{K-1,K-1} = 1 - t_{K-1,K-2}, \qquad t_{k,k} = 1 - t_{k,k-1} - t_{k,k+1} \tag{1.46}$$

$$t_{k,k} = 1 - t_{k,k-1} - t_{k,k+1}, \qquad k = 1,2,3,\ldots,K-2 \tag{1.47}$$

At this point the FSMC is completely defined, but it must be underlined that the choice of ***T***, ***p*** and ***e*** must not be arbitrary: all the constraints must be verified and the non-linear relation between the SNR and e_k makes the SNR ranges to be non-uniformly distributed. There must be a certain criterion to follow in order to obtain a FSMC that captures a real physical link evolution, as exposed in next paragraph.

1.4.3 A Signal to Noise Ratio partitioning scheme for tractable performance analysis

Different partitioning criteria have been proposed in the literature, but none of them was targeted to facilitating the analysis of the loss performance over the wireless link. In [19] a very simple partitioning criterion is proposed, by imposing that the steady state probabilities satisfy $p_0=p_1=\ldots=p_{K-1}$. In [22] the authors propose a new scheme taking into account the average fade duration at various SNR levels; in [23] both the number of states of the model and the partition levels are determined: the scheme leads to having equal average time duration for each state as a multiple of the duration of data packets being transmitted over the channel, but this is not true in reality. None of the previously mentioned partitioning schemes take into account the relation between the SNR and the fading phenomena itself; in addition none of them facilitates the packet performance analysis.

In [24] the authors utilize Jake's level-crossing analysis, the distribution of the received SNR and the elegant analytical structure of Mitra's producer-consumer fluid queueing model [25]. As exposed in previous paragraph, typically a FSMC model is constructed by partitioning the range of the received SNR into a set of nonoverlapping intervals. Each interval is represented by a nominal BER, which in turn represents a certain channel quality. None of the previous works was designed to enable tractable analysis of packet-level performance degradations. More specifically, the use of packet buffering at the transmitter side of a wireless link introduces variable queueing delays and occasional packet loss (due to buffer overflow). Hence, a good model should not only reflect the physical characteristics of the channel, but it should also facilitate analytical investigation of its performance.

A key problem in the design of wireless networks is how to efficiently allocate their scarce resources to meet applications' packet loss and delay requirements. Such efficient allocation can be achieved by using the concept of *Effective Bandwidth* (EB). In general, the EB refers to the minimum amount of network resources (in bits per second) that if allocated to a given traffic flow would guarantee a certain level of QoS (typically in terms of packet loss rate). Many previous works (as [26], [27]) derived a closed-form expression for the EB subject to packet loss and delay constraints, but the analysis was conducted with a GE channel model (that provides a coarse approximation of the channel with a high conservative estimate of the EB). For this reason a multi-state (>2) Markov model must be derived. Figure 1.8 illustrates the typical telecommunication system where the FSMC needs to be accurately modeled.

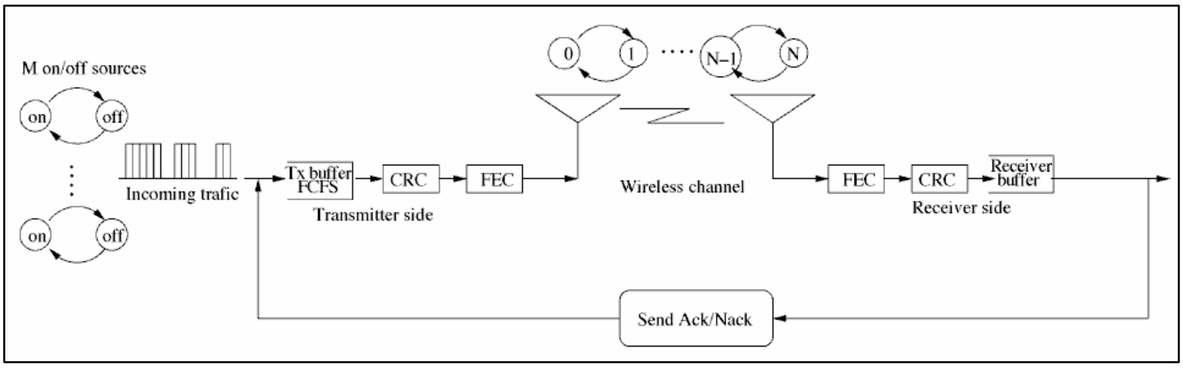


Figure 1.8. Wireless link model.

In the considered system, arriving packets at the transmitter are stored temporarily in a First-In First-Out (FIFO) buffer, which is drained at a rate that depends on the state of the channel at the receiver. We refer to the draining (or service) rate when the received SNR is *r* by *c(r)*. After departing the buffer, a packet undergoes a strong CRC encoding followed by partial FEC that allows for correcting only a fraction of packet errors.

In packet networks, a traffic source is often viewed as an alternating sequence of active and idle periods. During an active period a burst of packets is created. This so called ON-OFF model has the advantage of being able to capture the bursty nature of various types of network traffic. Accordingly, we consider *M* incoming traffic sources, each of which is modeled as a fluid source with exponentially distributed ON and OFF periods. The means of ON and OFF periods are $1/\alpha$ and $1/\beta$ respectively. When the source is active, it transmits at a peak rate σ. The channel is modeled by the *K=N+1* state FSMC model of figure 1.7.

Let p_i be the steady-state probability that the channel is in state s_i, *i=0...N*. The FSMC stays in state *i* for an exponentially distributed time with mean T_i. It is assumed that bit errors within any given state are mutually independent. For a FEC code with a correction capability of τ bits per code block (packet), the probability of an uncorrectable error in a received packet when the channel is in state s_i is given by:

$$P_{c_i} = \sum_{j=\tau+1}^{n} \binom{n}{j} \left(P_c(\hat{r}_i)\right)^j \cdot \left(1 - P_c(\hat{r}_i)\right)^{n-j}, \tag{1.48}$$

where *n* is the number of bits in a code block including the FEC bits, $P_c(r)$ is the BER when the istantaneous SNR is *r* (the form of $P_c(r)$ depends on the underlying modulation scheme), and r_i is the nominal SNR value in state s_i. The packet transmission-retransmission process can be approximated by a Bernoulli process [28]. It is assumed that the transmitter always gets the feedback message from the receiver before the next transmission slot, and a packet is retransmitted persistently until it is successfully received.

Wireless transmission of continuous waveforms in obstacle environments is prone to multipath, which results in randomly varying envelope of the received signal (to see paragraphs 1.4.1 and 1.4.2). It is assumed that the channel changes slowly (slow fading) with respect to symbol transmission rate; furthermore, we assume that transitions between channel states take place only at the end of a packet transmission.

As mentioned before, the FSMC model that represents the time-varying behaviour of the Rayleigh fading channel will be obtained by partitioning the received SNR into *K* intervals, with *K=N+1* and *N+2* thresholds. The steady-state probability that the FSMC is in state s_k is given by eq. (1.35).

Mitra's producers-consumers fluid queueing model [25] facilitates the analytical investigation of communication systems possessing randomly varying statistical properties. According to this model, the fluid produced by *M* producers is supplied to a FIFO buffer that is drained by *N* consumers. Each producer and consumer alternates between independent and exponentially distributed active and idle periods. Let λ^{-1} and μ^{-1} denote the mean of the idle and active periods of a consumer, respectively. When active, a consumer drains fluid from the buffer at a constant rate, which is the same for all consumers. It is easy to see that the number of active consumers fluctuates in time according to the Markov chain in fig. 1.9.

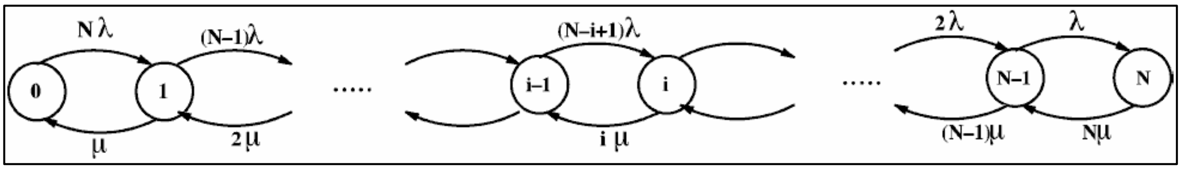


Figure 1.9. Markov Chain in Mitra's model.

In the considered wireless model, *M* corresponds to the number of incoming ON–OFF sources at the transmitting node, while *N* corresponds to the ratio between the nominal service rate when the channel is in the best state (state *N*) and the corresponding service rate when the channel is in the second worst state (state 1) (recall that the nominal service rate in state 0 is zero.) In other words, one can think of the nominal service rate in state 1 as the unit of bandwidth, and of *N* as the number of units of bandwidth that can be offered when the channel is in the best state. To analyze the packet loss and delay performance, we must first partition the wireless channel in a manner that produces the same Markovian structure of figure 1.9. In other words, we match the service rate at the transmitter buffer to the total instantaneous consumption rate in Mitra's model. This requires that we choose the partitioning thresholds such that each state corresponds to a given number of active consumers. Let $\theta(r)=(c(r)/c(\infty))$ be the ratio between the service rate at a received SNR r and its value when the channel is error-free (note that $c(\infty)<c$ due to the FEC overhead). We form a FSMC model based on the requirement that in each SNR interval there exists a point r_i, $r_i \le r_i < r_{i+1}$, that satisfies the relationship:

$$\theta(\hat{r}_i) = \frac{c(\hat{r}_i)}{c(\infty)} = \frac{i}{N} \qquad i = 0,\ 1,\ \ldots,\ N-1. \quad (1.49)$$

Note that in the producer-consumer model, the consumption rate in state i, $i=0,1,\ldots,N$, is i times the consumption rate in state 1. The BER that corresponds to $\hat{r}_i$ is called the *nominal* BER associated with state s_i. We set $\theta(\hat{r}_0)=\theta(r_0)=0$ and $\theta(r_N)=\theta(\infty)$. When $N=1$, $K=N+1=2$ the model defaults to the standard two-state GE model with

$\theta(\hat{r}_0)=0$ and $\theta(r_1)=1$ and no partitioning is needed. Hereafter, we concentrate on the case $N\geq 2$.

It is easy to see that in the producer–consumer model, the steady-state probability distribution is binomial [25]:

$$\pi_i = \binom{N}{i} \cdot \left(\frac{\lambda}{\mu}\right)^i \cdot \frac{1}{\left(1+\frac{\lambda}{\mu}\right)^N}, \quad i=0, \ldots, N. \tag{1.50}$$

For the considered wireless channel to fit the Markov chain in [25], the partitioning must be done so that no more than one state falls in the "good" (BER close to zero) and "bad" (BER close to one) regions of the SNR space, since either scenario will lead to an unrealistic number of states. This can be justified as follows. For $i=1,\ldots,N$, $\hat{r}_i$ must satisfy eq. (1.49). Selecting two states in the "bad" region implies that:

$$\frac{1}{N} = \theta(\hat{r}_1) < \theta(r_2) << 1 \tag{1.51}$$

which results in a very large N. Likewise, selecting two states in the "good" region will lead to:

$$\theta(\hat{r}_{N-1}) \geq \theta(r_{N-1}) = 1 - \varepsilon \tag{1.52}$$

where $0<\varepsilon<<1$. So:

$$\frac{N-1}{N} \geq 1 - \varepsilon \tag{1.53}$$

which also leads to a large N.

To completely specify the underlying Markov chain, we need to determine N, λ/μ and the thresholds $r_1, r_2, \ldots, r_N$. First, by equating eq. (1.35) with eq. (1.50) an expression for N in terms of r_1 and λ/μ is obtained. Using the level-crossing analysis and the structure of the embedded Markov chain, an expression for λ/μ in terms of r_1 and r_N only is obtained. Then, by selecting an appropriate value for r_N, the value of the threshold r_1 can be obtained. After obtaining r_1, the other thresholds can be obtained recursively, by applying the inverse of eq. (1.35):

$$r_{i+1} = -\rho \ln(e^{-\frac{r_i}{\rho}} - \pi_i), \qquad i = 1,2,\ldots, N-2. \tag{1.54}$$

The time spent in any state s_i, is exponentially distributed with mean $T_i=1/q_i$, where q_i represents the total rate out of state s_i. Let the total rate out of state 0 be approximated by the LCR at r_1:

$$N\lambda = L(r_1) = \sqrt{\frac{2\pi r_1}{\rho}} f_m e^{-\frac{r_1}{\rho}}. \tag{1.55}$$

The same treatment can be made for the total rate out of state N:

$$N\mu = L(r_N) = \sqrt{\frac{2\pi r_N}{\rho}} f_m e^{-\frac{r_N}{\rho}}. \tag{1.56}$$

Dividing eq. (1.55) by eq. (1.56) the expression of λ/μ is obtained. After many mathematical manipulations (for details to see [24]) the expression of *N* is obtained:

$$N = \frac{-\ln\left(1-e^{-\frac{r_1}{\rho}}\right)}{\ln\left(1+\sqrt{\frac{r1}{rN}}\cdot\frac{e^{-\frac{r_1}{\rho}}}{e^{-\frac{r_N}{\rho}}}\right)}, \quad N \geq 2 \tag{1.57}$$

with the relationship:

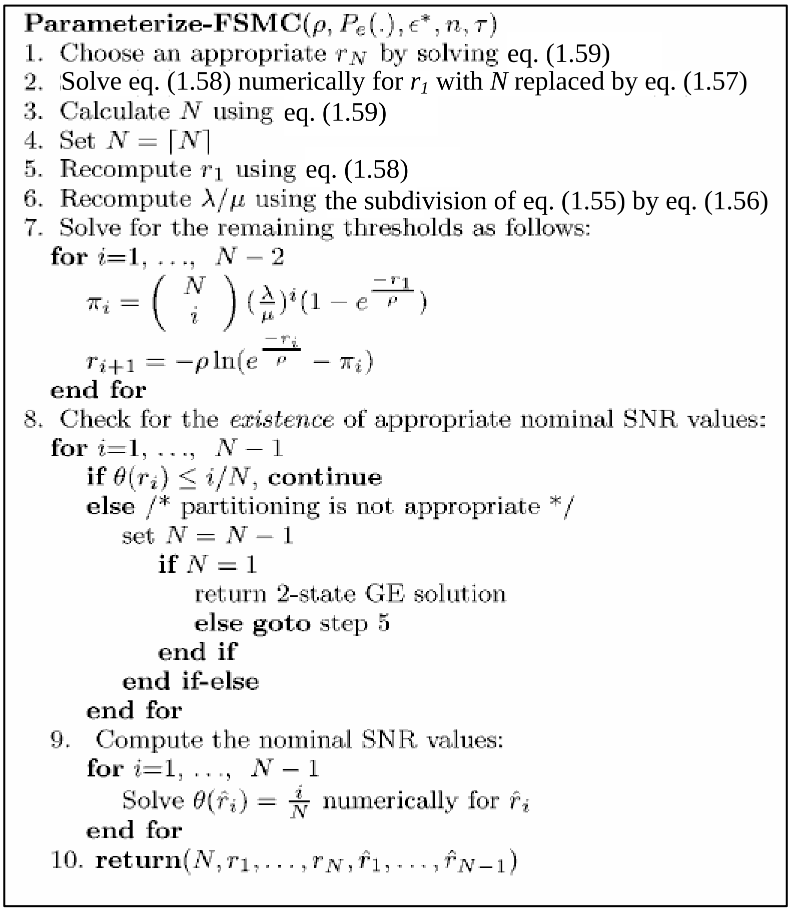
$$f(r_1, r_N, N) = \frac{r_N}{\rho} + N\ln\left(\sqrt{\frac{r_1}{r_N}}\,\frac{e-\frac{r_1}{\rho}}{e-\frac{r_N}{\rho}}\right) + \ln\left(1-e-\frac{r_1}{\rho}\right) = 0, \tag{1.58}$$

In this analysis, *r*r_N has been chosen such that the service rate at state *N* is almost equal to the error-free service rate (this depends on the relationship between SNR and the BER, which in turn is dependent on the modulation scheme). This is done by solving for r_N in:

$$\theta(r_N) = \sum_{j=0}^{\tau} \binom{n}{j} P_c(r_N)^j \cdot (1-P_c(r_N))^{n-j} = 1-\varepsilon^* \tag{1.59}$$

where $0<\varepsilon^*<<1$ is a predefined control parameter. In eq. (1.57) the expression for $P_c(r)$ depends on the deployed modulation scheme.

Using the obtained r_N, we numerically solve eq. (1.58) for r_1. Then from r_1 and r_N we can get *N* by eq. (1.57). The following algorithm, called *Parameterize-FSMC*, sumarizes the main steps of the partition method:



```
Parameterize-FSMC($\rho, P_e(.), \epsilon^*, n, \tau$)
1. Choose an appropriate $r_N$ by solving eq. (1.59)
2. Solve eq. (1.58) numerically for $r_1$ with $N$ replaced by eq. (1.57)
3. Calculate $N$ using eq. (1.59)
4. Set $N = \lceil N \rceil$
5. Recompute $r_1$ using eq. (1.58)
6. Recompute $\lambda/\mu$ using the subdivision of eq. (1.55) by eq. (1.56)
7. Solve for the remaining thresholds as follows:
   for $i$=1, ..., $N-2$
       $\pi_i = \binom{N}{i} (\frac{\lambda}{\mu})^i (1 - e^{\frac{-r_1}{\rho}})$
       $r_{i+1} = -\rho \ln(e^{\frac{-r_i}{\rho}} - \pi_i)$
   end for
8. Check for the existence of appropriate nominal SNR values:
   for $i$=1, ..., $N-1$
       if $\theta(r_i) \leq i/N$, continue
       else /* partitioning is not appropriate */
           set $N = N - 1$
               if $N = 1$
                   return 2-state GE solution
                   else goto step 5
               end if
           end if-else
       end for
   9. Compute the nominal SNR values:
       for $i$=1, ..., $N-1$
           Solve $\theta(\hat{r}_i) = \frac{i}{N}$ numerically for $\hat{r}_i$
       end for
   10. return($N, r_1, \ldots, r_N, \hat{r}_1, \ldots, \hat{r}_{N-1}$)
```

Figure 1.10. The Parameterize FSMC algorithm.

The algorithm in figure 1.10 summarizes the steps that are needed to obtain the parameters of the FSMC model. It takes as input the parameters of the coding scheme *(n, k, τ)*, the modulation-dependent BER function $P_e(.)$, ρ and ε^*. It returns the number of states *N*, the partitioning thresholds and the nominal SNR values. Note that in step 9 of the algorithm, if $\theta(r_i)>i/N$ for some state $i \in \{1,2,....,N\text{-}1\}$, then there is no $\hat{r}_i \in [r_i, r_{i+1})$ for which $\theta(\hat{r}_i)=i/N$ and the partitioning does not fulfill the requirements of the producer-consumer model. If that happens, we decrement the value of *N* and repeat the computations (decrementing *N* icreases the ranges of the various states, which improves the likelihood of finding appropriate nominal SNR values). Note that the algorithm is guaranteed to return a solution, since for *N=1*, the two nominal SNR values, $\hat{r}_0$ and $\hat{r}_1$, are given (the partitioning reduces to the two-state GE model).

Once the FSMC model has been obtained with all the desired parameters, the packet performance analysis must be made, introducing the concept of wireless *Effective Bandwidth* (EB), in terms of packet loss and packet delay. The general concept of an "effective resource" in wireless environments is related to the degradations introduced by a wireless link: if two stations (a source and a destination) are communicating through a wireless link, the amount of resources (e.g. the bandwidth, that is the transmission rate) dedicated by the source to the destination will not be completely received at the receiver side, because of the BER and/or *Packet Error Rate* (PER) characteristics of the wireless link. In other words, the EB can be viewed as the minimum amount of network resources that if allocated to a given traffic flow would guarantee a certain level of QoS. So, the EB amounts are obtained in order to determine the minimum value of *c* (the error-free service rate before accounting for the FEC overhead) that guarantees a desired QoS requirement. The computation of the EB has been performed in [26] and [27] for the case of a GE channel model, but it results in an unnecessarily conservative allocation of network bandwidth.

For the packet loss case, the QoS requirement is given by the couple *(x,p)*, where *x* is the maximum buffer size and *p* is the Packet Loss Ratio (PLR) with $Pr[Q>x]=p$, so the corresponding EB is defined by:

$$c^*_{loss}=min\{\ c : \textit{that results in } Pr[Q>x]=p\ \}, \tag{1.60}$$

while for the case of the delay requirement, the EB is defined as:

$$c^*_{delay}=min\{\ c : \textit{that results in } Pr[delay>l]=\varepsilon_d\ \}, \tag{1.61}$$

where the pair $(l,\ \varepsilon_d)$ represents the delay constraint. Hence, we use c^* to indicate either c^*_{loss} *or* c^*_{delay}, depending on the context. We obtain both quantities in terms of the source, the channel and error control parameters. Following the so-called “dominant eigenvalue” procedure applied to the queueing model in [24], the expressions of c^* are obtained as follows:

$$c^*_{loss}=\frac{\Omega-(\lambda+\mu)^2}{2\eta\xi(\mu-\lambda+\Omega)},\ \ with\ \ \Omega=\frac{M\left(-\alpha-\beta+\sigma\xi+\sqrt{(\alpha+\beta-\sigma\xi)^2+4\beta\sigma\xi}\right)}{N}, \tag{1.62}$$

where $\eta=k/(nN)$, *k* is the number of information bits in a code block, *n* is the total number of bits in a code block, $\xi=-(log p)/x$, μ and λ are related to the producer-consumer model as previously exposed, α and β are related to the ON and OFF periods of the *M* sources and σ is the peak transmission rate of a single source when it is active. For the delay case:

$$c^*_{delay}=\frac{\sigma M(\vartheta-\beta M)\cdot(\vartheta+N(\lambda+\mu))}{N\eta(\vartheta+N\lambda)\cdot(\vartheta-M(\alpha+\beta))}\ \ with\ \ \vartheta=(\log\varepsilon_d)/t. \tag{1.63}$$

In the whole treatment, it is assumed that the transmitter always gets a feedback message from the receiver before the next transmission slot and a packet is retransmitted persistently until it is successfully received. The nominal service rate in state *i*, c_i, can be approximated by the inverse of the mean of the geometrically distributed retransmission process:

$$c_i=c\cdot e\cdot(1-Pc_i), \tag{1.64}$$

where *c* is the error-free service rate, $e=k/n$ is the FEC overhead and P_{ci} is defined as in eq. (1.48).

1.4.4 An overview on channel coding and some interesting results

The analytical analysis of previous paragraphs aims to give a novel and practical partitioning method in order to make possible a performance analysis of the wireless data transmission over fading channels. A modulation scheme needs to be specified in order to obtain the BER curve and the corresponding service ratios. The attention has been focused on the Binary Phase Shift Keying (BPSK) and Complementary Code Keying (CCK) modulation schemes, as defined in the IEEE 802.11a and 802.11b standards respectively. The expressions of the BER for BPSK and CCK are given in eq. (1.37) and their courses are given in figure 1.11.

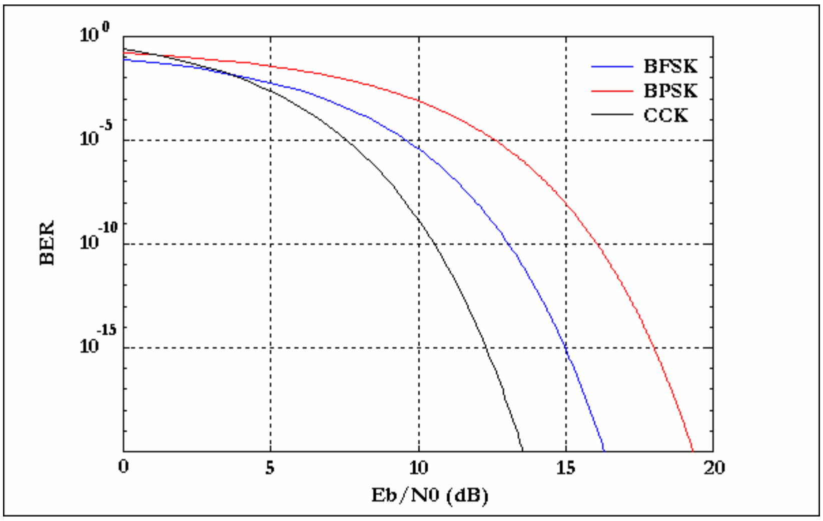


Figure 1.11. Different courses of BER vs. SNR for different modulation schemes in an uncoded channel.

In order to avoid high BER during transmission sessions, the channel coding technique can help the system, introducing redundancy into the transmission itself [5]. The use of error-correcting codes (that is to say channel coding) leads to a reduction of the BER and to an increasing of the coding gain. One way of classifying codes is to distinguish between *block codes* (redundancy is added to blocks of data) and *convolutional codes* (redundancy is added continuously). It has been shown, in the research field, that a Viterbi decoder [29] (the classical solution for convolutional codes) can be used to detect also block codes [30].

The block codes group the source data into blocks and, from the number of the bits in that block, compute a longer codeword that is actually transmitted. The smaller the code-rate (ratio between the number of bits in the original datablock to that of the transmitted block), the higher the redundancy and the higher the probability that errors can be corrected. Source data are parsed into blocks of k symbols and each of these uncoded datablock is then associated with a codeword of length n symbols; the ratio k/n is called the code rate R_c. A particular type of block code is the cyclic block code, where a cyclic shift of a codeword creates another valid codeword. This does not mean that all codewords can be created by shifting a single basis word alone. However, it does imply that all codewords can be obtained from a single codeword by shifting (due to the cyclic property) and addition (due to linearity). It is common to represent cyclic codes by their code polynomials; for example $X(x)=0\cdot x^5+1\cdot x^4+1\cdot x^3+0\cdot x^2+1\cdot x^1+0\cdot x^0$ is used for the bit sequence 011010. For this polynomial representation, a cyclic shift corresponds to a multiplication of $X(x)$ by x, taken $mod(x^n+1)$. There is one codeword polynomial, called generator polynomial, which has minimum degree $n\text{-}k$. It can be shown that this generator polynomial $G(x)$ must be a factor of x^n+1. Multiplying $G(x)$ by x^i, $i=1,...,k\text{-}1$ (no modulo operation necessary here) gives a basis set of codewords from which all other codewords can be generated by linear combination.

Reed-Solomon (RS) codes (an example of cyclic block codes) have the best error-correcting capability of any code of the same length and dimension. Having $n\text{-}k=2t$ parity check bits, they can correct t errors. As mentioned in [5] they are a special case of Bose-Chaudhuri-Hocquenghem (BCH) codes with blocklength $n=q\text{-}1$, where q typically is a power of 2 and an example of decoding algorithm for RS codes can be found in [31].

Now numerical results obtained based on the previously presented analysis are shown. The considered modulation schemes are BPSK and CCK, as in the IEEE 802.11 standards. In our studies we used $n=424$, $M=1$, with RS code for error correction, with f_m=50Hz. When not specified $t=5$, $k=414$ and $\varepsilon^*=1e^{-10}$ (from the relation $k=n-2t$). First of all the Parameterize-FSMC algorithm of figure 1.10 has been applied with different input values of ρ (dB), with the expressions of P_e as in eq. (1.37). This kind of study is necessary, in order to obtain some indications about the channel behaviour under different noise and average SNR conditions.

ρ=2		ρ=3		ρ=4		ρ=5		ρ=6	
BPSK N=7	CCK N=6	BPSK N=4	CCK N=4	BPSK N=3	CCK N=3	BPSK N=3	CCK N=3	BPSK N=3	CCK N=3
0	0	0	0	0	0	0	0	0	0
0.01078	0.01705	0.11658	0.14894	0.39567	0.47502	0.82993	0.96605	1.42993	1.63978
0.06743	0.12788	0.78392	0.84173	2.2262	2.37764	3.41748	3.67925	4.94178	5.36157
0.30179	0.53474	2.70872	2.8312	6.95965	6.46132	6.95965	6.46132	6.95965	6.46132
0.91995	1.51908	6.95965	6.46132	∞	∞	∞	∞	∞	∞
2.16993	3.43468	∞	∞	-	-	-	-	-	-
4.37788	6.46132	-	-	-	-	-	-	-	-
6.95965	∞	-	-	-	-	-	-	-	-
∞	-	-	-	-	-	-	-	-	-

Table 1.12. Partitioning thresholds for different average SNR values ρ.

Table 1.12 resumes the obtained thresholds (in the coloumns) for different values of the average SNR ρ. Recall that r_0=0 and r_{N+1}=∞ for every input parameter (r_0 and r_{N+1} assume their values by definition).

ρ=2		ρ=3		ρ=4		ρ=5		ρ=6	
BPSK N=7	CCK N=6	BPSK N=4	CCK N=4	BPSK N=3	CCK N=3	BPSK N=3	CCK N=3	BPSK N=3	CCK N=3
.0031	.006894	.03785	.03943	.09009	.09253	.09703	.09915	.102151	.1042
.02777	.053449	.19184	.196216	.332633	.33614	.34238	.34521	.349113	.3516
.10691	.172671	.36465	.366177	.40935	.40704	.402705	.4006	.397711	.3958
.22864	.29751	.30806	.303714	.16792	.1643	.15789	.15499	.151024	.1484
.293384	.28834	.09759	.094465	-	-	-	-	-	-
.22588	.14904	-	-	-	-	-	-	-	-
.096613	.03201	-	-	-	-	-	-	-	-
.01771	-	-	-	-	-	-	-	-	-

Table 1.13. Steady state probabilities for different average SNR values ρ.

Table 1.13 resumes the obtained steady state probabilities π_i of the $N+1$ states for different values of the average SNR ρ. The constraint of eq. (1.25) (the sum of the elements of each coloumn must be 1) is not perfectly verified due to the approximations on the last decimal digits, necessary to fill the table while respecting the page margins.

As example two transition probabilities matrices are shown below for ρ=5dB:

$$T_{BPSK} = \begin{pmatrix} 0.9961 & 0.0039 & 0 & 0 \\ 0.0011 & 0.9976 & 0.0013 & 0 \\ 0 & 0.0011 & 0.9981 & 0.0008 \\ 0 & 0 & 0.0020 & 0.9980 \end{pmatrix} \qquad T_{CCK} = \begin{pmatrix} 0.9968 & 0.0032 & 0 & 0 \\ 0.0009 & 0.9979 & 0.0012 & 0 \\ 0 & 0.0010 & 0.9984 & 0.0006 \\ 0 & 0 & 0.0015 & 0.9985 \end{pmatrix}$$

Figure 1.14. Transition probabilities matrices for BPSK and CCK modulations.

Before starting the analysis of the EB, some figures about the Parameterize-FSMC algorithm are shown below.

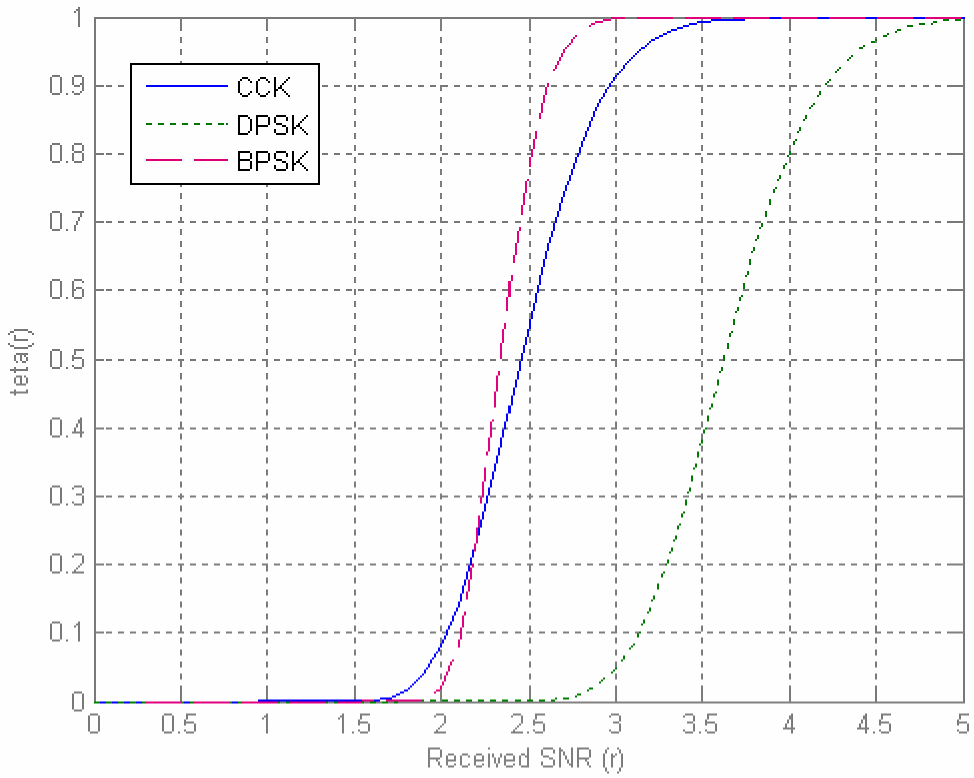


Figure 1.15. Service ratio $\theta(r)$ versus r, for different modulation schemes.

Figure 1.15 shows $\theta(r)$ versus r for different modulation schemes: BPSK, CCK and Differential Phase Shift Keying (DPSK). Because of their different behaviour, the channel partitioning is dependent on the modulation scheme and higher received SNR levels lead to optimum system performances (service rate near to 1).

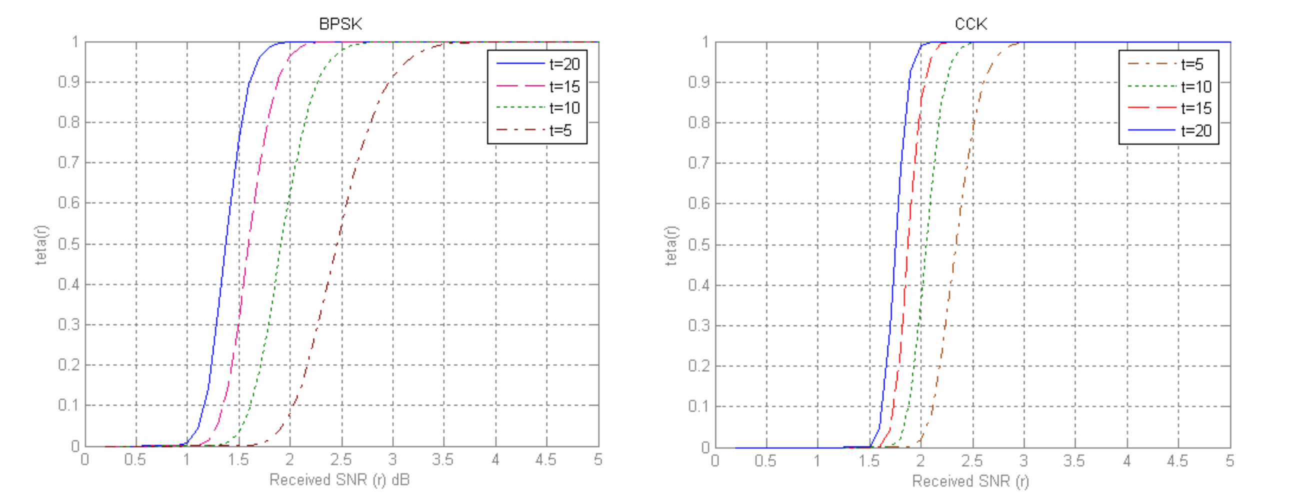


Figure 1.16. Service ratio $\theta(r)$ versus r for different correction capabilities t, BPSK (a) CCK (b).

From figure 1.16 it can be noticed that larger values of t (stronger FEC), make faster the service rate in approaching its asymptotic value, leading to a smaller value of N. As shown later, a larger N implies higher bandwidth allocation efficiency (less EB for a given QoS constraint).

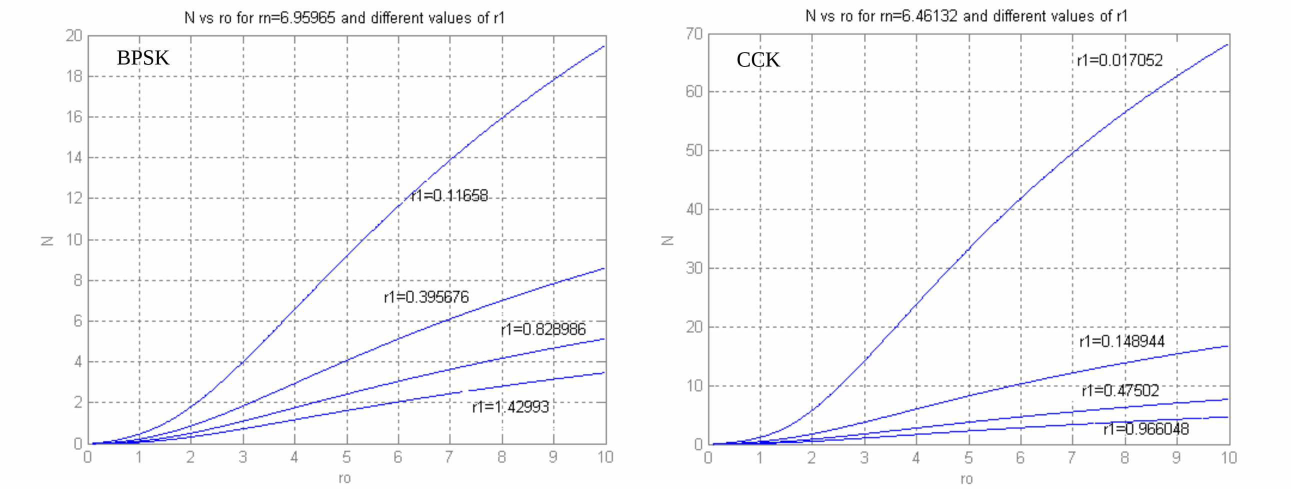


Figure 1.17. Service ratio $\theta(r)$ versus ρ for different values of r_1, BPSK (a) CCK (b).

Figure 1.17 shows the effect of ρ on N. It is observed that as ρ increases, so does N, suggesting the possibility of using ρ as a mean of controlling N (since ρ can be controlled by adjusting the signal power at the transmitter).

Figure 1.18 shows the relationship between r_1, r_N and ρ. When r_N is increased, wider ranges can be chosen during the partitioning procedure, so the value of r_1 goes decreasing, as well as when lower values of ρ are introduced in the partitioning procedure.

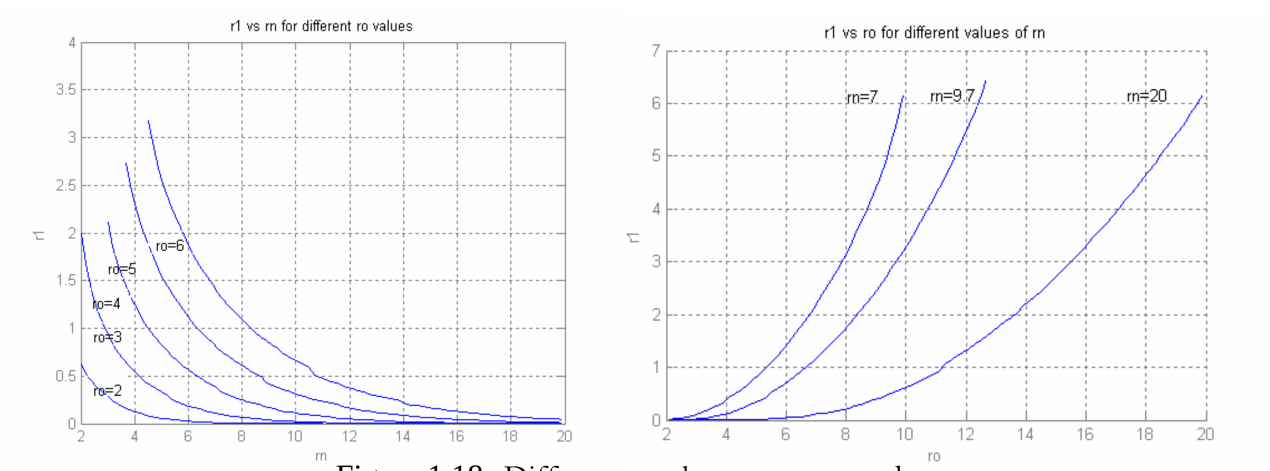


Figure 1.18. Different r_1 values versus r_N and ρ.

For a given r_N, increasing ro results in a larger r_1 and, subsequently, smaller N because the function $\theta(r)$ saturates faster. On the other hand, for a fixed ρ, an increase in r_N results in a decrease in r_1. This, in effect, leads to a larger separation between r_1 and r_N on the SNR axis and, consequently, to a higher N value.

Now the impact of the channel partitioning approach on EB-based allocation subject to either a packet loss or a packet delay constraint is presented. Results are shown for BPSK and CCK modulations. The source peak-rate is set to σ=2594.334 packets/second (about 1.1Mb/s when using 424-bit packets), with α^{-1}=0.02304s and β^{-1}=0.2304s.

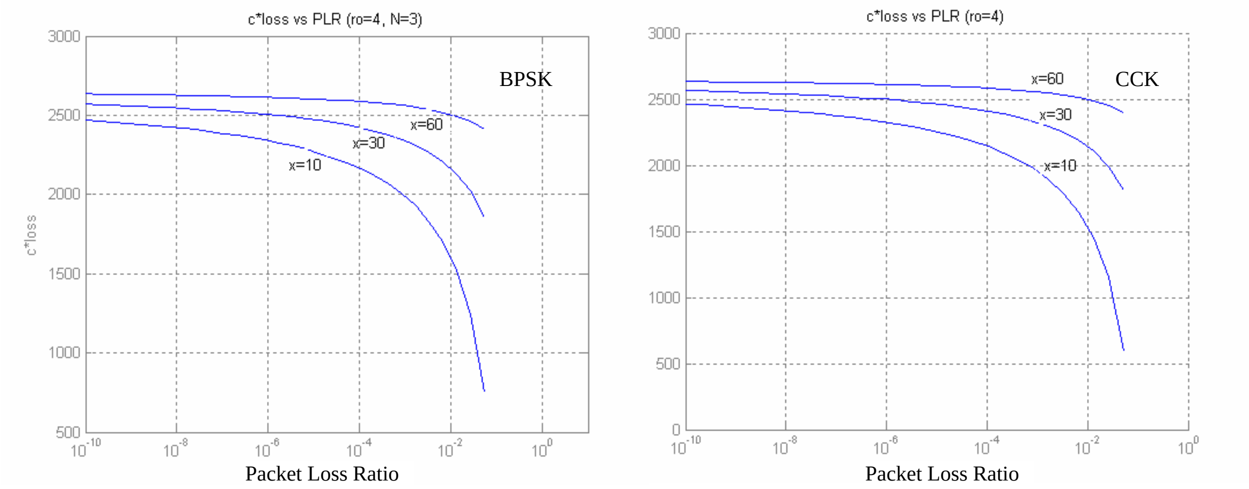


Figure 1.19. EB in terms of c^*_{loss} vs PLR.

Figure 1.19 depicts the EB as a function of the PLR constraint p using different buffer sizes (x) with N=3 (4-state channel model). The curves show that even with a small buffer, typical PLR requirements (10^{-6} or 10^{-3}) can be guaranteed using an amount of bandwidth that is less than the source peak-rate. The significance of the EB analysis is that it allows the operator to decide before-hand the amount of resources (buffer and bandwidth) needed to provide certain QoS guarantees. A reduction in the per-connection allocated bandwidth translates into an increase in the network capacity (measured in the number of concurrently active mobile users). The CCK modulation has a slightly worse performance for small buffer sizes and high PLR if compared with the BPSK modulation scheme.

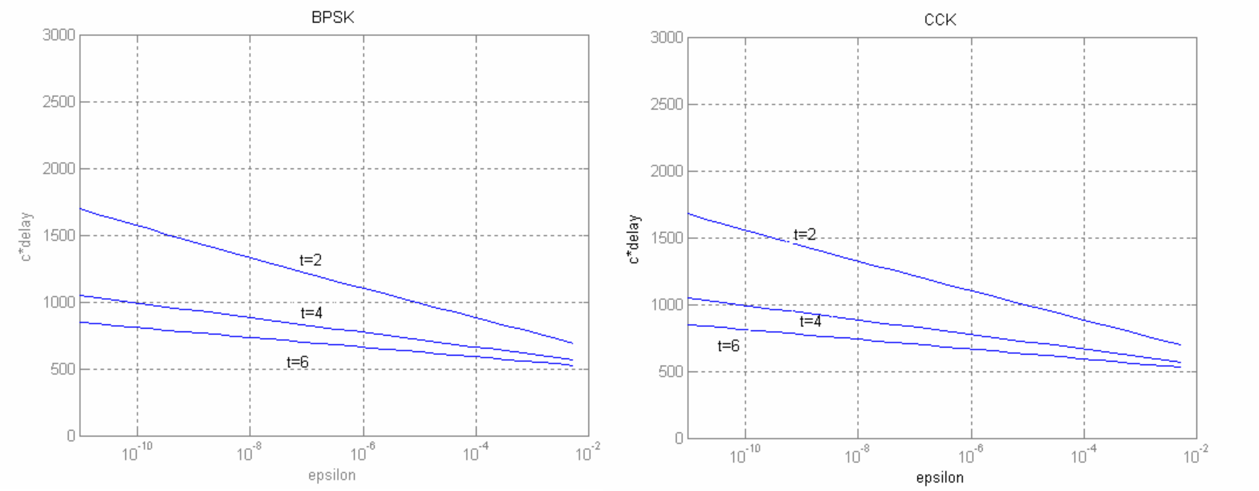


Figure 1.20. EB in terms of c^{*}_{delay} vs ε_d.

Figure 1.20 depicts a similar behaviour for the delay case. In fact, the reduction in the required bandwidth is even more pronounced in this case. There are no evident differences between the two modulation schemes.

1.5 Conclusions on Chapter 1

In this first chapter of the PhD thesis an overview of the main problems affecting WLANs environment has been made. A little history panorama has been given, in order to know how cellular systems evolved until our days. The main attention has regarded the modeling of the wireless link between transmitter and receiver; the first undsired effect in a WLAN environment is the fading, so the different proposed algorithms for resource management must deal with this propagation phenomenon. As it will be shown in next chapters, a good resource management algorithm must take into account channel conditions in order to make an effective bandwidth allocation. For these reasons channel modeling has a heavy impact in wireless communications systems. A good and usable Markov modeling for a WLAN channel has been illustrated and a new approach for partitioning the received SNR range that enable tractable analysis of the packet loss and delay performance over a time-varying wireless channel has been presented. This approach is based on adapting a multistate embedded Markov channel model to Mitra's producer-consumer fluid model, which has known queueing performance. Our analysis exploited several properties of a slowly varying wireless channel, including its LCR and the Rayleigh distribution of the signal envelope. The EB subject to packet loss and delay constraints has been investigated. Besides capturing channel fluctuations, the analysis also accomodates the inherent burstiness in the traffic through the use of appropriate fluid source models. Numerical examples showed that the allowable number of states (*N*) in the Markovian model depends on the underlying modulation scheme, the average SNR and the separation between r_1 and r_N. The larger the value of *N*, the higher is the channel efficiency (in terms of EB). The provided closed-form expressions for the EB can be used as part of admission control and service provisioning in cellular wireless packet networks.

Chapter 2 – The Quality of Service (QoS) in wireless environments and resources reservation

2.1 Introduction: from Best-Effort to QoS-based networks

The last studies about telecommunications and computer-science have increased the need and the availability of communication devices, such as laptops and palmtops; at the same time, the research in digital wireless communications has made possible the connection of notebooks to Internet wherever they are. In addition, many proposals have been made for real-time applications on Integrated Services Packet Networks (ISPNs), like audio-library, image-browsing, video-conferencing and video-on-demand (these kinds of applications need of some constraints on the packet delivery ratio, packet loss rate and so on). In [32] an architecture for ISPNs is described, with the support of real-time traffic; two main components of the proposed architecture are: the Call Admission Control (CAC) scheme and the reservation protocol. The CAC scheme deals with the admission of a new service request into the network, respecting the negotiated QoS levels of the already admitted flows, aiming to high system utilization.

For the Internet real-time traffic the ReSerVation Protocol (RSVP) is used as a resource reservation protocol. Users' mobility has a heavy impact on the QoS parameters and the existing architecture for real-time services with motionless hosts becomes inadequate for QoS management. A new architecture, able to operate with users' mobility, is needed. So, in this chapter the migration from Best-Effort to QoS-based services is investigated.

2.2 Why Internet Quality of Service (QoS): an overview

Quality of Service (QoS) ([33], [34]) has been one of the principal topics of research and development in packet networks for many years. An overview of Internet QoS is now given. QoS generally describes the assurance of sufficiently low delay and packet loss for certain types of applications or traffic. The requirements can be given by human factors, e.g., bounds on delay for interactive voice communications, or by business needs, e.g., the need to complete a transaction within a given time horizon. QoS can be described qualitatively (relative) or quantitatively (absolute). Relative QoS definitions relate the treatment received by a class of packets to some other class of packets, while absolute definitions provide metrics such as delay or loss, either as bounds or as statistical indications. Examples of absolute bounds are statements such as "no more than 5% of the packets will be dropped" or "no packet will experience a delay of more than 100 ms". A set of such statements, along with guarantees about reliability, are often called a Service Level Agreement (SLA – it can be considered as a service contract between a customer and a service provider). Proportional QoS ([35], [36]) tries to refine and quantify relative QoS. QoS guarantees can be made either over an aggregate of communication associations, or for an individual group of packet delineated in time. The latter is often called a "flow".

Applications differ in their QoS requirements. Most applications are loss-sensitive; while data applications can recover from packet loss via retransmission (losses above 5% generally lead to very poor effective throughput). Data applications such as file transfer are not generally delay-sensitive, although human patience imposes lower throughput bounds on applications such as web browsing. Continuous media applications such as streaming audio and video generally require a fixed bandwidth, although some applications can adapt to changing network conditions (as we will see in next sections).

This diversity of applications makes the current Internet approach of offering the same, "best-effort" service, to all applications inadequate. Internet Service Providers (ISPs) also see service differentiation as a way to obtain higher revenue for their bandwidth. In short, it is likely that at least portions of the Internet will see service differentiation in the near future ([37], [38]). Since best-effort service will continue to be dominant, all Internet QoS mechanisms are layered on top of the existing Internet, rather than replacing it with a new infrastructure. Internet design principles [39] such as connectionless service, robustness and end-to-end principles should serve as guidance for any proposed enhancement to current Internet.

In order to provide Internet QoS, we need to describe the properties of flows and aggregates as well as their service requirements. The token bucket is the most commonly used flow specification, for example in the form of the Traffic Specification

(TSpec). Service requirements can be specified in a variety forms, such as the Request (or Requirement) Specification (RSpec) which includes a service rate (*R*) and a delay slack term (*S*). The specifications of traffic and its desired service can be given on a per-flow basis or in SLA. For more details refer to [40].

Since today's Internet interconnects multiple administrative domains (Autonomous Systems - AS), it is the concatenation of domain-to-domain data forwarding that provides end-to-end QoS delivery. Although there are variety of choices, two major frameworks, Integrated Services (IntServ) and Differentiated Services (DiffServ), have emerged as the principal architectures for providing Internet QoS.

IntServ ([32], [41]) is a per-flow based QoS framework with dynamic resource reservation. Its fundamental philosophy is that routers need to reserve resources in order to provide quantifiable QoS for specific traffic flows. RSVP (Resource Reservation Protocol) [42] serves as a signaling protocol for application to reserve network resources. The IntServ architecture adds two service classes to the existing best-effort model: guaranteed service and controlled load service. Guaranteed service [43] provides an upper bound on end-to-end queuing delay. This service model is aimed to support applications with hard real-time requirements. It also provides an assured level of bandwidth, a firm end-to-end delay bound and no queuing loss for conforming packets of a data flow. In a perfect fluid model, a flow conforming to a token bucket of rate *r* and depth *b* will have its delay bound by b/R provided $R \geq r$. To allow for deviations from this perfect fluid model in the router approximation two error terms, *C* and *D*, are introduced; consequently, the delay bound now becomes ($b/R + C/R + D$). However, with guaranteed service a limit is imposed on the peak rate *p* of the flow, which results in a reduction of the delay bound. In addition, the packetization effect of the flow needs to be taken into account by considering the maximum packet size *M*. These additional factors result in a more precise bound on the end-to-end queuing delay as follows:

$$Q_{delayend2end} = \frac{(b-M)\cdot(p-R)}{R\cdot(p-r)} + \frac{(M+C_{tot})}{R} + D_{tot} \qquad for \quad (p > R \geq r) \tag{2.1}$$

$$Q_{delayend2end} = \frac{(M+C_{tot})}{R} + D_{tot} \qquad for \quad (R \geq p \geq r) \tag{2.2}$$

where C_{tot} and D_{tot} represent the summation of the *C* and *D* error terms, respectively, for each router along the end-to-end data path. In order for a router to invoke guaranteed service for a specific data flow, it needs to be informed of the traffic characteristics, T_{spec}, of the flow along with the reservation characteristics R_{spec}.

Controlled-load service [44] provides a quality of service similar to best-effort service in an underutilized network, with almost no loss and delay. It is aimed to share the aggregate bandwidth among multiple traffic streams in a controlled way under overload condition. By using per-flow resource reservation, IntServ can deliver fine-grained QoS guarantees. However, introducing flow-specific state in the routers represents a fundamental change to the current Internet architecture. Particularly in the Internet backbone, where a hundred thousand flows may be present, this may be difficult to manage, as a router may need to maintain a separate queue for each flow. Unlike guaranteed service, controlled-load service provides no firm quantitative guarantees. A T_{spec} for the flow desiring controlled-load service must be submitted to the router as for the case of guaranteed service, although it is not necessary to include the peak rate parameter. If the flow is accepted for controlled-load service, the router makes a commitment to offer the flow a service equivalent to that seen by a best-effort flow on a lightly loaded network.

Many people in the Internet community believe that IntServ framework is more suitable for intra-domain QoS or for specialized applications such as high-bandwidth flows. IntServ also faces the problem that incremental deployment is only possible for controlled-load service, while ubiquitous deployment is required for guaranteed service, making it difficult to be realized across the network.

To address some of the problems associated with IntServ, Differentiated Services (DiffServ) has been proposed in the scientific community with scalability as the main goal. DiffServ ([45], [46]) is a per-aggregate-class based service discrimination framework using packet tagging [47]. Packet tagging uses bits in the packet header to mark a packet for preferential treatment.

DiffServ has two important design principles, namely pushing complexity to the network boundary and the separation of policy and supporting mechanisms. The network boundary refers to application hosts, leaf (or first-hop) routers and edge routers. Since a network boundary has relative small number of flows, it can perform operations at a fine granularity, such as complex packet classification and traffic conditioning. In contrast, a network core router may have a larger number of flows, it should perform fast and simple operations. The differentiation of network boundary and core routers is vital for the scalability of DiffServ.

The separation of control policy and supporting mechanisms allows these to evolve independently. DiffServ only defines several per-hop packet forwarding behaviors (PHBs) as the basic building blocks for QoS provisioning and leaves the control policy as an issue for further work. The control policy can be changed as needed, but the supporting PHBs should be kept relatively stable. The separation of these two components is a key for the flexibility of DiffServ. A similar example is Internet routing. It has very simple and stable forwarding operations, while the construction of routing tables is complex and may be performed by a variety of different protocols. Currently, DiffServ provides two service models besides best effort. Premium service [48] is a guaranteed peak rate service, which is optimized for very regular traffic patterns and offers small or no queuing delay. This model can provide absolute QoS assurance. One example of using it is to create "virtual leased lines", with the purpose of saving the cost of building and maintaining a separate network. Assured service [49] is based on statistical provisioning. It tags packets as "In" or "Out" according to their service profiles. "In" packets are unlikely to be dropped, while "Out" packets are dropped first if needed. This service provides a relative QoS assurance.

Having outlined the frameworks, the details of Internet QoS mechanisms can be considered along two major axes: data path and control path. Data path mechanisms are the basic building blocks on which Internet QoS is built. They implement the actions that routers need to take on individual packets, in order to enforce different levels of service (packet classification, marking, metering, policing, shaping, queueing and scheduling). Control path mechanisms are concerned with configuration of network nodes with respect to which packets get special treatment and what kind of rules are to be applied to the use of resources. For more details about queueing and scheduling ([50] – [53]) can be helpful.

The main control path mechanism is the Call Admission Control ([54] – [57]) which implements the decision algorithm that a router or host uses to determine whether a new traffic stream can admitted without impacting QoS assurances earlier granted. As each traffic stream needs certain amount of network resources (link bandwidth and router buffer space) for transferring data from source to destination, admission control is used to control the network resource allocation. The goal is to correctly compute the admission region, since an algorithm that unnecessarily denies access to flows that could have been successfully admitted will underutilize network resource; while an algorithm that incorrectly admits too many flows will induce QoS violations. There are three basic approaches for admission control: deterministic, statistic, and measurement-based. The first two use a priori estimation, while the later one is based on the current measurement of some criteria parameters. The deterministic approach uses a worst-case calculation which disallows any QoS violation. It is acceptable for smooth traffic flows, but it is inefficient for bursty flows and leads to a lower resource utilization. Both statistical and measurement-based approaches allow a small probability of occasional QoS violation to achieve high resource utilization.

The last important concept that is related to the control path mechanisms is the Bandwidth Broker (BB): it is a logical resource management entity that allocates intra-domain resources and arranges inter-domain agreements. A bandwidth broker for each domain can be configured with organizational policies and controls the operations of edge routers.

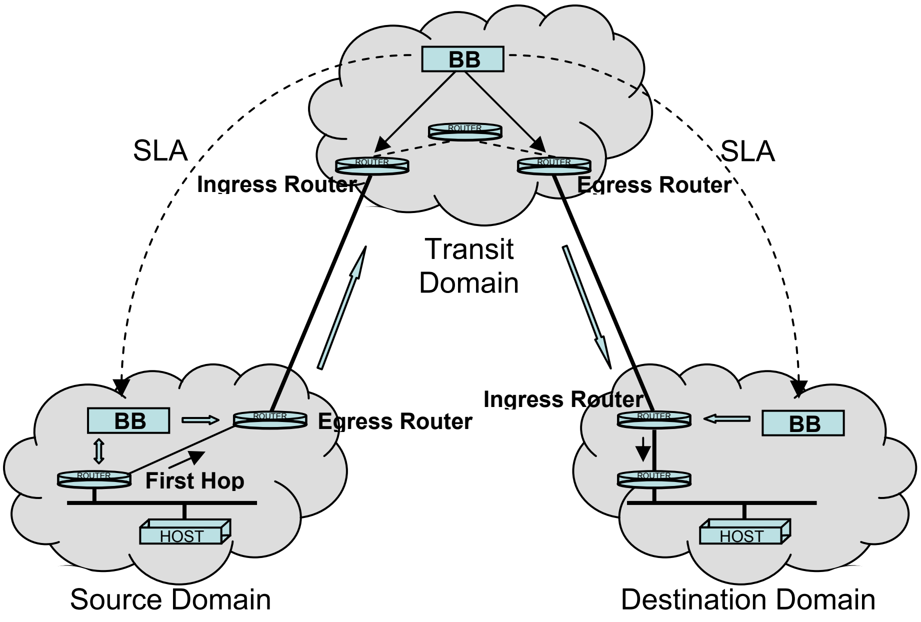


Figure 2.1. Bandwidth Brokers (BBs).

In its inter-domain role, a BB negotiates with its neighbour domains, sets up bilateral agreement with each of them and sends the appropriate configuration parameters to the domain's edge routers (figure 2.1). Bilateral agreement means that a bandwidth broker only needs to coordinate with its adjacent domains. End-to-end QoS is provided by the concatenation of these bilateral agreements across domains, together with adequate intra-domain resource allocation. Within a domain, a bandwidth broker performs resource allocation through admission control. The choice of the intra-domain algorithm is independent of the inter-domain negotiation. The architecture of a bandwidth broker bears some similarity to current Internet routing, in which Border Gateway Protocol v4 (BGP4) serves as the standard inter-domain router protocol, many choices are available for intra-domain routing, and the concatenation of AS-to-AS (Autonomous Systems) forwarding provides end-to-end data delivery.

It must be underlined that a certain grade of policy is necessary to the network: the policy specifies the regulation of access to network resources and services based on administrative criteria. Policies control which users, applications or hosts should have access to which resources and services and under what conditions [58]. Instead of configuring individual network devices, ISPs and corporate administrators would like to regulate the network through policy infrastructure, which provide supports for allowing administrative intentions to be translated into differential packet treatment of traffic flows

2.3 Integrated Services Packet Networks

All the proposed PhD thesis is based on an IntServ architecture able to manage QoS for mobile hosts in a 2D environment; after having argued about the QoS concept in wireless networks it is now necessary to study thoroughly the ISPN framework in order to well describe how the RSVP (sctrictly related to the ISPNs) can be used when a QoS guarantee is needed. For sake of brevity, now we only give the main concepts regarding the IntServ acrhitecture

As we have seen in previous paragraphs, the Integrated Services (IS) are dedicated to the supporting of real-time as well as the current non-real-time service of IP. The extension is necessary to meet the growing need for real-time services for a variety of new applications, including teleconferencing, remote seminars, telescience and distributed simulation.

The multicasts of Internet Engineering Task Force (IETF) meetings across the Internet have formed a large-scale experiment in sending digitized voice and video through a packet-switched infrastructure. These highly-visible experiments have depended upon three enabling technologies:

- Many modern workstations now come equipped with built-in multimedia hardware, including audio codecs and video frame-grabbers and the necessary video gear is now inexpensive;
- IP multicasting, which is not yet generally available in commercial routers, is being provided by the MBONE, a temporary "multicast backbone";
- Highly-sophisticated digital audio and video applications have been developed.

These experiments also showed that an important technical element is still missing: real-time applications often do not work well across the Internet because of variable queueing delays and congestion losses. The Internet, as originally conceived, offers only a very simple QoS, point-to-point best-effort data delivery. Before real-time applications such as remote video, multimedia conferencing, visualization and virtual reality can be broadly used, the Internet infrastructure must be modified to support real-time QoS, which provides some control over end-to-end packet delays. This extension must be designed from the beginning for multicasting; simply generalizing from the unicast (point-to-point) case does not work as earlier discussed.

We use the term Integrated Services (IS) for an Internet service model that includes best-effort service, real-time service, and controlled link sharing. The requirements and mechanisms for IS have been the subjects of much discussions and research over the past several years. The IS model proposed in [41] includes two sorts of service targeted towards real-time traffic: guaranteed and predictive service. It integrates these services with controlled link-sharing and it is designed to work well with multicast as well as unicast. Some assumptions are aldo made: resources (e.g., bandwidth) must be explicitly managed in order to meet application requirements (this implies that "resource reservation" and "admission control" are key building blocks of the service); an alternative approach, which we reject, is to attempt to support real-time traffic without any explicit changes to the Internet service model. The essence of real-time service is the requirement for some service guarantees and it is argued that *guarantees cannot be achieved without reservations*: the user must be able to get a service whose quality is sufficiently predictable that the application can operate in an acceptable way over a period of time determined by the user. There is an inescapable requirement for routers to be able to reserve resources, in order to provide special QoS for specific user packet streams, or "flows". This in turn requires flow-specific state in the routers, which represents an important and fundamental change to the Internet model. The Internet architecture was been founded on the concept that all flow-related state should be in the end systems and designing the TCP/IP protocol suite on this concept led to a robustness that is one of the keys to its success. So, an IS extension that includes additional flow state in routers and an explicit setup mechanism is necessary to provide the needed service. A partial solution short of this point would not be a wise investment. The IS extensions preserve the essential robustness and efficiency of the Internet architecture and they allow efficient management of the network resources; these will be important goals even if bandwidth becomes very inexpensive in the future.

In the ensuing discussion, we take back the definition of "flow" given in paragraph 2.2 and making the abstraction as a distinguishable stream of related datagrams that result from a single user activity and requires the same QoS. For example, a flow might consist of one transport connection or one video stream between a given host pair. It is the finest granularity of packet stream distinguishable by the IS. We define a flow to be simplex, i.e., to have a single source but *N* destinations. Thus, an *N*-way teleconference will generally require *N* flows, one originating at each site. In today's Internet, IP forwarding is completely egalitarian; all packets receive the same quality of service and packets are typically forwarded using a strict FIFO queueing discipline. For IS, a router must implement an appropriate QoS for each flow, in accordance with the service model. The router function that creates different qualities of service is called "traffic control". Traffic control in turn is implemented by three components: the packet scheduler, the classifier and admission control. The final component of the implementation framework of [41] is a reservation setup protocol, which is necessary to create and maintain flow-specific state in the endpoint hosts and in routers along the path of a flow (our studies are based on the RSVP, as explained later in this chapter). In order to state its resource requirements, an application must specify the desired QoS using a list of parameters that is called a "flowspec". The flowspec is carried by the reservation setup protocol, passed to admission control for acceptability and ultimately used to parametrize the packet scheduling mechanism. Figure 2.2 shows how these components might fit into an IP router that has been extended to provide integrated services. The router has two broad functional divisions: the forwarding path below the double horizontal line and the background code above the line. The forwarding path of the router is executed for every packet and must therefore be highly optimized. Indeed, in most commercial routers, its implementation involves a hardware assist. The forwarding path is divided into three sections: input driver, internet forwarder and output driver. The internet forwarder interprets the internetworking protocol header appropriate to the protocol suite, e.g., the IP header.

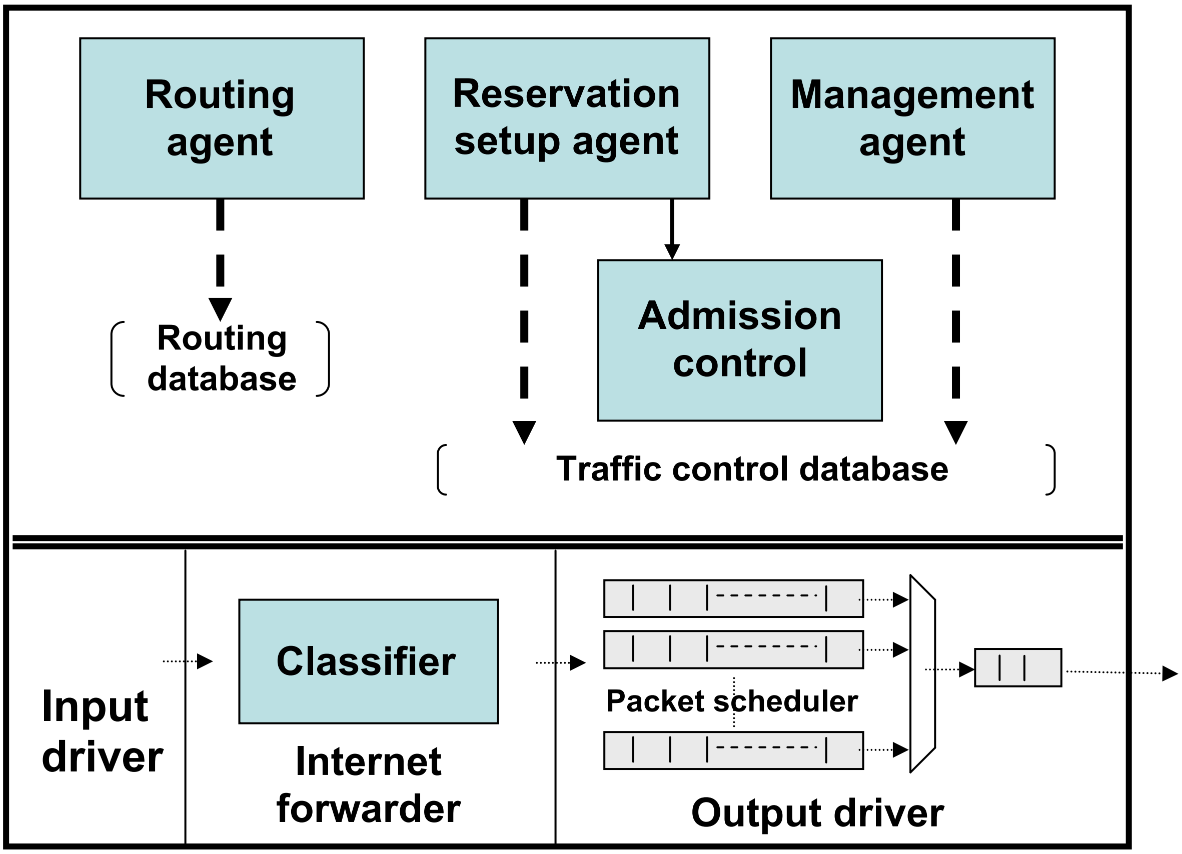


Figure 2.2. Implementation reference model for routers.

For each packet, an internet forwarder executes a suite-dependent classifier and then passes the packet and its class to the appropriate output driver. A classifier must be both general and efficient. For efficiency, a common mechanism should be used for both resource classification and route lookup. The output driver implements the packet scheduler. The background code is simply loaded into router memory and executed by a general-purpose CPU. These background routines create data structures that control the forwarding path. The routing agent implements a particular routing protocol and builds a routing database. The reservation setup agent implements the protocol used to set up resource reservations. If admission control gives the "OK" for a new request, the appropriate changes are made to the classifier and packet scheduler database to implement the desired QoS. Finally, every router supports an agent for network management. This agent must be able to modify the classifier and packet scheduler databases to set up controlled link-sharing and to set admission control policies. The implementation framework for a host is generally similar to that for a router, with the addition of applications. Rather than being forwarded, host data originates and terminates in an application. An application needing a real-time QoS for a flow must somehow invoke a local reservation setup agent. The best way to interface to applications is still to be determined. For example, there might be an explicit API for network resource setup, or the setup might be invoked implicitly as part of the operating system scheduling function. The IP output routine of a host may need no classifier, since the class assignment for a packet can be specified in the local I/O control structure corresponding to the flow.

A service model is embedded within the network service interface invoked by applications to define the set of services they can request. While both the underlying network technology and the overlying suite of applications will evolve, the need for compatibility requires that this service interface remain relatively stable (or, more properly, extensible; we do expect to add new services in the future but we also expect that it will be hard to change existing services). Because of its enduring impact, the service model should not be designed in reference to any specific network artifact but rather should be based on fundamental service requirements.

The per-packet delay is the central quantity about which the network makes quality of service commitments. We make the even more restrictive assumption that the only quantity about which we make quantitative service commitments are bounds on the maximum and minimum delays. The degree to which application performance depends on low delay service varies widely, and we can make several qualitative distinctions between applications based on the degree of their dependence. One class of applications needs the data in each packet by a certain time and, if the data has not arrived by then, the data is essentially worthless; we call these “real-time” applications. Another class of applications will always wait for data to arrive; we call these “elastic” applications. We now consider the delay requirements of these two classes separately.

So far, we have implicitly assumed that all packets within a flow were equally important. However, in many audio and video streams, some packets are more valuable than others. We therefore propose augmenting the service model with a "preemptable" packet service, whereby some of the packets within a flow could be marked as preemptable. When the network was in danger of not meeting some of its quantitative service commitments, it could exercise a certain packet's "preemptability option" and discard the packet (not merely delay it, since that would introduce out-of-order problems). By discarding these preemptable packets, a router can reduce the delays of the not-preempted packets. Furthermore, one can define a class of packets that is not subject to admission control. In the scenario described above where preemptable packets are dropped only when quantitative service commitments are in danger of being violated, the expectation is that preemptable packets will almost always be delivered and thus they must included in the traffic description used in admission control. However, we can extend preemptability to the extreme case of "expendable" packets (the term expendable is used to connote an extreme degree of preemptability), where the expectation is that many of these expendable packets may not be delivered. One can then exclude expendable packets from the traffic description used in admission control; i.e., the packets are not considered part of the flow from the perspective of admission control, since there is no commitment that they will be delivered.

The "reservation model" describes how an application negotiates for a QoS level. The simplest model is that the application asks for a particular QoS and the network either grants it or refuses. Often the situation will be more complex. Many applications will be able to get acceptable service from a range of QoS levels, or more generally, from anywhere within some region of the multi-dimensional space of a flowspec. For example, rather than simply refusing the request, the network might grant a lower resource level and inform the application of what QoS has been actually granted. A more complex example is the "two-pass" reservation model: in this scheme, an "offered" flowspec is propagated along the multicast distribution tree from each sender S_i to all receivers R_j. Each router along the path ecords these values and perhaps adjusts them to reflect available capacity. The receivers get these offers, generate corresponding "requested" flowspecs and propagate them back along the same routes to the senders. At each node, a local reconciliation must be performed between the offered and the requested flowspec to create a reservation and an appropriately modified requested flowspec is passed on. This two-pass scheme allows extensive properties like allowed delay to be distributed across hops in the path. Further work is needed to define the amount of generality, with a corresponding level of complexity, which is required in the reservation model.

The various tools and considerations described until now can be combined to support three main kinds of service:

- *Guaranteed delay bounds*: a theoretical result ([59], [60]) shows that if the router implements a Weighted Fair Queue (WFQ) scheduling [36] discipline and if the nature of the traffic source can be characterized (e.g. if it fits within some bound such as a token bucket) then there will be an absolute upper bound on the network delay of the traffic in question. This simple and very powerful result applies not just to one switch, but to general networks of routers;
- *Link sharing*: the same WFQ scheme can provide controlled link sharing. The service objective here is not to bound delay, but to limit overload shares on a link, while allowing any mix of traffic to proceed if there is spare capacity. This use of WFQ is available in commercial routers today and is used to segregate traffic into classes based on such things as protocol type or application;
- *Predictive real-time services*: this service is actually more subtle than guaranteed service. Its objective is to give a delay bound which is, on the one hand, as low as possible, and on the other hand, stable enough that the receiver can estimate it. The WFQ mechanism leads to a guaranteed bound, but not necessarily a low bound. In fact, mixing traffic into one queue, rather than separating it as in WFQ, leads to lower bounds, so long as the mixed traffic is generally similar (e.g., mixing traffic from multiple video coders makes sense, mixing video and FTP does not).

2.4 Resource ReSerVation Protocol (RSVP)

There are a number of requirements to be met by the design of a reservation setup protocol. It should be fundamentally designed for a multicast environment and it must accommodate heterogeneous service needs. It must give flexible control over the manner in which reservations can be shared along branches of the multicast delivery trees. It should be designed around the elementary action of adding one sender and/or receiver to an existing set or deleting one. It must be robust and scale well to large multicast groups. Finally, it must provide for advance reservation of resources and for the preemption that this implies. The RSVP has been designed to meet these requirements [61].

The Resource Reservation Protocol (RSVP) is a network-control protocol that enables Internet applications to obtain special qualities of service (QoSs) for their data flows. RSVP is not a routing protocol; instead, it works in conjunction with routing protocols and installs the equivalent of dynamic access lists along the routes that routing protocols calculate. RSVP occupies the place of a transport protocol in the OSI model seven-layer protocol stack. RSVP originally was conceived by researchers at the University of Southern California (USC) Information Sciences Institute (ISI) and Xerox Palo Alto Research Center. The IETF is now working toward standardization through an RSVP working group. RSVP operational topics discussed in this chapter include data flows, quality of service, session startup, reservation style, and soft state implementation.

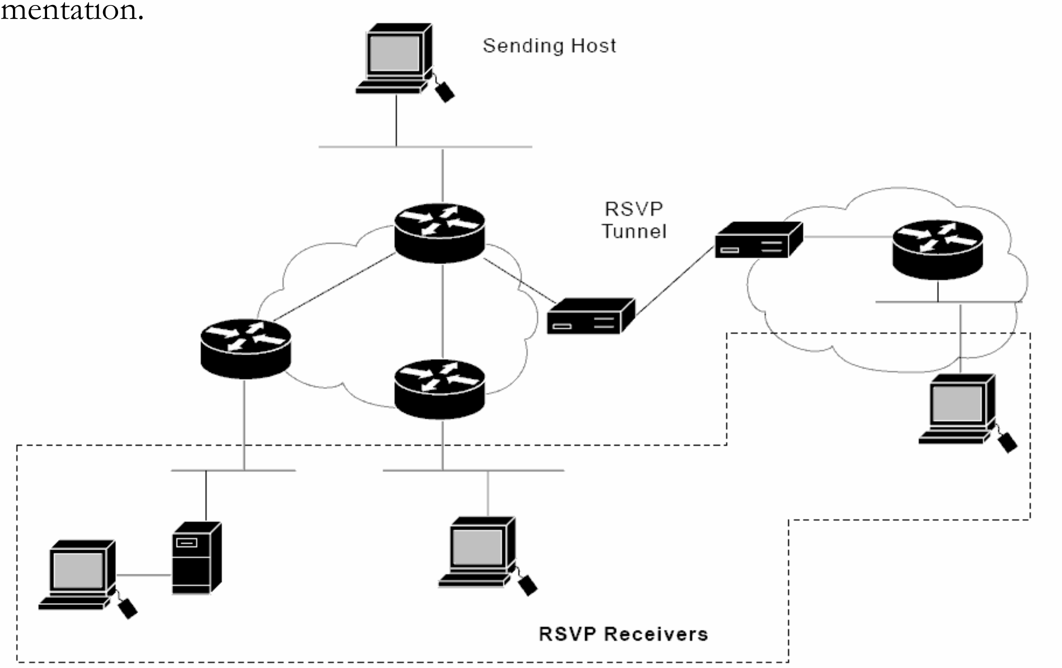


Figure 2.3. A typical RSVP environment.

In RSVP, a data flow is a sequence of messages that have the same source, destination (one or more), and quality of service. QoS requirements are communicated through a network via a flow specification, which is a data structure used by internetwork hosts to request special services from the internetwork. A flow specification often guarantees how the internetwork will handle some of its host traffic. RSVP supports three traffic types: best-effort, rate-sensitive and the delay-sensitive. The type of data-flow service used to support these traffic types depends on QoS implemented. The following sections address these traffic types and associated services. For details see [61].

RSVP data flows are generally characterized by sessions, over which data packets flow. A session is a set of data flows with the same unicast or multicast destination and RSVP treats each session independently. RSVP supports both unicast and multicast sessions (where a session is some number of senders talking to some number of receivers), whereas a flow always originates with a single sender. Data packets in a particular session are directed to the same IP destination address or a generalized destination port. The IP destination address can be the group address for multicast delivery or the unicast address of a single receiver. A generalized destination port can be defined by a UDP/TCP destination port field, an equivalent field in another transport protocol or some application-specific information. RSVP data distribution is handled via either multicasts or unicasts. Multicast traffic involves a copy of each data packet forwarded from a single sender toward multiple destinations. Unicast traffic features a session involving a single receiver. Even if the destination address is unicast, there might be multiple receivers, distinguished by a generalized port. Multiple senders also might exist for a unicast destination, in which case, RSVP can set up reservations for multipoint-to-point transmission. Each RSVP sender and receiver can correspond to a unique Internet host. A single host, however, can contain multiple logical senders and receivers, distinguished by generalized ports.

The RSVP was designed to enable the senders, receivers and routers of communication sessions (either multicast or unicast) to communicate with each other in order to set up the necessary router state to support the services described previously. RSVP identifies a communication session by the combination of destination address, transport-layer protocol type and destination port number. It is important to note that each RSVP operation only applies to packets of a particular session; therefore, every RSVP message must include details of the session to which it applies. In addition, although RSVP is applicable to both unicast and multicast sessions, we concentrate on the more complicated multicast case. Also, we do not discuss the security issues of RSVP or any billing that may be necessary to exert backpressure on the use of reservations. RSVP is not a routing protocol; it is merely used to reserve resources along the existing route set up by whichever underlying routing protocol is in place.

Figure 2.4 shows an example of RSVP for a multicast session involving one sender S1 and three receivers RCV1, RCV2, RCV3. The primary messages used by RSVP are the Path message, which originates from the traffic sender and the Resv message, which originates from the traffic receivers.

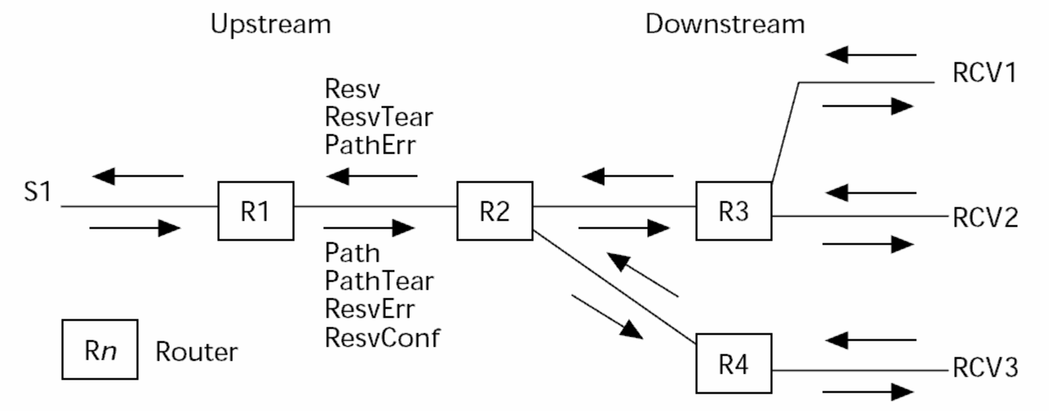


Figure 2.4. Direction of RSVP messages.

The primary roles of the Path message are first to install reverse routing state in each router along the path and second to provide receivers with information about the characteristics of the sender traffic and end-to-end path so that they can make appropriate reservation requests. The primary role of the Resv message is to carry reservation requests to the routers along the distribution tree between receivers and senders. Returning now to figure 2.4, as soon as S1 has data to send it begins periodically forwarding RSVP Path messages to the next hop R1, down the distribution tree. RSVP messages can be transported "raw" within IP datagrams using protocol, although hosts without this raw input/output (I/O) capability may first encapsulate the RSVP messages within a UDP header. For details about RSVP messages and their format refer to [61].

Each intermediate RSVP-capable router along the distribution tree intercepts Path messages and checks them for validity. If an error is detected, the router will drop the

Path message and send a PathErr message upstream to inform the sender who can then take appropriate action.

The router is also responsible for generating Path messages based on the stored path state and forwarding them down the routing tree, making sure that for each outgoing interface the Adspec and Phop objects are updated accordingly [61]. Path messages will be generated and forwarded whenever RSVP detects any changes to stored path state or is informed by the underlying routing protocol of a change in the set of outgoing interfaces in the data forwarding path. Otherwise, a Path message for each specific path state entry is created and forwarded every refresh period timeout interval in order to refresh downstream path state. The refresh period timeout interval is several times smaller than the cleanup timeout interval so that occasional lost Path messages can be tolerated without triggering unnecessary deletion of path state. However, it is still recommended that a minimum network bandwidth be configured for RSVP messages to protect them from congestion losses. Although all path state would eventually timeout in the absence of any refreshes via Path messages, RSVP includes an additional message, PathTear, to expedite the process. PathTear messages travel across the same path as Path messages and are used to explicitly tear down path state. PathTear messages are generated whenever a path state entry is deleted, so a PathTear message generated by a sender will result in deletion of all downstream path state for that sender. It is recommended that senders do this as soon as they leave the communications session. Also, deletion of any path state entry triggers deletion of any dependent reservation state.

RSVP protocol mechanisms provide a general facility for creating and maintaining a distributed reservation state across a mesh of multicast and unicast delivery paths. In order to maintain a reservation state, RSVP tracks a soft state in router and host nodes. The RSVP soft state is created and periodically refreshed by path and reservation-request messages. The state is deleted if no matching refresh messages arrive before the expiration of a cleanup timeout interval. The soft state also can be deleted as the result of an explicit teardown message. RSVP periodically scans the soft state to build and forward path and reservation-request refresh messages to succeeding hops. When a route changes, the next path message initializes the path state on the new route. Future reservation-request messages establish a reservation state. The state on the now-unused segment is timed out (the RSVP specification requires initiation of new reservations through the network two seconds after a topology change). When state changes occur, RSVP propagates those changes from end to end within an RSVP network without delay. If the received state differs from the stored state, the stored state is updated. If the result modifies the refresh messages to be generated, refresh messages are generated and forwarded immediately.

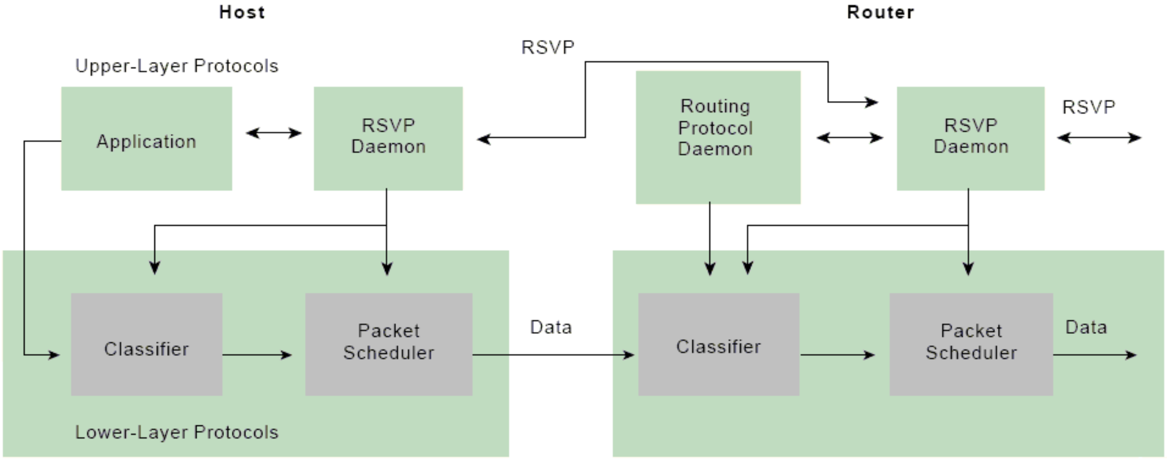


Figure 2.5. The RSVP operational environment reserves resources for unidirectional data flows.

Under RSVP, resources are reserved for simple data streams (that is, unidirectional data flows). Each sender is logically distinct from a receiver, but any application can act as a sender and receiver. Receivers are responsible for requesting resource reservations. Figure 2.5 illustrates this general operational environment, while the subsequent section provides an outline of the specific sequence of events. The RSVP resource-reservation process initiation begins when an RSVP daemon consults the local routing protocol(s) to obtain routes. A host sends IGMP messages to join a multicast group and RSVP messages to reserve resources along the delivery path(s) from that group. Each router that is capable of participating in resource reservation passes incoming data packets to a packet classifier and then queues them as necessary in a packet scheduler. The RSVP packet classifier determines the route and QoS class for each packet. The RSVP scheduler allocates resources for transmission on the particular data link layer medium used by each interface. If the data link layer medium has its own QoS management capability, the packet scheduler is responsible for negotiation with the data-link layer to obtain the QoS requested by RSVP. The scheduler itself allocates packet-transmission capacity on a QoS-passive medium, such as a leased line and also can allocate other system resources, such as CPU time or buffers. A QoS request, typically originating in a receiver host application, is passed to the local RSVP implementation as an RSVP daemon. The RSVP protocol then is used to pass the request to all the nodes (routers and hosts) along the reverse data path(s) to the data source(s). At each node, the RSVP program applies a local decision procedure called admission control to determine whether it can supply the requested QoS. If admission control succeeds, the RSVP program sets the parameters of the packet classifier and scheduler to obtain the desired QoS. If admission control fails at any node, the RSVP program returns an error indication to the application that originated the request.

2.5 Mobile RSVP (MRSVP)

As portable computers become more powerful and the accessibility of a fixed network from a mobile host becomes easier, the number of mobile users will grow and additionally the mobile users will demand the same real-time services available to fixed hosts. Applications such as Internet cellular phone, access to real-time data through the Web (e.g., call center application), require that the network offer quality of service to moving users. Mobility of hosts has a significant impact on the QoS parameters of a real-time application. The existing system architecture for real-time services in a network with fixed hosts is not adequate for supporting mobile hosts and new system architecture is required to handle the effects of mobility (fading, Doppler shift, see chapter 1 for more details). In this paragraph a resource reservation protocol called Mobile RSVP proposed for a network with mobile hosts in [62] is described.

The main QoS parameters for real-time services are packet delay, packet loss rate, delay jitter and throughput. To provide real-time services, a network is designed to provide sufficient guarantees on these QoS parameters. Mobility of a host has a significant impact on these QoS parameters. When a mobile host moves from a location to another one with an active flow, the data flow path changes. As a result the packet delay may change due to changes in the path length and different congestion levels at the routers along the new path. If the new location into which the mobile host moves is overcrowded, the available bandwidth in the new location may not be sufficient to provide the throughput it was receiving at the previous location. In addition, the mobile user may suffer temporary disruption of service during hand-off while the connection is teared down along the old path and it is established along the new path. Therefore, the mobile users may have to adapt to these changes as they move with their flows active. In some extreme cases, some connections to the mobile users may have to be dropped if the minimum QoS requirements of all users cannot be satisfied. To obtain a quality of service which is not affected by mobility, it is necessary to make resource reservation from many locations where the mobile host may visit. We consider a network architecture in which a mobile host can make advance resource reservation along the data flow paths to and from the locations it may visit during the lifetime of the connection. The mobile host can be a sender in a flow, a receiver in a flow or both sender and receiver in the same flow simultaneously. Other than these, the reservation model of RSVP is used. In our reservation model, a mobile host can make advance reservations from a set of locations, called Mobility Specification MSPEC). Ideally, the MSPEC should be the set of locations the mobile host will visit while it participates in the flow. The advance determination of the set of locations to be visited by a mobile host is an important research problem (in chapter 4 a new prediction algorithm is proposed). Also, in many situations, a mobile host can specify its own MSPEC as part of its mobility profile. In any case, we assume that, the mobile host has acquired its MSPEC, either from the network or from its mobility profile, when it initiates a reservation. In our reservation model, the MSPEC of a mobile host can be changed dynamically while the flow is open. In such a case, resources will be reserved at the newly added locations of the MSPEC only if enough resources are available on the data flow paths to/from those locations.

In the network of figure 2.6, link (N1, N2) has a capacity 2B and all other links have a capacity B each. Locations C1, C2 and C3 are in the mobility specification of Mobile Host MH1 and locations C4 and C5 are in the mobility specification of mobile host MH3. In a flow, MH1 is the receiver and MH3 is the sender. MH1 requires a bandwidth B; for MH1 an active reservation is setup to C3 and passive reservations are setup to C1 and C2. For MH3, an active reservation is setup from C5 and a passive reservation is setup from C4. In another flow, Mobile host MH2 requires a weaker QoS guarantee and successfully makes an active reservation from the fixed sender S1 to C1 for a bandwidth B. The considered model supports two types of reservations: active and passive.

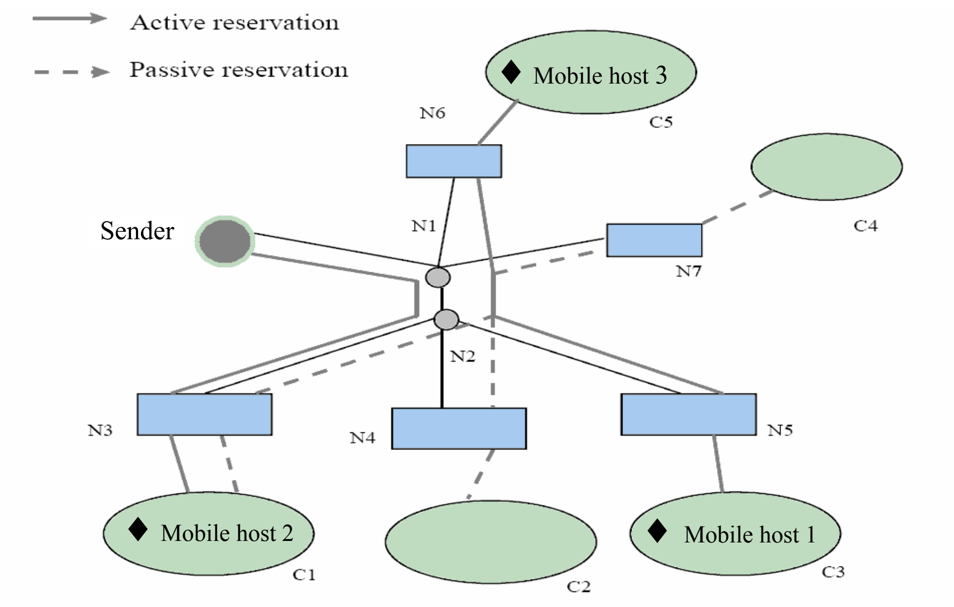


Figure 2.6. Reservation model in mobile environment.

A mobile sender makes an active reservation from its current location and it makes passive reservations from the other locations in its MSPEC. Similarly, a mobile receiver makes an active reservation to its current location and passive reservations to the other locations in its MSPEC. On a link, active and passive reservations for a flow are merged. However, either of the active and passive reservations for the same flow on a link can be removed without affecting the other. To improve the utilization of the links, bandwidth of passive reservations of a flow can be used by other flows requiring weaker QoS guarantees or best effort service. However, when a passive reservation becomes active (i.e. when the flow of the mobile host who made the passive reservation moves into that link), these flows may be affected. So, in the considered model of [62], the resources of passive reservation are multiplexed among the different classes of users. This multiplexing has significant impacts on the network performance and QoS parameters. In an earlier research work [63], we have shown that, if sufficient degree of multiplexing is allowed in the network, the network performance do not degrade significantly when compared to system in which no advance passive reservation is allowed. A unicast packet is delivered to a mobile host by using the Mobile-IP [64] routing protocol. In such a case, resource reservations for a mobile host must be established along the route determined by Mobile-IP. This implies that, when the mobile host is located in a foreign subnet and the unicast packets for the mobile host is delivered via its home agent by IP tunnelling, resource reservations must also be established over the tunnel (provided the routers on the tunnel are RSVP capable).

The reservation mechanism in RSVP is not adequate to support the reservation model of MRSVP described above. This is due to the following reasons:

- RSVP does not have any provision for passive reservation;
- In RSVP, reservation can be initiated from a location, only when the sender or the receiver is present at that location. Thus, in RSVP, a mobile host cannot make an advance reservation from a location where it is not currently present. For a mobile host, Path or Resv messages must originate from the locations where it wants to make advance reservation;
- In RSVP, the sender IP address and port number is used to identify the senders in the FILTER SPEC. A Path message carries the IP address of its origin in the SENDER TEMPLATE. This ensures that the Path message is properly routed to the destination by the routing protocols in which routing decision depends on the source address of the packet. As a consequence, if Path messages originate from several locations in the MSPEC of a mobile sender, a receiver or an intermediate router cannot determine the identity of the mobile host from SENDER TEMPLATE object of the message. As a result, Resv message forwarding for the different reservation styles becomes difficult;

In this paragraph, an overview of the MRSVP protocol is given. Just as Mobile-IP protocol requires home agents and foreign agents to aid in routing, MRSVP requires proxy agents to make reservations along the paths from the locations in the MSPEC of the sender to the locations in the MSPEC of the receiver. The proxy agent at the current location of a mobile host is called the local proxy agent; the proxy agents at the other locations in its MSPEC are called Remote Proxy Agents (RPAs).

In figure 2.7 it is assumed that mobility specification (MSPEC) of the mobile host MH1 is C2, C3 and C4. MH1 sends Receiver_Spec message to the remote proxy agents at nodes N2 and N3. MH2 sends Sender_Spec message to the remote proxy agent at node N6. Sender S and MH2 sends Active Path message; the proxy agent at N6 sends Passive Path message. Active Resv message is sent from the mobile host MH1 via node N4. Proxy agents at nodes N2 and N3 send Passive Resv messages. The remote proxy agents will make passive reservations on behalf of the mobile host. The local proxy agent of a mobile host acts as a normal router for the mobile host and an active reservation is set up from the sender to the mobile host (or from the mobile host to the sender) via its local proxy agent. An important issue is how the mobile host determines who will be the proxy agents.

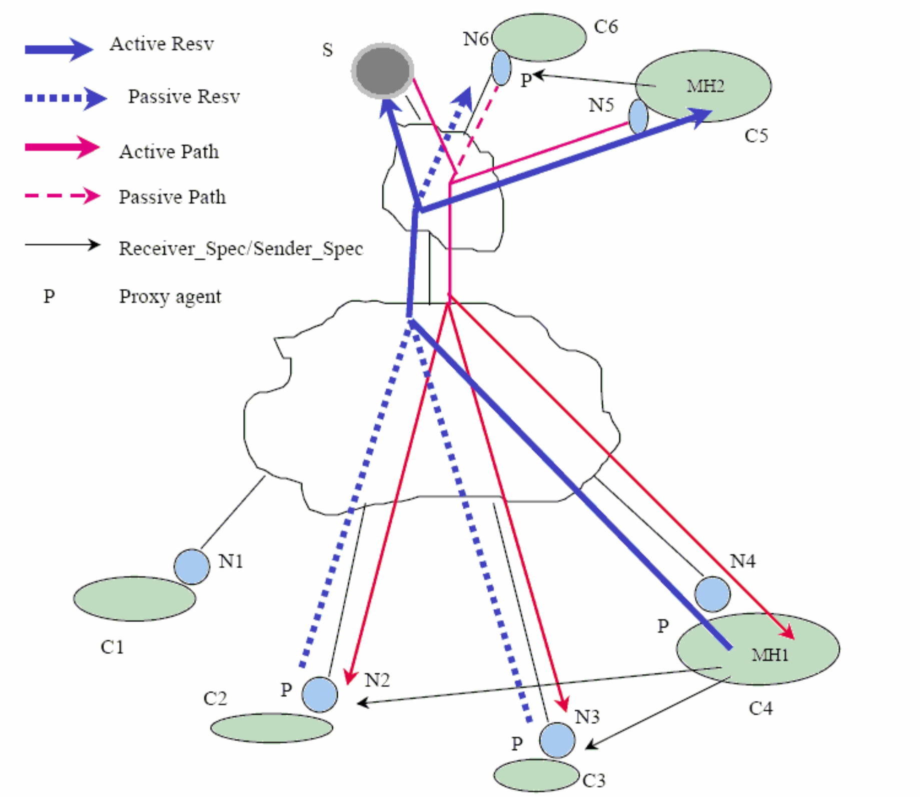


Figure 2.7. Active and Passive reservations with RPAs.

We assume that in the MSPEC of a mobile host, each location is represented by the subnet address of the subnetwork covering that location. Then it can use a proxy discovery protocol (described in the next section) to determine the IP addresses of the proxy agents. After the mobile host knows the IP addresses of its proxy agents, the most important task is to set up the paths of active and passive reservations. If the mobile host is a sender of the flow, the paths of active reservation from the current location of the mobile host and the paths of passive reservations from its proxy agents are determined by the routing mechanism of the network. When the mobile host is a receiver, the paths of active and passive reservations to its current location and the proxy agents depend on the flow destination as follows:

1. The mobile host joins a multicast flow: in this case the mobile host directs the proxy agents to join the multicast group and the data flow paths are set up along the multicast routes;

2. The mobile host initiates a unicast flow: in this case the paths may be set up by unicast routing or by multicast routing.

A sender host periodically sends active Path messages to flow destination. In addition, if the sender is mobile, its proxy agents will send passive Path messages. After the routes of active and passive reservations are set up, the mobile host and the proxy agents will start receiving the Path messages. On receiving a Path message the mobile host will send a Resv message for active reservation. If a proxy agent receives Path messages for a multicast group, for which it is acting as a proxy agent or for a mobile host from which it has received a request for acting as a proxy, it will make a passive reservation on the downstream link to which the mobile host will attach when it arrives in its subnet and then send a Resv message to make a passive reservation. Resv messages for active reservations are converted to Resv messages for passive reservation when they are forwarded towards nodes which contain only proxy agents of mobile senders and no active sender (figure 2.7). In addition to the messages present in RSVP, some additional messages are required in MRSVP. These are:

1. *Join group*: this message is sent by a mobile receiver to its remote proxy agents to request them to join a multicast group. It contains the multicast address of the group to join;

2. *Receiver Spec*: this message is used by a mobile receiver to send the FLOWSPEC and the flow identification (i.e. the SESSION object) to its remote proxy agents;

3. *Sender Spec*: a mobile sender uses this message to send its SENDER TSPEC, ADSPEC and the destination address of a flow to a proxy agent;

4. *Receiver Mspec*: this message is used by a mobile host to send its MSPEC to the appropriate node who sets up the routes of active and passive reservations. It contains the addresses of proxy agents of the locations in the MSPEC of the mobile host;

5. *Sender Mspec*: this message is used by a mobile sender to send its MSPEC to a proxy agent, which sets up the routes of active and passive reservations for the mobile sender.

6. *Forward Mspec*: this message is used by a mobile sender to forward the MSPEC of a mobile receiver to its local proxy agent;
7. *Anchor Spec*: this message is used by a sender anchor to forward the flow-specific information to the mobile sender and its proxy agents;
8. *Terminate*: this message is used by the mobile host to request its remote proxy agents to terminate reservation.

Proxy agents set up passive reservations on behalf of a mobile host. Hence, a mobile host needs to discover these proxy agents. We assume that a mobile host knows the subnet addresses for locations that are in its MSPEC. The mobile host still needs to know the addresses of the proxy agents in those subnets. When the mobile host has a foreign agent care-of address in a foreign subnet, the foreign agent acts as the proxy agent in that subnet. As per IETF Mobile-IP [64], these care-of-addresses may be preallocated or they may be dynamically acquired by the mobile host when it moves into that subnet. However, the protocols for acquiring care-of address by a mobile host in a foreign subnet works only when the mobile host is present in that subnet. These protocols cannot be used by a mobile host to acquire a care-of address remotely in a foreign subnet where it is not currently present. Hence, a mobile host needs to use a mechanism such as Service Location Protocol [65] to acquire care-of addresses remotely. The details of this mechanism are out of the scope of this PhD thesis.

In the following, we describe an alternative protocol for dynamically acquiring foreign agent care-of-address in a foreign subnet where the mobile host is not currently located. The protocol uses two messages, Remote Agent Solicitation and Remote Agent Advertisement. The general procedure is:

1. The mobile host sends a Remote Agent Solicitation message in which the destination address is the subnet directed broadcast address of the foreign subnet;
2. On receiving a Remote Agent Solicitation message, the foreign agent will reply with a Remote Agent Advertisement message containing the care-of address to the mobile host.

The RSVP operations at other nodes in the network (except the mobile nodes and mobility agents) need not be changed. Functionally, there are several issues in the protocol. These are: setting up paths for active and passive reservation, reservation setup, merging reservation messages, hand-off management, soft state maintenance, tearing down of reservation, confirmation and handling of error messages. For details refer to [62].

2.6 Service classes and mobility management in ISPNs

As earlier discussed, recent progress in computing technology and wireless digital communication has made portable computers easily available. This has led to an intensive research in the area of Mobile Computing to provide mobile users access to an inter network. The research, so far, has focused on the problem of maintaining connectivity at the network and transport layer inspite of the mobility of the mobile hosts. As exposed in previous paragraphs, in order to handle real-time services an enhanced architecture has five key components:

1) The nature of service commitments that the network can provide;
2) The service interface parameters passed between the network and the flow end-points; this includes both the characterization of the quality of service the network can provide and the characterization of the behaviour of the end-points pf the flow;
3) The functionality of the network elements, namely scheduling algorithms, required to meet the service commitments;
4) The admission control mechanism by which the service commitments are established at each network element;
5) The reservation protocol to setup the required states in the network elements for providing the required service.

In this paragraph the way of accommodating mobile hosts in this architecture is discussed [66]. In an ISPN, the main QoS parameter is the delay experienced by a packet travelling from a sender to a receiver. It is composed by three main components: a) propagation delay (the propagation time of the packet at the speed of light; this is fixed once the data path is chosen; usually the data-path is composed of multiple hops, connected by switches or routers), b) the delay in transmission at each switch waiting for the entire packet to arrive before it can be transmitted onto the next link (this delay, obviously, depends on the packet size), c) congestion delay (it arises to the statistical sharing of the finite link bandwidth by packets belonging to multiple flows: packets belonging to multiple flows: packets arriving at a switch are buffered in a service queue until the outgoing link is available; it depends on the number of flows using a link, the link capacity and the traffic generation rate of different sources).

To provide real-time services this congestion delay must be bounded or minimized. Two service models can be defined: *guaranteed* service for intolerant applications and *predictive* service for tolerant applications. In a mobile computing network, a geographic area is divided into several cells. Each cell is served by a Base Station (BS) or by an AP (they are called mobility agents). A mobile host maintains connectivity with the fixed network through the base station of the cell in which it is currently located. When a mobile host moves from one location to another, the delivery delay of a packet is affected in two ways. The first factor is that the propagation delay may change in the length of the path from the sender to the receiver. Secondly, the congestion delay at the switches along the new path may be different. There are two broad ways to provide service guarantees to mobile hosts. The first approach is location dependent: the QoS is guaranteed to a specific location, i.e., QoS guarantee is maintained as long as the mobile host stays at the location from where it initiated the session. To obtain such a service the mobile host makes a reservation for a certain QoS from its current location. As soon as it moves to a new location, the QoS guarantee is no longer valid. The application has to renegotiate its desired QoS at the new location. If sufficient resources are not available along the new data flow path, the mobile host may suffer service degradation, which we call hand-off failure. The second approach is location independent: the mobile host receives the same QoS guarantee at all locations (given by a mobility profile, like the MSPEC specified earlier).

A key concept of this PhD thesis is the pre-reservation: to obtain mobility independent QoS, a mobile host must make reservations along the data flow path from the sender to each location it may visit. However this leads to very low utilization of resources because, although data is physically flowing to all locations, data is being used only at the current location of the mobile. In the following of the thesis it will be shown that the utilization of the network resources can be significantly improved if the reserved bandwidth of unused data flows could be used by other flows and, in addition, if the MSPEC profile is obtained from a deep prediction analysis of user mobility. However, the users who are utilizing the unused resources may suffer service degradation when the original reservers start using the resources.

We based our analysis on two service models, derived from those of [32]. Three service classes to which mobile users may subscribe are defined:

1) Mobility Independent Guaranteed service (MIG class): a mobile user admitted to this service class will receive guaranteed service with respect to packet delay bounds as long as its moves are limited to its mobility specification and it is conforming to its traffic characterization. This class is appropriate for the intolerant applications which require absolute bound on packet delay;
2) Mobility Independent Predictive service (MIP class): a mobile user admitted to this class will receive predictive service with respect to packet delay bound as long as its moves are limited to its mobility specifications and it is conforming to its traffic characterization. This class is appropriate for those tolerant applications which require fairly reliable delay bounds in all cells it might visit and does not want to be affected by mobility of hosts;
3) Mobility Dependent Predictive service (MDP class): a mobile user admitted to this service class will receive predictive service with high probability in all cells it may visit during the lifetime of its connection as long as its moves are limited to its mobility specification and it is conforming to its traffic characterization. However it may occasionally fail to get predictive service and experience severe degradation of QoS. This class is appropriate for tolerant applications, which can tolerate the effects of delay violations due to mobility of hosts.

The predictive flows (both in MIP and MDP classes) are categorized into a number of levels with widely separated delay bounds. The delay bounds of the corresponding levelsof MIP and MDP flows are same, i.e. MIP level *j* and MDP level *j* have the same delay bounds. To implement the above service models for mobile hosts in ISPN, it is not enough to reserve resources along the path from the sender to the current location of the mobile host; it is necessary to make reservation along the paths to other locations, where the mobile host may visit. However, it is not necessary to initiate the data flow along each of those paths; data flow is initiated only along the path from the sender to the current location of the mobile host. Thus we define two types of flows:

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

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

One approach to reservation in an ISPN is to use the a priori worst-case traffic characterization of each reserving flows to compute the resource requirements to provide the requested services. However, if the flow traffic is bursty, the average flow rate is significantly less than their a priori worst-case characterization and as result the network utilization will be very low. In next chapters it will be shown as this problem can be overcomed. In addition, when a mobile host with an existing flow in a particular service class moves into a new cell within its mobility specification, the delay bounds of the already existing flows in the new cell may be violated or the utilization target may be exceeded. Since the flows of MIG and MIP service classes cannot tolerate the consequences of mobility, we move flows of MDP classes to service classes with higher delay or convert them to best-effort traffic. Whenever a mobile host leaves a cell or a flow is terminated, the existing flows are given preferences to upgrade to their requested class rather than admitting new flows. When a mobile host with an MDP class flow enters a new cell within its mobility specification which cannot provide its existing QoS, the QoS of the mobile host degrades.

In our work, MIP and MDP service classes are considered. At the end of this chapter, after the description of another extension of the RSVP, a performance analysis is made under a one-dimensional mobility model, only to demonstrate the importance and the effectiveness of the passive reservation policy. The mobility model will be extended in next chapters, where a full and deeper performance analysis of a wirelessLAN network is shown.

2.7 Another extension of RSVP: the Dynamic RSVP (DRSVP)

As discussed in chapter 1, it is known that wireless links are subject to variations in transmission quality due to factors such as interference and fading, which cause changes in transmission quality. If the lower layers do not detect or respond to these changes, the network layer sees an increase in lost or corrupted packets. This makes it difficult to apply network layer QoS mechanisms, which have been designed mainly to deal with congestion loss and network layer queuing effects, rather than packet loss due to link errors. Therefore, the variations in transmission quality are best addressed within the physical or link layers, which can react in several ways: possibilities include dynamic changes in modulation, automatic repeat-request and adaptive forward error correction mechanisms. In general, the techniques employed within the link and physical layer will trade off link throughput in order to maintain low error rate, creating variable bandwidth as seen from the network layer. Another source of variable bandwidth in nomadic networks is node movement, which has several consequences. First, it exacerbates the problem of variable link characteristics, as nodes move in and out of areas of good signal strength. Second, nodes may have to switch to different media as they move in and out of coverage. Again, this illustrates the need for QoS mechanisms to deal with variable bandwidth. Node movement also means that the network topology can change. In the simple case, this consists of the movement of end systems through a fixed network infrastructure. Mobile end systems are "handed off" between fixed access points. However, in a more general case of a mobile ad-hoc network, intermediate systems (routers) also move, resulting in relatively rapid changes in network topology. This makes the general routing problem difficult and QoS-aware routing extremely difficult. It also means that end-to-end bandwidth can change even when individual links remain stable, as topology changes can result in a new route through the network that traverses links with different available resources. As mentioned above, a solution for providing QoS support in nomadic networks must work in the face of topology changes (either the constrained case of mobile end systems or the more general case of a mobile ad-hoc network). A QoS solution for these environments must also be capable of handling variations in bandwidth, both on individual links and end-to-end. In this paragraph, we discuss a resource reservation-based approach for providing QoS support by performing admission control and reserving resources for flows or connections on an end-to-end basis. A resource reservation-based approach is problematic in a variable bandwidth environment: if available resources change after admission control has been performed, the network may not be able to meet commitments for flows that have been successfully admitted. Nevertheless, we believe that a resource-reservation based approach is important for applications that need a per-flow, end-to-end QoS solution. One issue to be considered in the variable network environment is how closely routing and QoS mechanisms should interact. One approach is to have them tightly coupled, in other words, support QoS routing. In principle, given a sufficiently rapid QoS-aware routing algorithm, whenever link conditions or network topology change, the routing algorithm would immediately find new routes through the network with sufficient resources to allow QoS commitments made by the network to be maintained. QoS routing is a challenging problem even in a static network; it is especially challenging in a dynamic one. Another option is to completely decouple QoS and routing. This approach is less difficult than QoS routing and is taken with RSVP in traditional networks, which use "soft state" to reserve and release resources on a given path. Traditional routing protocols are used to route both RSVP and data traffic; when a change in routing occurs, RSVP traffic will follow the new path and reserve resources on the new path, while resources on the old path will "time out".

The previous MRSVP approach tackles the problem of maintaining QoS during hand-offs but assumes that, at least for some users (like those belonging to MIP class), mobility will be predictable [66].

In this paragraph, a new dynamic approach is presented: it is assumed that the network provides service at a signaled QoS allocation point within the range requested in the QoS reservation request. Service is guaranteed at the allocated point, but the network may change this allocation point at any time. As long as the application is capable of adapting its transmission characteristics to stay within its allocated level, it receives QoS support from the network. One reason that we find the notion of QoS ranges expecially attractive is that it facilitates the decoupling of routing and QoS maintenance. If a change in network topology causes a new route to be computed or if throughput changes on one of the links within a route, having a range rather than a single value increases the likelihood that QoS can be maintained at some point within the range. If resources decrease, the current allocation within the range can be decreased, rather than having to fail and tear down the reservation and if resources increase, the current allocation can be increased accordingly. It is important to note that this approach relies on "soft state" to reestablish QoS along the new route when a change occurs. When this happens, the protocol relies on the concept of adaptive QoS to deal with the fact that a different level of resources may be available along the new route. However, the reliance on soft-state mechanisms means that when a route changes, there will be a period during which the traffic receives only best effort service. To demonstrate the feasibility of the dynamic QoS approach discussed above, a distributed network protocol called Dynamic-RSVP (DRSVP) has been proposed in [67]. As the name suggests, this protocol is an extension of RSVP. The authors implemented DRSVP by modifying and extending Information Sciences Institute's (ISI's) implementation of RSVP. This implementation includes the controlled load service model and the key managed resource at each router is interface bandwidth.

The DRSVP protocol was created by making the following extensions and modifications to standard RSVP. An additional flow specification (FLOWSPEC) has been added in Resv messages and an additional traffic specification (SENDER_TSPEC) in Path messages, so that they describe ranges of traffic flows. A "measurement specification" (MSPEC) has been added to the Resv messages, which is used to allow nodes to learn about "downstream" resource bottlenecks and a new reservation notification (ResvNotify) message, which carries a "sender measurement specification" (SENDER_MSPEC) that is used to allow nodes to learn about "upstream" resource bottlenecks has been created. The admission control processing has been changed in order to deal with bandwidth ranges and a bandwidth allocation algorithm that divides up available bandwidth among admitted flows, taking into account the desired range for each flow as well as any upstream or downstream bottlenecks for each flow has been considered.

2.7.1 DRSVP description

Figure 2.8 illustrates a simple network in which node S sends data to node R through intermediate nodes N_1, N_2, N_3, and N_4. The nodes are connected by links, shown in the figure as wide bars, with the width of the bar corresponding to the bandwidth available on the link. The adaptive application running on node S can generate data at rates within the range from s_l to s_h. These values are communicated in Path messages, which flow through the network hop by hop, following the same route as the data messages, to the receiver R. Upon receipt of the Path messages, the receiving application on R requests a reservation for this flow, with QoS range (s_l, s_h).

The request is carried through the network in Resv messages, which travel the reverse of the route followed by the Path messages (assuming bi-directional links). Finally, ResvNotify messages flow through the network from S to R. We will examine the operation of the protocol in detail, using the figure to illustrate how the protocol would operate at node N_2 for this simple example. Each node receives Path and ResvNotify messages from upstream nodes, and Resv messages from downstream nodes (the "upstream" and "downstream" directions are defined relative to the flow of data from S to R, not relative to the flow of protocol messages). In the simple example shown in the figure, there is only a single flow and each node has only one upstream and one downstream interface for this flow.

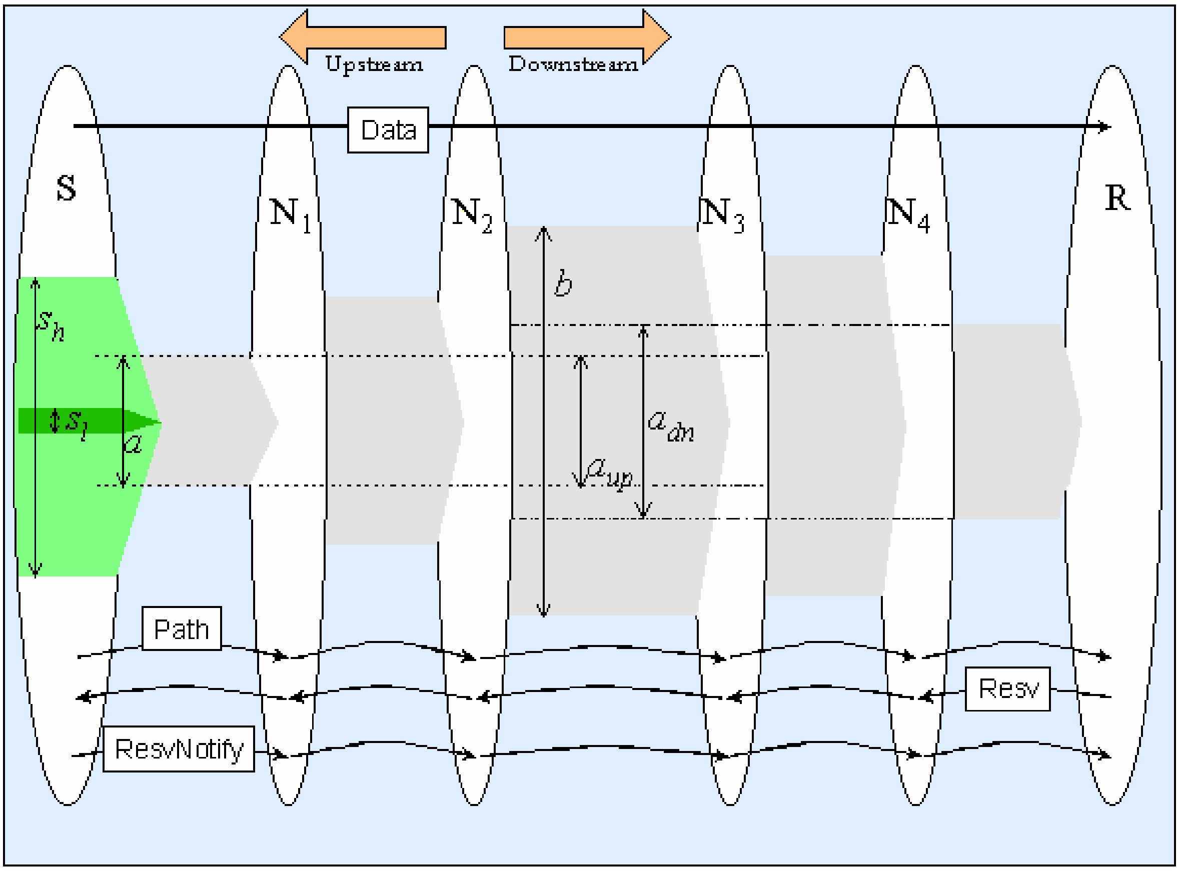


Figure 2.8. A simple application of the DRSVP.

In general, however, there will be multiple flows, and each flow may be multicast, so each flow can have multiple upstream interfaces and multiple downstream interfaces. In this case, a node could possibly receive different values of s_l and s_h in Resv messages from downstream receivers. Each node aggregates and stores the received values. We use $s_h(f)$ and $s_l(f)$ to denote the aggregated value of s_h and s_l for flow f at a given node.

This aggregation is performed by setting $s_h(f)$ to the maximum value of received for flow *f* on any interface at the node and setting $s_l(f)$ to the minimum value received. When a node receives a Resv message on interface *i* for flow *f*, requesting a resource reservation in the range (s_l, s_h), it must determine how much bandwidth within this range it can allocate for the flow on that interface. It does this by executing a bandwidth allocation algorithm that divides up the available bandwidth on interface *i*, denoted $b(i)$, among all the flows that are utilizing this interface. This bandwidth allocation algorithm is a key part of the DRSVP protocol operation. The following discussion describes how we compute the bandwidth allocation for flow on interface *i*, denoted $a(f, i)$. First, if we have enough bandwidth on interface *i* to provide every flow on that interface with the maximum desired bandwidth, the bandwidth allocation algorithm will be simple because there is plenty of bandwidth to spare. Let $F(i)$ denote the set of all flows that have been admitted on downstream interface *i*. Then, the amount of bandwidth needed to satisfy the maximum requested for all flows is:

$$H = \sum_{g \in F(i)} s_h(g). \tag{2.3}$$

If $H \le b(i)$, we simply allocate for *f* on interface *i* the maximum requested bandwidth:

$$a(f,i) = s_h(f). \tag{2.4}$$

This is the case at node N2 in figure 2.8. There is only one flow present and there is sufficient bandwidth available on the downstream interface from N2 to satisfy the maximum requested for the flow. If there is not enough bandwidth for this, i.e. $H>b(i)$, we look to see if there are flows that do not need the maximum requested bandwidth allocated, because they cannot utilize it due to bottlenecks elsewhere in the network. Resv messages received from downstream nodes contain a parameter, denoted m_r, which provides an indication of downstream bottlenecks. Similarly, ResvNotify messages received from upstream nodes contain a parameter m_s that provides an indication of upstream bottlenecks. We use the notation $a_{dn}(g,j)$ to denote the value of m_r that we have received for flow *g* on downstream interface *j*. Similarly, $a_{up}(g,j)$ denotes the value of m_s that we have received for flow *g* on upstream interface *j*. Note that senders such as S in figure 2.8 do not have any upstream nodes. In this case, we simply set a_{up} to s_h, the maximum rate requested for the flow. Similarly, at receivers we set a_{dn} to s_h. A multicast flow may have multiple upstream and downstream interfaces, so we need to aggregate the a_{up} and a_{dn} values for these different interfaces to determine the bottlenecks that may affect this flow elsewhere in the network. If we denote the set of downstream and upstream interfaces for flow *g* as *D(g)* and *U(g)*, respectively, then we perform this aggregation as follows:

$$a_{dn}(g) = \min_{j \in D(g)} \left[a_{dn}(g, j) \right], \qquad a_{up}(g) = \max_{j \in U(g)} \left[a_{up}(g, j) \right] \tag{2.5}$$

We use the minimum when aggregating downstream values because we assume that the sending application will back-off to the rate that can be reliably delivered to all receivers. As a result, there will be no need to reserve more bandwidth than could be delivered to the most constrained receiver. We use the maximum when aggregating upstream values because we want to ensure that we have reserved enough capacity to allow us to deliver the traffic received from the most aggressive transmitter.

Figure 2.8 illustrates the values of a_{up} and a_{dn} at node N2 for the single flow in the example. Using the aggregated downstream and upstream bottleneck parameters, we can now obtain a single estimate of the bottlenecks that affect a flow elsewhere in the network. We refer to this estimate as the "external allocation" and it is computed simply as:

$$a_{ext}(g) = \min[a_{up}(g), a_{dn}(g)]. \tag{2.6}$$

If we have enough bandwidth on interface *i* to provide every flow on that interface with at least as much as its external allocation, then we are not creating a bottleneck for any flow. The amount of bandwidth needed to satisfy the external allocation for all flows is given by:

$$A_{ext} = \sum_{f \in F(i)} a_{ext}(f). \tag{2.7}$$

If $A_{ext} \leq b(i)$, then we can give each flow at least its external allocation. To avoid creating a bottleneck, we only need to reserve at least $a_{ext}(f)$ for flow *f*. However, even though we do not need to reserve more than the external allocation, we do want to advertise the fact that we could reserve more if we needed to. This is crucial to fast convergence of the distributed algorithm when bottlenecks in the network are removed. We need to report the fact that, at this node, the maximum reservation that we could give to each flow is its external allocation plus a share of the "excess" bandwidth available at this node. We assume that the excess bandwidth will be divided up among all the flows in a proportionate manner, so the allocation at this node for flow *f* is given by:

$$a(f,i) = a_{ext}(f) + \beta[s_h(f) - a_{ext}(f)], \tag{2.8}$$

Here, β is a factor that determines how much each flow can be given of its requested range above the external allocation, computed as follows:

$$\beta = \frac{b(i) - A_{ext}}{H - A_{ext}}. \tag{2.9}$$

Finally, if $A_{ext} > b(i)$, we indeed have a bottleneck on interface *i*. In this case, we compute *L*, the bandwidth needed in order to provide each flow with the minimum required bandwidth:

$$L = \sum_{f \in F(i)} s_l(f). \tag{2.10}$$

If $L \leq b(i)$, then we give each flow the minimum of its range:

$$a(f,i) = s_l(f) + \beta[a_{ext}(f) - s_l(f)], \qquad \beta = \frac{b(i) - L}{A_{ext} - L}. \tag{2.11}$$

On the other hand, if $L > b(i)$, there is insufficient capacity to maintain even the minimum. In this case, we reject flow *f*. If it is a new flow, it is considered to have failed admission control. If it is an existing flow, link bandwidth has decreased to the point that we cannot maintain the minimum requested bandwidth for all flows and some existing flow must be torn down. Our implementation simply tears down the first flow for which this condition is detected. A more sophisticated implementation would select a flow to tear down based on some policy, e.g., tearing down flows that are under utilizing their reservation or are somehow considered to be of lower priority than other flows. Having computed the allocation for flow *f*, we know what level of resources to reserve. We also must determine what values to report as m_r and m_s for this flow in Resv and ResvNotify messages that we send upstream and downstream for this flow. To do this, we first take the minimum of the allocations on all of the downstream interfaces for the flow:

$$a(f) = \min_{i \in D(f)} [a(f,i)]. \tag{2.12}$$

Then, the value of m_r we will report upstream is the minimum of the allocation we have made and the allocation made by other nodes downstream of us:

$$m_r = \min[a(f), a_{dn}(f)]. \tag{2.13}$$

Similarly, the value of m_s we will report downstream is the minimum of the allocation we have made and the allocation made by other nodes upstream of us:

$$m_s = \min\lfloor a(f), a_{up}(f) \rfloor \tag{2.14}$$

In the example of figure 2.8, node N2 is not a bottleneck, so it simply forward the values of a_{up} and a_{dn} in its reports upstream/downstream. Observe that N2 knows of the existence and magnitude of the bottlenecks that are present in the network upstream and downstream from it.

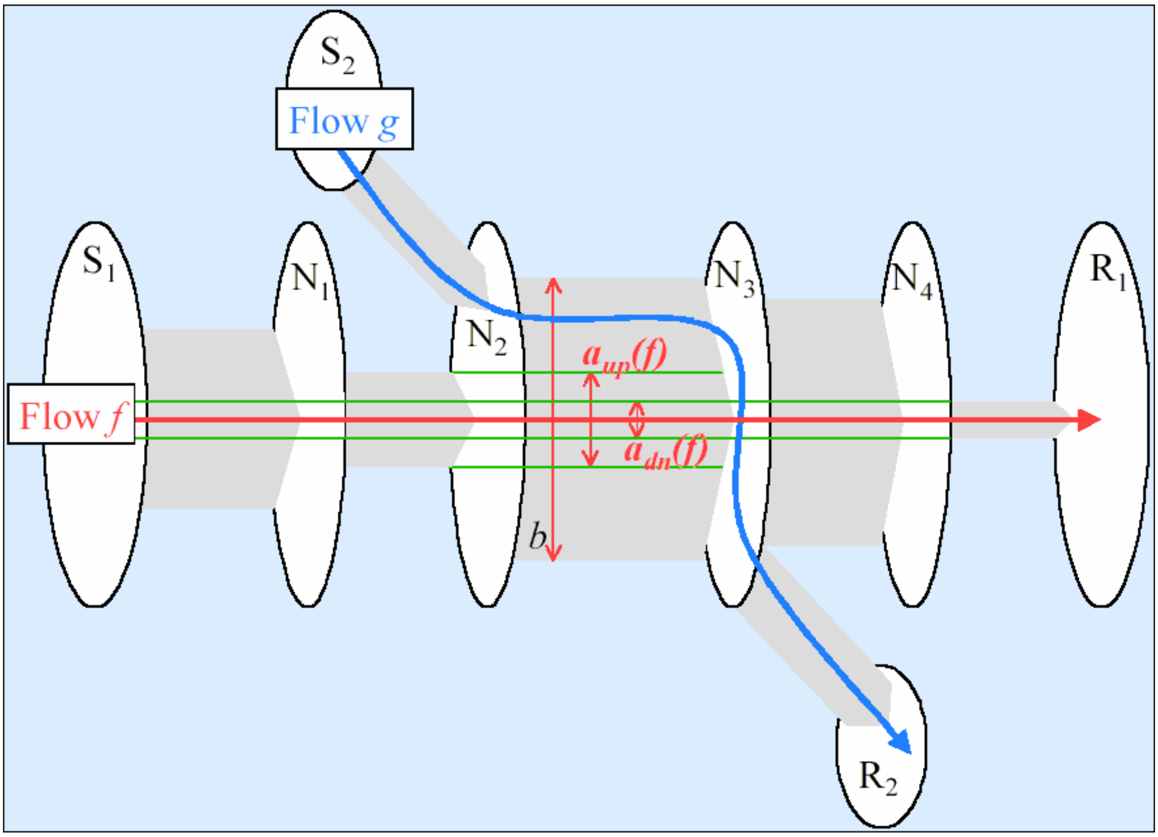


Figure 2.9. The DRSVP for a couple of flows.

Figure 2.9 shows the situation that would occur if a new flow were added that also traversed the link from *N2* to *N3*. Node *N2* must now decide how to divide the bandwidth of this interface between flows f and g. Since flow *f* is limited by external bottlenecks, *N2* knows that all the bandwidth above level $a_{dn}(f)$, as shown, can be allocated to flow *g* without affecting the end-to-end reservation for *f*. If the resources requested by *g* were high enough or if additional flows were added across this link or if the bandwidth of the link were to decrease, N2 might find that the bandwidth it could allocate to *f* was less than the value of a_{dn}. Node N2 would then become the new bottleneck for flow *f*, which would be reported in Resv and ResvNotify messages forwarded by N_2, affecting the end-to-end reservation level for the flow.

For details and implementation of DRSVP see [67].

2.8 The importance of passive reservations

In previous paragraphs a detailed description of how the QoS can be guaranteed in wireless networks has been given. This PhD thesis also faces the problem of how a certain level of service continuity can be granted to mobile hosts when they belong to MIP class service, by pre-reserving a certain amount of bandwidth when a mobile host is admitted into the network, as described in paragraphs 2.5 and 2.6.

The curves illustrated in the last paragraphs of chapter 1 have been obtained by using MATLAB tool and by implementing the proposed model in [19]. The FSMC has also been implemented in C++ and integrated with a complete network simulator based on the MRSVP, where users can request MIP or MDP services. By now, let us ignore the particulars of implementation that will be given in the last chapter, where a full and deep analysis of the simulator characteristics and of the proposed architecture is presented. For sake of simplicity, let us now consider a one-dimensional mobility environment (it will be extended in a two-dimensional one), as the one depicted in figure 2.10. As illustrated, there are 5 wireless cells, each one covered by an AP; a MRSVP sender is connected to the APs through an "infite-bandwidth" wired switching-subnet. The total bandwidth of each link is 5.5Mbps. Each mobile host starts its flow (after the CAC) in a certain current cell (e.g. mobile host 1, MH1, in cell C1), then it moves straight in a circular way (e.g. if it starts in the cell C4, it will visit C4, C5 then C1, C2, etc.), until it has visited all the cells or the connection has finished. The proposed Call Admission Control and Bandwidth Reallocation schemes will be discussed in chapter 4. Users moves according to the Random Way Point Mobility Model (RWPMM) described in [68]. Some important simulation parameters are:

- mean of requests arrival rate (Poisson process) λ_a: 3 flows/s;
- exponentially distributed call duration with mean μ = 180s;
- admissible bandwidth levels for each flow (Kbps): 512, 640, 768, 896;
- token bucket size (bit): 896000;

- token bucket rate (bit): 512000;
- token bucket peak-rate (bit): 896000;
- packet size (bit): 512;

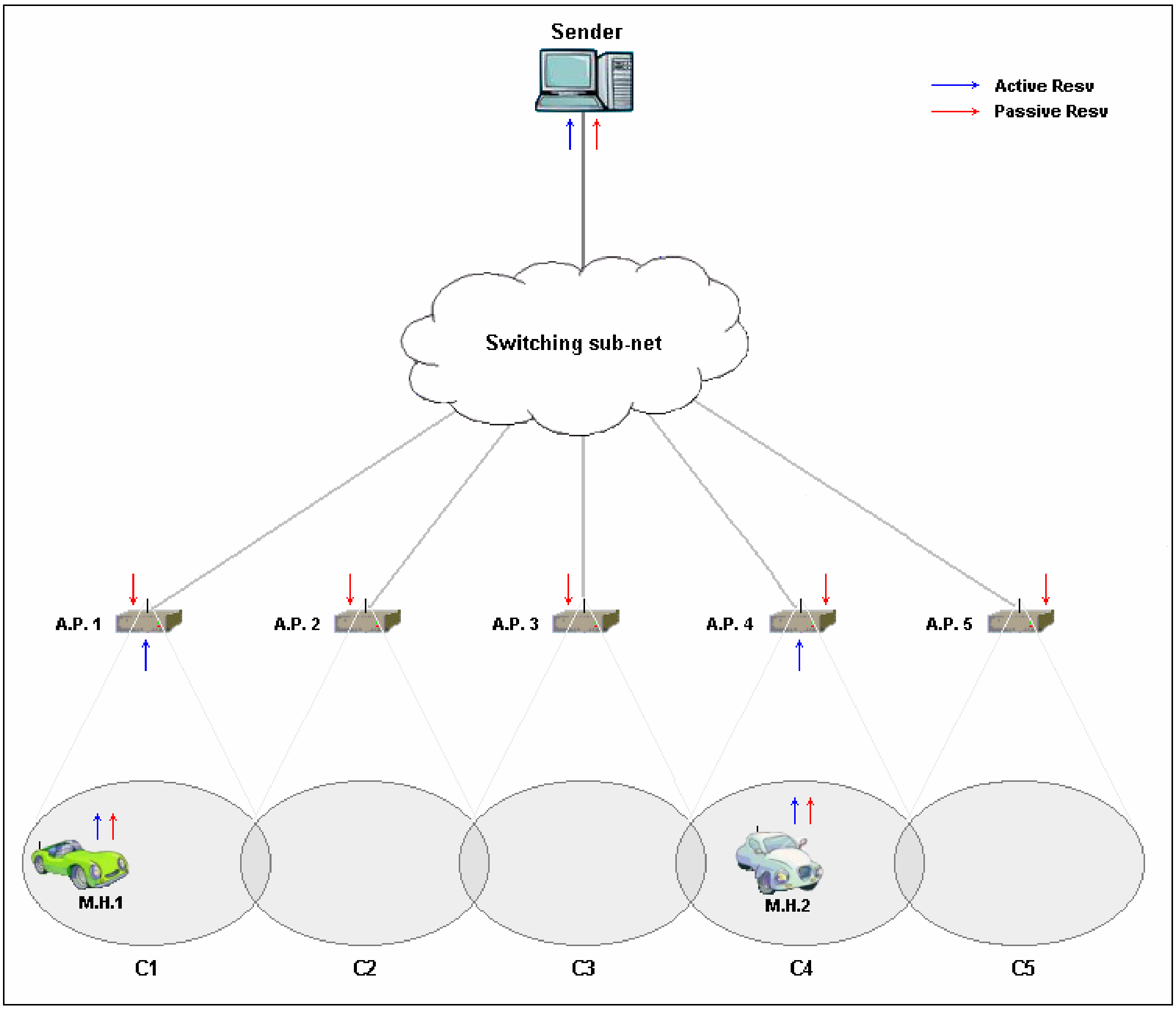


Figure 2.10. A simple wireless environment for MRSVP performances evaluation.

The FSMC has been tuned for the CCK modulation with an average SNR ρ=4dB (4-state FSMC). The performance of the system are investigated in terms of received bandwidth, user satisfaction level (this concept will be full analysed in chapter 4, through the introduction of the utility functions), average system utilization and average number of admitted and dropped flows. These are the typical parameters that are usually observed in order to evaluate the correctness of the proposed policy. Figure 2.11 depicts the trend of the average received bandwidth for MIP and MDP users. It must be outlined that the received bandwidth differs from the assigned bandwidth by the APs: it takes into account the degradation introduced by the wireless link (according to the model proposed in chapter 1); if at the time $t=t_0$, B is the assigned bandwidth introduced by an AP to user i, then it will receive $R=B\cdot(1-d_j)$, where d_j is the nominal BER of the FSMC when it stays in state j (for more details, see chapter 1). After this little introduction, let us now explain the courses of the curves of figure 2.11.

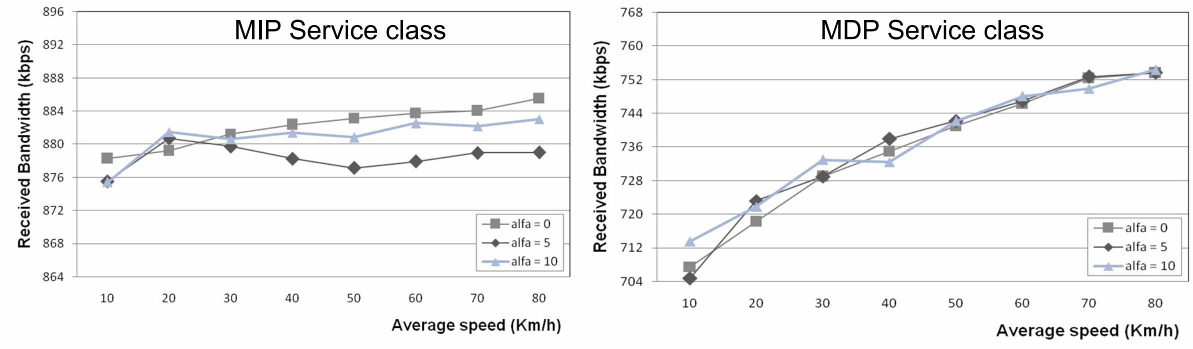


Figure 2.11. Average received bandwidth for MIP and MDP users.

There are many considerations that can be made: first of all it must be noticed how user mobility does not affect MIP services performance; the highest gap is found for v_{avg}=80Km/h, where the difference between the curves for a=0 and a=5 is about 8Kbps, that can be disregarded, considering a constant value for the received bandwidth, either for different values of a or for different values of average speed. As we will see later, the a value represents the variation around the average speed v_{avg} of the Random Way-Point Mobility Model (RPMM): higher values of a leads to a higher standard deviation around the average speed value v_{avg}. In the MDP case the trend is quite different: no evident differences can be observed for different values of a but the received bandwidth is strictly related to the average hosts speed: when it increase, better services can be offered to the admitted flows; that is the average time spent in a cell decreases, each AP does not reach a stable admission configuration (that is a stable number of admitted flows) so there is always a higher amount of bandwidth that can be shared among the active MDP flows; the main difference between the cases of v_{avg}=10Km/h and v_{avg}=80Km/h is about 50Kbps. Thus, the first analyzed performance parameter gives a first description of how the MIP and MDP services are differentiated.

Figure 2.12 illustrates how MIP and MDP users are satisfied for the received service; for the MIP case the values are contained in the range [2.4, 2.9], while for the MDP one the range is [1.2, 1.5]. First of all, it must be outlined that the used utility function takes its values in the range [1, 4] with a non-decreasing trend (this concept will be well explained later, in chapter 4).

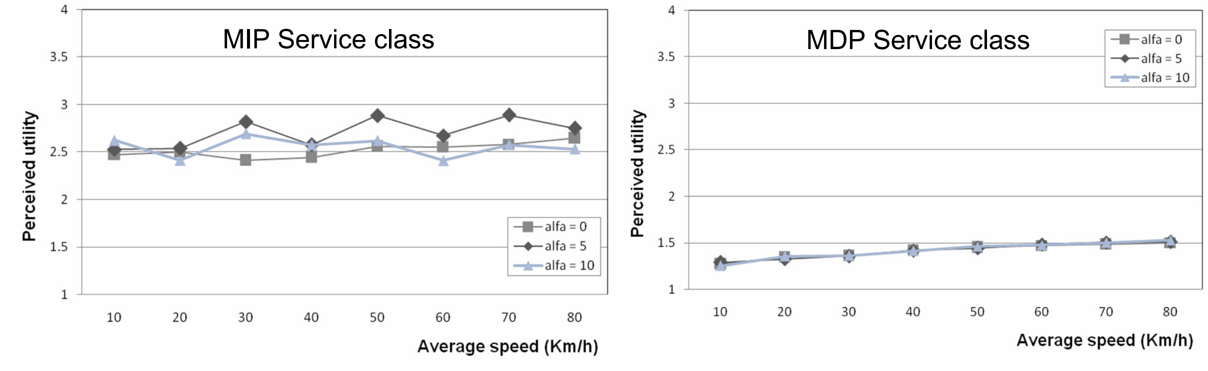


Figure 2.12. Average perceived utility for MIP and MDP users.

So it is evident that for MDP users the service offers lower QoS guarantees than for MIP ones, so the satisfaction level is near (but not below) the minimum bound. For the MIP services, the perceived values do not reach the maximum because of the presence of channel multipath fading, that introduces some kind of service degradation. One common consideration must be underlined: in each case, MIP or MDP, the bandwidth allocation algorithm and the CAC scheme must ensure that the minimum service guarantees will be always guaranteed (in the case of the utility values, figure 2.12, the perceived level must not go below the lower bound). In the next chapters there will be given a complete description of how this can be done.

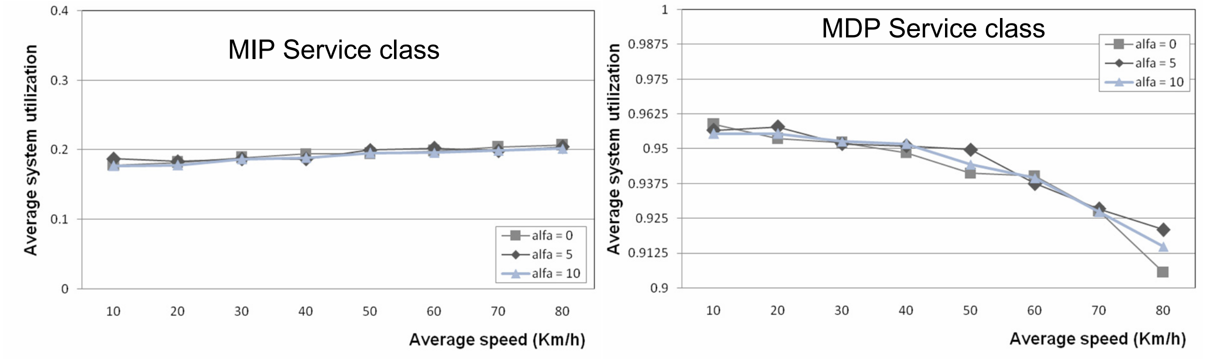


Figure 2.13. Average system utilization for different types of traffic, MIP and MDP.

Figure 2.13 confirms the independence of MIP users from mobility: there are no variations either for different values of *a* or for different average speeds. That is to say, the passive-reservation policy enhances the offered service, but maintains the network at a lower level of utilization: in figure 2.13, for MIP users, the value is around 18%-20% and can be considered constant or independent from mobility parameters. As early announced, the low system utilization is due to the presence of a heavy amount of passive reservations, that is a large amount of reserved but unused bandwidth. Our studies demonstrated that this problem can be avoided if a certain grade of resource multiplexing (MDP over MIP passive reservations or MIP over MIP passive reservations) is introduced. This will be shown in the fourth chapter of this PhD thesis. When only MDP services are available in the network, the utilization increases until to the maximum (around 95%-96%) for lower speeds, while there is a decreasing of about 5% for higher speeds: it is due to the higher presence of hand-off events (lower cell residence time) and the higher presence of signaling overhead (for the hand-off management). So, offering MDP services will maintain the system at a good level of utilization while, for better QoS, MIP services can be offered, paying in terms of system utilization.

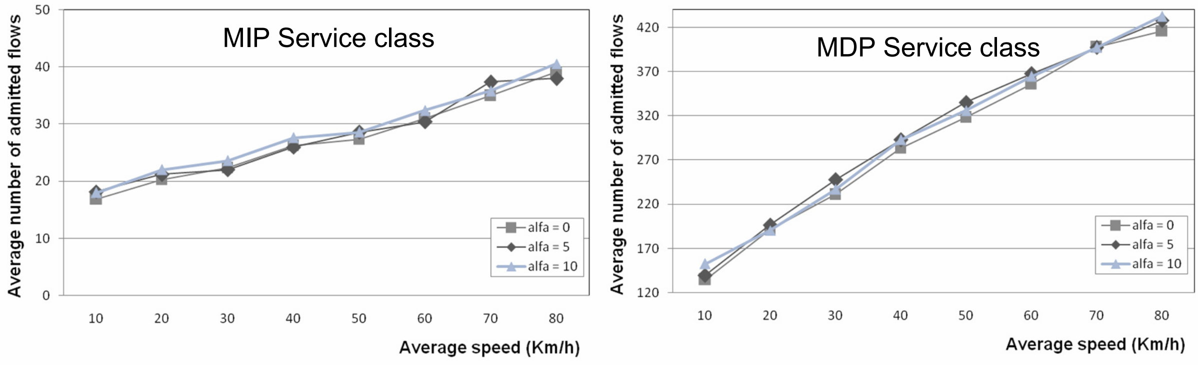


Figure 2.14. Average number of admitted flows.

As it can be expected, the number of admitted flows for MIP and MDP services is very different: from 18 to 40 for MIP case and from 135 to 420 for MDP case (figure 2.14). The number of admitted flows is always increasing for higher speeds, due to the lower cell residence time and to the higher bandwidth availability for new calls; this effect is more evident for the MDP case (a difference of 280 serviced calls, against the MIP difference of 22 serviced calls) because there is a less restrictive admission control (as explained later) without the presence of passive reservations; for MIP services, when a user leaves a cell, the current active bandwidth will be released, a new passive-to-active switch will be made, but the other passive reservations will influence the admission control of the next cells (APs). The values for the MIP case are lower than the MDP one for two reasons: the MIP CAC scheme must ensure the bandwidth availability not only on the current cell but also on the future-probably-visited cells; in addition, the already-existing passive reservations will limit the amount of available free bandwidth. Our PhD study also regards the optimization of QoS and system utilization through the introduction of some statistical and predictive considerations (see next chapters).

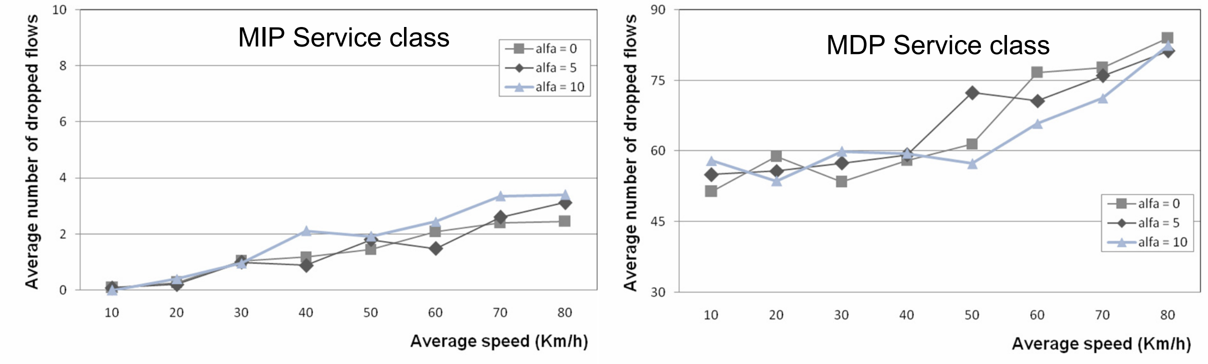


Figure 2.15. Average number of dropped flows.

Figure 2.15 depicts the trend of the average number of dropped flows during active sessions: it strictly depends on the adopted bandwidth reallocation algorithm. For the MIP case, there is a negligible amount of dropped flows: as it will be explained when the reallocation algorithm is presented (chapter 4), a MIP flow can be dropped only when an outage event occurs (the outage probability is a priori fixed and in these campaigns of simulation we set p_{outage}=0.4), so the curves regard some events that are independent from the passive-reservation policy. The slight increasing for higher speeds is due to the higher value of Doppler shift, which makes the channel more fluctuating among its states. Nevertheless, the average number of dropped flows for the MIP case is negligible (below 8%). The MDP management leads to higher percentage of dropped flows than the MIP case: the percentage is contained in the range [22%, 45%], depending on the average host speed. In this case, the absence of pre-reservation policies makes the offered service nearer to the "Best-Effort". As in the MIP case, the dropping probability is affected by the outage event and, in addition, there is the pre-emption of MDP flows by the MIP ones, when reallocating the bandwidth.

2.9 Conclusions on Chapter 2

In conclusion, in the second chapter of the PhD thesis the main mechanisms of QoS guaranteeing have been exposed. It has been noticed how the migration from "Best-Effort" services to those wit QoS specifications, based on some service constraints, is possible, using the adequate protocols that depend on the considered environment. After an overview of the QoS concept for telecommunication systems, the RSVP and its extensions have been presented, in order to acquire the needed information about the QoS mechanisms. In the three years of doctoral activity, the attention has been focused on the management of QoS for mobile host, so the MRSVP has been implemented for this purpose. In particular two classes of service have been considered: MIP (for those tolerant applications which require fairly reliable delay bounds in all cells it might visit and does not want to be affected by mobility of hosts) and MDP (for tolerant applications, which can tolerate the effects of delay violations due to mobility of hosts). After a first addicted campaign of simulations under the RWPMM some curves have been shown, illustrating how a certain grade of QoS can be reached for MIP service class, by using the passive reservation policy. The pre-reservation policy is necessary if the independence from mobility must be granted, offering a service continuity also during hand-off events. The main problem of the pre-reservation of passive bandwidth is the a-priori knowledge of the number of cells that the user will visit (in a one-dimensional environment) or what cells the user will visit (in a two-dimensional environment), in order to build a correct MSPEC object in the MRSVP packets. These problems will be assessed when the prediction algorithm is proposed, as it will be illustrated in chapter 4.

Chapter 3 – The rate adaptation in wireless networks

3.1 Introduction

In the last years, the demand for wireless communications has been exponentially increasing: the propagation of portable devices has led to a constant increasing of Quality of Service requests; in addition, users mobility has a heavy impact on real-time applications service parameters so, as exposed in previous chapter, the existing architecture for fixed hosts is not able to support and to manage the mobility effects of mobile hosts and there is the need of employing new protocols as the MRSVP. In an ISPN with mobile hosts, if a high number of users are trying to be covered by the same AP, a certain level of links congestion may be suffered; as consequence, not all the QoS constraints are respected and the received service will be surely degraded. In order to avoid such kind of problems, the admission control must limit the number of user that can be admitted into the coverage and/or, alternatively, different bandwidth levels can be used for the applications (for example, there are also some hardware codifiers that can adapt the bandwidth of the transmitted packets from 512Kbps to 1024Kbps). So, it is also possible to diminish the effects of congestion by reducing the assigned bandwidth of single flows (adaptive services) [69]. In this PhD thesis, a novel approach to both admit new calls and adaptively manage their bandwidth is proposed, in respect of some important criterion, like the fairness or the high system utilization. The need of dynamically change the transmission rate during active sessions introduces implicit overhead. It will be also shown that introducing flexibility in the assigned rate leads to better system performances, in terms of admitted calls and system utilization.

3.2 The concept of multi-rate transmission

Let us now hypothesize that a transmission source can dynamically adjust its transmission rate during an active session with the flow; considering this feature, when wireless network is in congestion temporarily, some services' bandwidth can be adjusted (we call it as rate adaptation), to avoid disconnecting some connections, resulting in an improved QoS for all the users. Because wireless link is always the bottleneck of the whole communication system, we now consider the bandwidth allocation and rate adaptation problems on wireless link in a cell (only for sake of simplicity). Suppose there are *n* total Mobile Nodes (MN) in a cell. So the set of active connections is $S = \{f_0, f_1, ..., f_{n-1}\}$. Then the *k-th* connection ($0 \leq k \leq n\text{-}1$) can operate at any of the j_k bandwidth levels $l_{k0}, l_{k1}, ..., l_{k(jk-1)}$, where $l_{k0} > l_{k1} > ... > l_{k(jk-1)}$. It is also assumed that both sides of the connection have negotiated the bandwidth level. We call l_{k0} and $l_{k(jk-1)}$ as the maximal bandwidth requirement and minimal bandwidth requirement respectively. We assume that the *k-th* connection currently operates at l_{ki}, with ($0 \leq i \leq j_{k-1}$). Let *C* be the total capacity of the cell. When $\sum_{f_k \in S} l_{k(j_k-1)} < C < \sum_{f_k \in S} l_{k0}$, the cell is in congestion. All the connections operate at a level between the maximal and minimal bandwidth requirements. If there is any change of the network condition in the cell, rate adaptation is needed to adjust the bandwidth of each connection accordingly. There are two kinds of rate adaptation: rate degrade and rate upgrade, which mean to change MN's bandwidth from high level to low level or from low level to high level respectively. The rate degrade can be described as:

"when $C < \sum_{f_k \in S} l_{ki}$ find a subset $S' \subseteq S$ and change the bandwidth of each connection in S' to $l_{ki'}$, so that $\sum_{f_k \in S} l_{ki} - l_{ki'} \geq \sum_{f_k \in S} l_{ki} - C$".

Rate upgrade can be described in a similar way.

When cell is in congestion, the network condition changes which will trigger rate adaptation include:

- MN's *hand-in*: base station needs to degrade the bandwidth of current connections to ensure that the new comer can operate at the minimal bandwidth requirement at least; otherwise the handoff will be rejected;
- MN's *hand-out*: its bandwidth can be shared by other connections;
- *Connection termination*: this case is the same of the previous case;
- *New call arrival*: though the cell is in congestion, there is still new call arrival. To ensure a high successful hand-off probability, we still reject the new call though it can be satisfied by rate adaptation. So this case needs no rate adaptation.

Two essential problems in rate adaptation are: "which connections should be adjusted?" (that is to say find the *S'* set), "how much bandwidth should be adjusted for each connection in S''(that is to say calculate each $l_{ki'}$). In this chapter a new bandwidth adaptation algorithm is proposed, in order to face the problem of dynamically guarantee the QoS in ISPN with high system utilization. Two main performance properties in rate adaptation are: the network overhead and fairness. The rate adaptation procedure includes two parts: rate calculation and bandwidth notification. So the overhead of rate adaptation also includes two parts: computation overhead and message overhead (also refer to as network overhead). The former means network node needs to calculate how to adjust the bandwidth, while the latter means messages should be sent to notify the corresponding node of the adjusted connection to ensure it to operate at the bandwidth after rate adaptation. So the network overhead is proportional to the dimension of S'. To ensure each connection to adjust bandwidth fairly during bandwidth allocation and adaptation, some fairness criteria have been defined, the most famous of which is the max-min fairness [70]. The bandwidth allocation procedure of this criterion is: firstly, allocate the bandwidth among the connections passing the bottleneck link equally; secondly, remove these connection and this link, considering the left network and connections, equally allocate the bandwidth on the bottleneck link at this time; repeat this operation till all the bandwidth in the network are allocated. The whole procedure needs iteration.

Fairness and network overhead are in contradiction with each other: to ensure good fairness, frequent rate adaptations are needed, which will result in huge network overhead; to reduce network overhead, the fairness will deteriorate. Many rate adaptation schemes are proposed to get a tradeoff between the two properties.

3.3 The importance of utility functions in wireless systems

The utility concept in telecommunications [72] generally refers to a function which describes the degree of user satisfaction with a certain amount of allocated resource. Utility functions are often introduced in order to maximize some indexes when a reallocation algorithm is introduced in the wireless bandwidth management. Utility functions are also used to describe the characteristics of different kinds of traffic (Best Effort – BE, Real Time – RT, Tolerant Applications – TA) and different courses are obtained, depending on the specific application: in this way bandwidth reallocations depends on traffic types, available resources and the channel quality, rather than solely dependent on the channel quality or traffic types as assumed in most existing works.

As previously exposed, resource allocation has been an active research topic in wireless networks [69], [71]. In such networks, radio resource is limited and the channel quality of each user may vary. Given channel conditions and total available bandwidth, system resource is allocated to users according to some performance metrics such as throughput and fairness or according to the types of traffic. "Throughput" and "fairness," however, are conflicting performance metrics. To maximize the system throughput, the system may allocate more resource to users with better channel conditions. This may cause radio resource monopolization by a small number of users, leading to unfairness. However, if the system attempts to provide fair treatment to all users, users with worse channel conditions tend to be allocated more resource so as to compensate for their channel conditions. Thus, the system throughput may be degraded dramatically. When utility functions are introduced, the "throughput-fairness" dilemma is avoided and the attention is focused on "user satisfaction" for resource allocation. Since it is unlikely for the system to fully satisfy all users with different demands, the total degree of user satisfaction should be maximized. The degree of user satisfaction with a given amount of resource is described by a utility function. A utility function *U(l)* is a non-decreasing function with respect to allocated resource *l*. The more resource allocated, the more satisfied the user. The marginal utility function defined by:

$$u(l) = \frac{\partial U(l)}{\partial l}, \qquad (3.1)$$

so it is the derivative of the utility function *U(l)* with respect to the given resource *l*. The exact expression of a utility function may depend on the traffic type and can be obtained by studying the behaviour and feeling of users (this kind of study is made by psychologists and economists). Some examples of utility functions applications are [73] and [74]. In [73], a power control scheme based on the utility function with respect to channel quality is proposed, while in [72], a utility-based scheduler together with a Forward Error Correction (FEC) and an ARQ scheme is proposed. Utility functions have been widely used in Internet pricing and congestion control [75]: many bandwidth pricing schemes have been proposed for wireless networks and the typical approach is to set the price for radio resource and allocate tokens to users so as to maximize the "social welfare" based on a users' bidding process. The bidding schemes, while useful for Internet pricing and congestion control, may not be practical for wireless networks. Since the traffic type, the number of users and channel conditions are all time-varying in wireless environments; the bidding process is very expensive as the users must keep sending additional bidding data back and forth for real-time bidding. Besides, the control protocols of the wireless system must be modified to accommodate this process. Optimizing the total utility at the base station can also maximize the "social welfare" as in a bidding scheme but in a simpler way.

Suppose that *n* users are served by a base station (or by an AP). Let *C* denote the total radio resource available at the base station, and l_i the resource to be allocated to user *i*. Each user may have a different degree of satisfaction with a given resource, guided by the respective utility function of the traffic. Users with the same kind of traffic may not have the same utility function in a wireless network, because the wireless channel quality of each user may not be identical (see chapter 1 for details). Let q_i denote the channel quality of user *i*, with $0 \leq q_i \leq 1$ and $i = 1,2, \ldots, n$. A smaller value of q_i indicates a worse channel quality. Given an amount of resource l_i and channel quality q_i , the amount of resource actually beneficial to user *i* is equal to $\theta_i=l_i \cdot qi$. Therefore, the utility function of user *i* can be expressed as $U_i(l_i)=U(l_i \cdot qi)$, where *U(.)* is the utility function of the traffic under consideration and U_i *(.)* is the utility function for the type of traffic described by *U(.)* but taking into account the channel quality of user *i*. Thus, the marginal utility function of *U(.)* is:

$$u(l) = \frac{\partial U(l_i \cdot q_i)}{\partial l_i} = q_i \cdot u(l_i \cdot l_i), \tag{3.2}$$

While for $U_i(.)$ we simply have $u_i(l_i)$. The main goal of a general utility-based objective is to maximize:

$$\sum_{i=1}^{n} U_i(l_i), \ \ subject \ \ to \ \ \sum_{i=1}^{n} l_i < C \ \ and \ \ l_i > 0 \ \ \forall i, \ \ i \in \{1,2,...,n\}. \tag{3.3}$$

Let us recall the definition given in [72]:

Def. 3.1: "*A resource allocation* $L^*=\{l_1,l_2,...,l_n\}$ *for n users is called an optimal allocation if for all feasible allocations* $L_a=\{l'_1,l'_2,...,l'_n\}$,

$$U(L^*) \geq U(L_a), \tag{3.4}$$

where $U(L^*) = \sum_{i=1}^{n} U_i(l_i)$ *and* $U(L_a) = \sum_{j=1}^{n} U_j(l'_j)$".

Note that the optimal allocation may not be unique in the system. To achieve optimal resource allocation in a wireless network, traffic of different utility functions should be treated differently. In this PhD thesis, MIP and MDP users (see chapter 2) has been considered with different types of traffic. An empirical study [76] shows that the utility functions of "*Hard QoS*" (e.g., CBR) and elastic (i.e., Best Effort) traffics are a unit-step function and a concave function, respectively, with respect to allocated bandwidth, as depicted in Fig. 3.1.

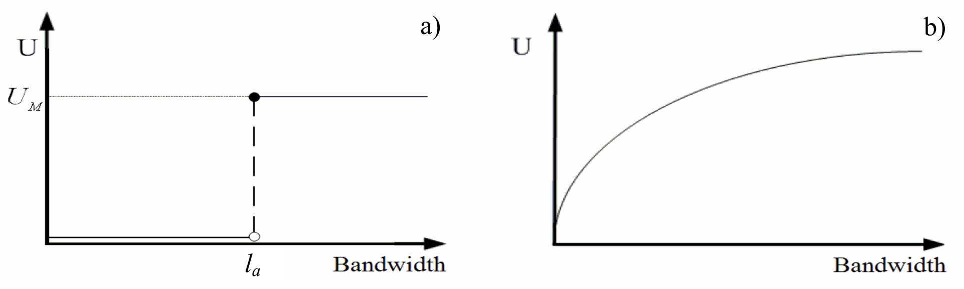


Figure 3.1. Different courses of utility functions: a) Hard QoS, b) Best Effort.

These curves can be characterized by the definitions given in [72]:

<u>Def. 3.2</u>: "*A unit-step utility function for QoS user i is described as* $U_i(l)=U_{Mi} \times f_u(q_i.l - l_{Mi})$*, where* $f_u(.)$ *is a unit-step function,* q_i *is the channel quality of this user,* M_i *is the kind of QoS traffic for this user,* l_{Mi} *is the preferred allocated resource and* U_{Mi} *is a normalization factor of this traffic*"; U_{Mi} is shown in figure 3.1 (a).

<u>Def. 3.3</u>: "*A concave utility function U(l) refers to a utility function with u(l)>0 and u'(l) < 0 for all l, where u(l) is defined as* $u(l) = \frac{\partial U(l)}{\partial l}$ *and u'(l) is defined as* $u'(l) = \frac{\partial U(l)}{\partial l} = \frac{\partial^2 u(l)}{\partial l^2}$".

By definition, a unit-step function is a discrete function and a concave utility function must be a non-increasing and continuous function with respect to resource *l*. Figure 3.2 plots the marginal utility functions of the traffic shown in figure 3.1.

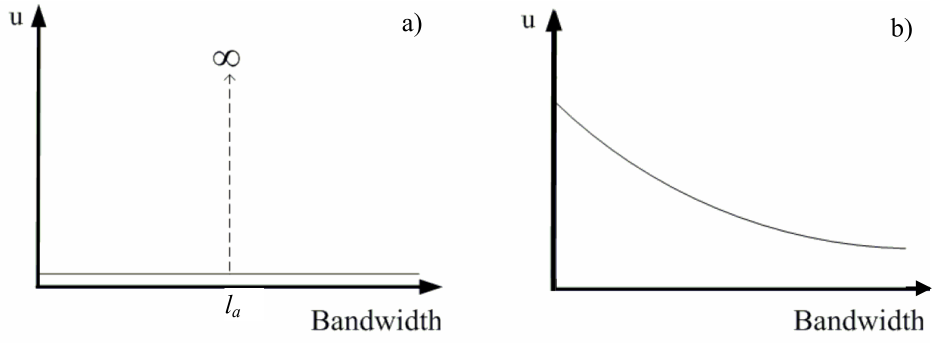


Figure 3.2. Marginal utility functions of two types of traffic.

Taking into account the above considerations, a new utility-based Call Admission Control and Reallocation Algorithm is proposed in this PhD work. In the following, the main steps of the proposed algorithms are proposed.

3.4 A new Utility-Based (UB) Bandwidth Allocation alGorithm (BAG) and a new threshold-based CAC scheme

As introduced in chapter 1, wireless links are usually subject to two types of variations, i.e. slow variations (shadowing) and fast variations (fading). For typical cellular communications, the duration of shadowing is in the order of seconds or tens of seconds, while fading usually lasts for milliseconds or shorter. To handle physical link variability, low level mechanisms such as error correction coding and swapping transmission opportunities in packet scheduling [36] are usually used. However, such mechanisms work for relatively small time scales, e.g. the duration of a symbol or a packet, which are comparable to the duration of fast fading. For slow link variations, such mechanisms alone are inadequate. In adaptive coding, a user with very bad link quality may waste a great amount of bandwidth on coding overhead. Swapping transmission slots in packet scheduling tries to improve effective bandwidth utilization. In most wireless scheduling schemes, where a two-state Markov channel model is used, a user will not receive any bandwidth when its link experiences a long-lasting shadowing degradation. However, in reality, the capacity of a wireless link will have more than two states. When slow variations are dominant, a more desirable approach is to change both the code length and the amount of bandwidth allocated to a user as its link changes state. Thus, a user will still be able to receive some service when its link quality degrades and the overall bandwidth will be utilized more effectively by allocating more bandwidth to users who can better utilize it. When fading and shadowing occur simultaneously, the fast variations are superimposed on slow variations, with the latter actually determining the short-term (in the order of seconds) average link quality. To improve bandwidth utilization, in addition to the swapping mechanism at the packet level, a high-level bandwidth allocation scheme should adjust the average bandwidth share (e.g. the scheduling weight) of each user as the average link quality changes. The main focus of this paragraph is to establish a very general modelling framework of the high-level bandwidth allocation problem based on an adaptive QoS model and design an adaptive allocation scheme to deal with link variations, taking into account the basic principles of a good bandwidth management scheme, like high throughput, fairness and low packet error rate (PER).

3.4.1 Proposal of the UB-BAG

In chapter 1, it has been deeply described and proven that the Finite-State Markov Channel (FSMC) can accurately model both fading and shadowing channels. Now a brief resume of the analytical treatment is given, in order to demonstrate the importance of the issues described in chapter 1. Each channel state corresponds to some channel quality and/or response at the receiver. To completely describe such a Markov channel, we need the state transition probabilities, average state-holding times and some parameter that reflects the physical characteristics of each state. The transition probabilities can be specified by a transition probability matrix $\boldsymbol{T} = [t_{i,j}]$, where $t_{i,j}$ is the transition probability from state i to state j, as illustrated in eq. (1.18). Each wireless link is modelled by a FSMC. Assuming that all the users move freely in the same region, all the links are independent and identical. To capture the link characteristics of each state of the Markov chain, we associate each state m with a parameter called bandwidth degradation ratio D_m, where $0 \leq D_m < 1$, with $1 \leq m \leq K$. The bandwidth degradation ratio represents the overall degree of bandwidth wastage incurred by unsuccessful transmissions, coding overhead and other factors. More specifically, if the bandwidth allocated to the user is l and its link is currently in state m, $D_m \cdot l$ of bandwidth will be wasted. We call $(1-D_m) \cdot l$ the effective bandwidth received by the user.

Utility, a concept originally used in economics, has been brought into networking research in recent years; it represents the "*level of satisfaction*" of a user or the performance of an application. A utility function, which is monotonically non-decreasing, describes how the utility perceived by a user changes with the amount of effective bandwidth it receives. The key advantage of the utility function is that it inherently reflects a user's QoS requirements and can quantify the adaptability of a user or an application. In an adaptive QoS model, user applications are required to be adaptable to service degradations and the bandwidth allocated to the user is not fixed, but adjusted according to the condition of the network. In this work, we propose a utility-oriented adaptive QoS service model for wireless networks: each user *i* signals its utility function $U_i(l)$, minimum utility level $u_{i,min}$ and maximum utility level $u_{i,max}$ to the network, where *l* is the amount of effective bandwidth received by the user. At any time instance, the instant utility value of the user is either zero or in the range of $[u_{i,min};u_{i,max}]$.

As described earlier, the communication link of each user can be modelled by a *K*-state Markov chain: we can indicate the average state-holding time of each state *m* with t_m and if, at a particular time instance, l_i is the amount of bandwidth that the network is allocating to user *i*, we define the received instant utility as:

$$u_i = U_i\left(\left(1 - D_{i,m}\right) \cdot l_i\right). \tag{3.5}$$

One of the objectives of the bandwidth allocation scheme is the QoS: in terms of users satisfaction, the minimum utility level must be guaranteed for each user *i*; if we define utility outage as the event that user *i*'s instant utility level falls below its minimum level, the scheme should guarantee that the probability of a utility outage is smaller than a certain threshold p_{outage}. To fully utilize the bandwidth, no bandwidth is reserved at any time, i.e. $\sum_{i=1}^{n} l_i = R$ is always satisfied. In our work it is assumed that each user *i*, *i = 1, 2, ..., n*, always can receive enough traffic to fully consume the allocated bandwidth as long as the effective bandwidth it receives does not exceed $u_{i,max}$.

In addition, the fairness criterion should also be based on utility functions: considering users *i* and *j* with average utility $u_{i,avg}$ and $u_{j,avg}$ respectively, we can define the Normalized Gap (NG) of the average received utility and the minimum level $u_{*,min}$ as:

$$G_i = \frac{u_{i,avg} - u_{i,\min}}{u_{i,\min}}, \tag{3.6}$$

so we want all users to have the same normalized gap in the long run ($G_i \cong G_j$, $\forall$i,j). In addition, the total effective utility delivered, $\sum_{i=1}^{n} u_{i,avg}$, is a criterion for measuring the bandwidth utilization.

Figure 3.3 represents the main phases of the proposed reallocation algorithm; it is carried out by each AP when the bandwidth must be redistributed after a channel link quality variation, a user admission or a user call termination.

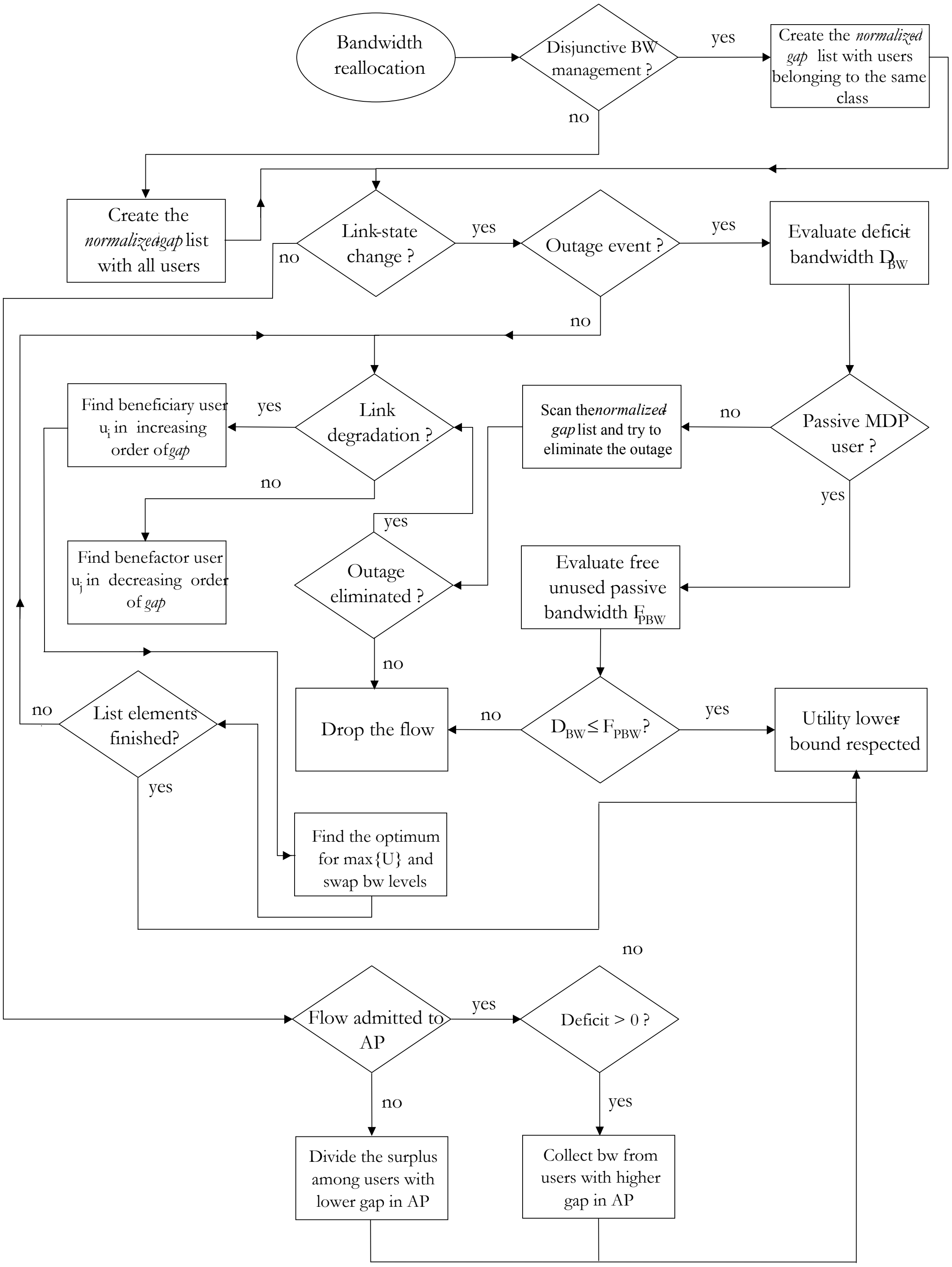

Bandwidth reallocation
Disjunctive BW management ?
yes
no
Create the *normalized gap* list with users belonging to the same class
Create the *normalized gap* list with all users
Link-state change ?
yes
no
Outage event ?
yes
no
Evaluate deficit bandwidth D_{BW}
Passive MDP user ?
no
yes
Scan the *normalized gap* list and try to eliminate the outage
Link degradation ?
yes
no
Find beneficiary user u_i in increasing order of *gap*
Find benefactor user u_j in decreasing order of *gap*
Outage eliminated ?
yes
no
Evaluate free unused passive bandwidth F_{PBW}
$D_{BW} \leq F_{PBW}$?
yes
no
Drop the flow
Utility lower bound respected
List elements finished?
no
yes
Find the optimum for max{U} and swap bw levels
no
Flow admitted to AP
yes
no
Deficit > 0 ?
yes
Divide the surplus among users with lower gap in AP
Collect bw from users with higher gap in AP

Figure 3.3. Bandwidth reallocation algorithm state flow diagram.

In order to manage different classes of service, the bandwidth allocation can be differently made for users belonging to MIP or MDP classes; so, in the first step, the algorithm must determine the amount of bandwidth that will be interested in the reallocation: if a *disjunctive* bandwidth management has been chosen, only the portion of bandwidth that is allocated to users belonging to the same service class of the user that causes the reallocation is considered; on the contrary, if a *conjunctive* policy is adopted, all the active bandwidth of the access point is considered for the reallocation. Once the amount of bandwidth that must be reallocated is determined, the algorithm determines the set of users that will be interested in the reallocation: all admitted users in an AP are sorted in a normalized-gap list.

When a wireless link changes to a state with a larger $D_{i,m}$, it is said to be degraded. When a link changes to a state with a smaller $D_{i,m}$, it is said to be upgraded. The basic idea of the bandwidth allocation scheme is that when a user's link degrades, it may surrender some bandwidth to another user with a smaller normalized gap, such that there is a net gain in the combined instant utility. When a link upgrades, the user may receive some bandwidth from another user with a larger normalized gap to achieve a gain in the combined instant utility. We assume that accurate knowledge of link states is available.

If at a particular time user *i*'s link state changes to state *p*, the following steps are performed, obtaining a sorted normalized-gap list:

1) all users' average utility level and normalized gap are updated;
2) users are sorted in increasing order of normalized gap;
3) if the instant utility level of user *i* is below the minimum (outage event), some users' bandwidth will be reduced and reallocated to user *i* to meet its $u_{i,min}$;
4) if there is no step three, user *i* may give up part of its bandwidth to another user if the link degrades, whereas it may receive some bandwidth if the link upgrades. We call the user who gives up part of its bandwidth to others the *benefactor*, and the user who receives bandwidth from others the *beneficiary*. In the third step, to satisfy user *i*'s $u_{i,min}$, the scheme searches for *benefactor(s)* starting from the user with the largest normalized gap.

In the third step, to satisfy user *i*'s $u_{i,min}$, the bandwidth allocation scheme searches for benefactor(s) starting from the user with the largest normalized gap. Suppose the user with the largest normalized gap is user *j*, whose link is currently in state *q* and it is above $u_{j,min}$. User *j* will yield:

$$\min(\frac{l_{i,\min}}{1-D_{i,p}}-l_i,l_j-\frac{l_{j,\min}}{1-D_{j,q}}) \tag{3.7}$$

amount of bandwidth to user *i*, where l_i and l_j are the bandwidth allocated to users *i* and *j*, respectively, before the link state transition. If min(.) takes the value of the second term in the parenthesis, it means user *j* can provide enough bandwidth to user *i* to satisfy $u_{i,min}$ while maintaining $u_{j,min}$. If min(.) takes the value of the first term in the parenthesis, it means user *j*'s surplus bandwidth alone is insufficient for user *i* to reach $u_{i,min}$. In this case, the instant utility of user *j* is kept at the minimum level and all surplus bandwidth is allocated to user *i.* Then the user with the second largest normalized gap will be the next candidate for benefactor. This procedure will be repeated until $u_{i,min}$ is reached or all the users have been checked.

When user i's link degrades, the bandwidth allocation scheme will search for an appropriate *beneficiary* to receive some bandwidth from user *i* and decide the amount to be transferred. Users are checked in increasing order of normalized gap. Suppose user *j*, whose link is in state *q*, is the beneficiary candidate being checked. Then:

$$\begin{cases} u_i = U_i\left(\left(1-D_{i,p}\right)\cdot l_i\right) \\ u_j = U_j\left(\left(1-D_{j,q}\right)\cdot l_j\right) \end{cases} \tag{3.8}$$

The bandwidth allocation scheme tries to maximize the combined instant utility of the two users with some constraints; the optimization problem is:

$$\begin{cases} \max(u'_i + u'_j), \quad where: \\ u'_i = U_i\left(\left(1-D_{i,p}\right)\cdot\left(l_i - x\right)\right) \\ u'_j = U_j\left(\left(1-D_{j,q}\right)\cdot\left(l_j + x\right)\right) \\ x \geq 0 \\ u'_i \geq u_{i,\min} \\ u'_j \geq u_{j,\min} \end{cases} \tag{3.9}$$

The constraints of the optimization problem can be simplified as:

$$\max\left(0, \frac{l_{j,\min}}{1-D_{j,q}} - l_j\right) \leq x \leq l_i - \frac{l_{i,\min}}{1-D_{i,p}} \tag{3.10}$$

Since the utility functions are bounded and monotonically non-decreasing and the constraints are linear, the above optimization problem is guaranteed to have solution(s), which can easily be solved by numerical methods. If the solution of $x > 0$ then the bandwidth of users *i* and *j* must be reallocated as: l_i=l_i-x and l_j=l_j+x.

If x=0, then the same procedure is repeated for the user with the next smallest normalized gap. This procedure is repeated until one beneficiary is found or all users with smaller normalized gap than user *i*'s have been searched.

The main difficulty in allocating bandwidth is how to combine utilization and fairness considerations and strike a balance between achieving high bandwidth utilization and fairness among users. If achieving high bandwidth utilization is the only one objective, some users may suffer starvations. If absolute fairness is maintained at all times, bandwidth utilization is sacrificed. The operations described actually combine the considerations of both long-term fairness and short-term maximization of bandwidth utilization. First, only users who are lagging behind user *i* in normalized average utility are in the beneficiary candidate list. Considering long-term fairness objective, when a user gives up its bandwidth, such bandwidth is transferred to the users who have received less utility than its fair share, so that they can catch up. The smaller the user's normalized gap, the higher its priority in the candidate list. Second, reallocating bandwidth between the benefactor and the beneficiary is aimed at maximizing the combined instant utility and, hence, the bandwidth utilization.

Similarly, when user *i*'s link upgrades, user *i* becomes the beneficiary and users with larger normalized gap are the candidates for *benefactor*. The scheme checks the candidates in decreasing order of normalized gap and, when the *benefactor* is found, the scheme decides the amount of bandwidth to exchange, maximizing the combined utility of the two users. Suppose user j, whose link is in state q, is the benefactor candidate being checked. The optimization problem becomes:

$$\begin{cases} \max(u'_i + u'_j), \quad where: \\ u'_i = U_i\left(\left(1-D_{i,p}\right)\cdot\left(l_i + x\right)\right) \\ u'_j = U_j\left(\left(1-D_{j,q}\right)\cdot\left(l_j - x\right)\right) \end{cases} \tag{3.11}$$

subject to:

$$\max\left(0, \frac{l_{i,\min}}{1-D_{i,p}} - l_i\right) \le x \le l_j - \frac{l_{j,\min}}{1-D_{j,q}} \ . \tag{3.12}$$

If $x > 0$, then the algorithm reallocates the bandwidth of user *i* and *j* as: l_i=l_i+x and l_j=l_j-x. If x = 0, then the same procedure is repeated for the user with the next largest normalized gap. This is repeated until one benefactor is found or all the users with larger normalized gap than user *i*'s have been searched. Besides the link state changes, adjustments in bandwidth allocation are also needed when the following events take place. Bandwidth needs to be collected from (or distributed to) the users in the network when the overall available bandwidth decreases (or increases) or a new user arrives (or departs). If *l* is the amount of deficient bandwidth which needs to be collected from the current users, either because of a decrease in overall bandwidth or a user's arrival, user *j* having the largest normalized gap G_j is to give up

$$\min(\max(0,l_j-\frac{l_{j,\min}}{1-D_{j,q}}),l) \qquad (3.13)$$

amount of bandwidth, where *q* is the current link state of user *j*. This procedure will be repeated until enough bandwidth has been collected or all the current users have been checked. If, after searching all the current users, the collected bandwidth is still not enough, the scheme will start a second round of collection, again starting from the user with the largest normalized gap, but, this time, each chosen user will be dropped out from the network. Similarly, if there is surplus bandwidth, the users with the first *k* smallest normalized gaps are chosen to receive the surplus bandwidth. Each user can increase its effective bandwidth up to the maximum effective bandwidth level. As it will be shown in next sections, a MDP user can use free available bandwidth in the current Access Point (active-MDP) or a certain amount of passive bandwidth that is reserved for MIP flows that will come in the current Access Point in the future (passive-MDP, in next chapter it will be shown how to make passive resevrations). Observing the previous algorithm description and the Data Flow Diagram (DFD) of figure 3.3 the time complexity of the bandwidth reallocation algorithm can be evaluated: let us hypothesize that there are *n* admitted users; first of all an update of the normalized gap must be made (its complexity is *θ(n)*), then the list of users must be sorted (the best performance can be obtained in *O(nlogn)*); at this time the exchange of bandwidth can be made: the benefactor and the beneficiary can be found with two list scans in *O(2n)* time and for each iteration the optimization problem can be numerically solved, in the worst case, with *h* steps, where *h=[(max_bw-min_bw)/bw_level_gap]* ; the complexity of the two scans becomes *O(2h·n)*; so the algorithm performs with a time complexity of *O(nlogn)* in the "worst case".

At this point, an overview of the proposed CAC scheme is given

3.4.2 Proposal of the UB-CAC

In order to guarantee users' minimum utility level $u_{*,min}$, an admission control policy should be enforced to limit the number of users in the system. Given a FSMC's transition probability matrix $\boldsymbol{T}$, the steady state probabilities vector $\boldsymbol{p} = [p_1, p_2, ..., p_K]$ can be calculated by solving the following equation:

$$\boldsymbol{Tp}^{\tau}=\boldsymbol{p}^{\tau}, \tag{3.14}$$

where τ is the transpose operator. If state m's average holding time is t_m, then at a particular time the probability of the link being in state m is:

$$\pi_i = \frac{p_i \cdot t_i}{\sum_{i=1}^{k} p_i \cdot t_i} \tag{3.15}$$

Recalling that when a user's instant utility falls below its minimum level there is a utility *outage* for the user, the probability p_0 of such event at any time is:

$$p_0 = P_r\left\{\sum_{i=1}^{n} \frac{l_{i,\min}}{1-D_{i,m_i}} > L_a\right\}, \tag{3.16}$$

where n is the total number of users including the new one and L_a represents the available bandwidth associated to the wireless cell c.

Modelling the wireless channel through a FSMC, it is possible to know the value of p_0 in the worst case, accounting the channel state conditions in the following way:

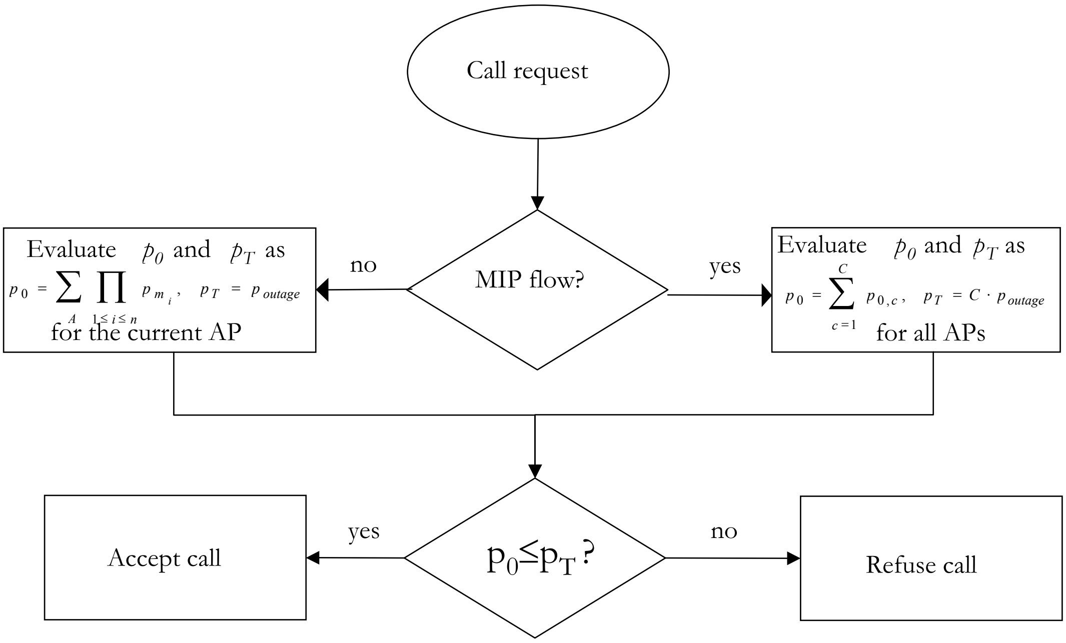
$$p_0 = \sum_{A} \prod_{1 \le i \le n} \pi_{m_i}, \quad A = \{m_1, m_2, ..., m_n \mid 1 \le m_1, ..., m_n \le K, \sum_{t=1}^{n} \frac{l_{t,\min}}{1-D_{t,m_t}} > L_a\}, \tag{3.17}$$

where m_i is user i's link state at the time istance and π_{mi} is the probability of the user i's link to being in state m at a particular time.

The CAC scheme works differently if the new request belongs to MDP or MIP class: for MDP service requests the value of p_0 is evaluated through eq. (3.17), then it is compared with the chosen threshold p_{outage}; if $p_0 \le p_{outage}$ then the call is admitted, it is rejected otherwise. That is to say the control is made only on the current cell, where the request has been made (active cell). For MIP requests the admission control algorithm works differently; the flow is admitted if:

$$\sum_{c=1}^{C} p_{0,c} \leq C \cdot p_{outage}, \tag{3.18}$$

where C is the number of cells that mobile host will probably visit (the way to determine it is explained in next chapter) and p_{outage} is the outage probability of the wireless system. So, when a new user arrives, the scheme calculates $p_{0,c}$ as described in eq. (3.17) and if $p_{0,c} \leq p_{outage}$ for each cell, the new user is admitted, otherwise it is rejected. If a user j is admitted, it is initially allocated $\frac{l_{j,\min}}{(1-D_{j,q})}$, where q is user j's current link state. The assigned amount of bandwidth to j is contributed by the users currently in the network following the algorithm we described previously. Figure 3.4 resumes all the phases of the admission control algorithm.



Call request
MIP flow?
no
yes
Evaluate p_0 and p_T as $p_0 = \sum_A \prod_{1 \leq i \leq n} p_{m_i}$, $p_T = p_{outage}$ for the current AP
Evaluate p_0 and p_T as $p_0 = \sum_{c=1}^{C} p_{0,c}$, $p_T = C \cdot p_{outage}$ for all APs
$p_0 \leq p_T$?
yes
no
Accept call
Refuse call

Figure 3.4. Admission Control Algorithm Data Flow Diagram.

Eq. (3.17) indicates the evaluation that must be performed in order to decide whether a new user can be admitted into the system: the product of the n terms (time complexity $\theta(n)$) must be repeated for all the states combinations for which the condition $\sum_{t=1}^{n} \frac{l_{t,\min}}{1-D_{t,m_t}} > L_a$ is verified, that is n^K in the "*worst case*" (supposing that the inequation is verified for all combinations) where n is the number of admitted users (including the new one) and K is the number of states of the chain model; so the time complexity of the CAC scheme for an active admission is $O\left(n \cdot n^{K}\right) = O\left(n^{K+1}\right)$; for passive reservation there is the multiplicative factor *C* that can be disregarded in the "*worst case*" complexity analysis. It is noticed that the temporal complexity is proportional to the number of states of the chain model and the number of current admitted users, so the needed time increases when *n* or *K* increase; the needed time may become unacceptable, so the on-the-fly evaluation of the eq. (3.17) must be avoided; if the number of chain states *K* is known, as well as the *t* bandwidth levels $B=\{l_1,l_2,\ldots,l_t\}$, the evaluation of eq. (3.17) can be a-priori made and stored in a CAC matrix, where the columns indicate the number of current admitted users and the rows indicate the amount of available bandwidth as explained in the performance evaluation section. In this way, the space-complexity is slightly increased, but the admission can be made in a constant time (only a selection on the matrix and a comparison must be led-out). An example of the obtained matrix is given in the last chapter of this thesis.

As for the bandwidth management, the considered bandwidth for L_a value must be accurately determined: the selected outage threshold can be the same for both service classes (*conjunctive* bandwidth management) or there can be two values of p_{outage}, (*disjunctive* bandwidth management): in the first case MIP flows can pre-empt the MDP ones, because all flows share the same resource, while, in the second case, MIP (MDP) reallocations do not influence MDP (MIP) assigned bandwidth. These considerations will be applied in the last chapter of this thesis, where a complete performance evaluation of the simulated system is deeply described.

3.5 UB BAG and CAC for ISPNs

Let us now particularize the general behavior of the proposed schemes for a given system model (this is also necessary to evaluate the performances of the algorithms); this thesis work is based on the ISPN model described chapter 2. Now a brief overview of the main characteristics of ISPNs and the way of particularize the proposed schemes are given.

ISPNs aim to offer a certain level of QoS with some fixed constraints; they provide different classes of service, such as the *predictive* class: there are MIP users (they require mobility independent services) and MDP ones (they require mobility dependent services that can be also dropped). The MRSVP minimizes the mobility effects for

MIP users through the employment of the passive reservations policy: in this way the continuity of service can be guaranteed after different hand-off events (with low system utilization). So, in a generic AP, the bandwidth at time $t=t_0$ is divided in two main sets:

- Active bandwidth, used by MIP or MDP users that belongs to the coverage of the considered AP;
- Passive bandwidth, reserved to MIP users that will hand into the considered cell in the future (the set of probably visited cells is created by an addicted prediction algorithm that will be well explained in next chapter).

These considerations are necessary for the application of CAC and BAG; there are two main ways to manage the available bandwidth:

- Conjunctive management: there is a single outage threshold that will be used for both service classes;
- Disjunctive management: two thresholds for MIP and MDP users are chosen, having the chance to set two different admission ratios.

Each bandwidth management policy splits the overall AP capacity B into a certain number of subsets, which will be differently affected by admitted users and the service class they belong to.

Let us now see the main differences between the two proposed management policies and how the passive bandwidth can be "re-used" in order to increase system utilization.

3.5.1 Conjunctive resource management

The available bandwidth is subdivided as depicted in figure 3.5:

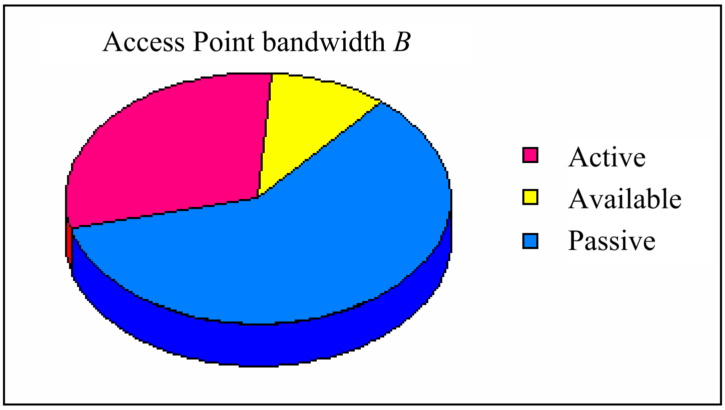


Figure 3.5. AP bandwidth *B* subdivision for conjunctive management.

That is to say $B=Active\cup Passive\cup Available$, with zero-intersection sets. So, the proposed schemes work in the following way:

- *in the call admission phase:* the CAC module takes into account all the active flows and all the active bandwidth of the AP; in particular, for the eveluation of eq. (3.17):

  MDP request: $L_a=Active\cup Available;\ n=MDPadmittedNumber+1$;

  MIP request (active or passive): $L_a=Active\cup Available$; $n=MIPadmittedNumber+$ $+MDPadmittedNumber+1$.

  The obtained p_0 value is compared with the single chosen p_{outage} threshold; MIP requests may be active or passive and they can pre-empt MDP admitted flows.
- *in the re-allocation phase:* there is no difference between the arrival/departure of MIP or MDP flows, so the service class is not taken into account.

3.5.2 Disjunctive resource management

The available bandwidth is subdivided as depicted in figure 3.6:

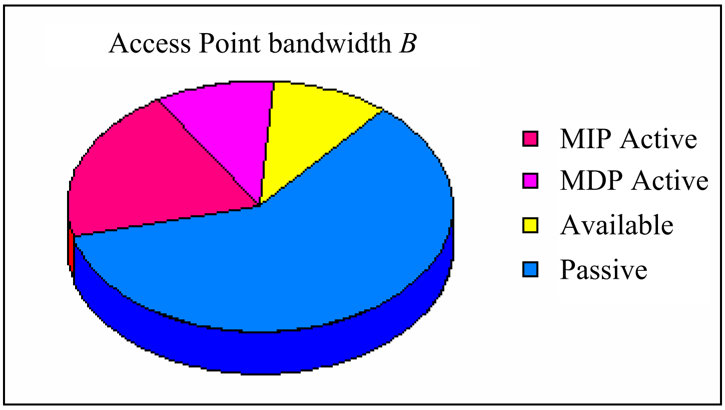


Figure 3.6. AP bandwidth *B* subdivision for disjunctive management.

That is to say $B=MIPactive\cup MDPactive\cup Passive\cup Available$, with zero-intersection sets. So, the proposed schemes work in the following way:

- *in the call admission phase:* the CAC module takes into account only the active bandwidth used by MIP/MDP flows, depending on the belonging class of the new request; in particular, for the eveluation of eq. (3.17):

  MDP request: $L_a=MDPactive\cup Available;\ n=MDPadmittedNumber+1$;

  MIP request (active or passive): $L_a=MIPactive\cup Available;\ n=MIPadmittedNumber+1$.

The difference between this policy and the previous one is evident: now there are two separate ways to evaluate eq. (3.17) because two thresholds are fixed as input parameters: $p_{outageMIP}$ and $p_{outageMDP}$ and the values of L_a and n must be accurately evaluated;

- *in the re-allocation phase:* now there is a difference between arrival/departure of MIP/MDP users: in the benefactors/beneficiaries list there will be only the users that belong to the same service class of the flow that caused the deficit/surplus of bandwidth, so the bandwidth is picked/given from/to users that belong to a specific service class.

3.5.3 Reuse of the passive bandwidth

It is evident that passive reservations of MIP flows are unused until a user makes a hand-in from an adjacent cell: at this point, the previous reserved bandwidth is switched into active resource, guaranteeing the service continuity. A way to overcome the bandwidth wastage, due to the presence of passive reservations, is to give to MDP users the chance to reuse the passive bandwidth of a cell, but when a MIP user arrives into the current AP coverage, its reserved passive bandwidth must be switched into active one and, if there is no free bandwidth availability, one ore more MDP users must be dropped: employing this new policy a higher increase in system utilization is observed as shown in the last chapter).

In both cases (with or without MDP passive bandwidth reuse) the AP capacity B is subdivided as illustrated in figure 3.7:

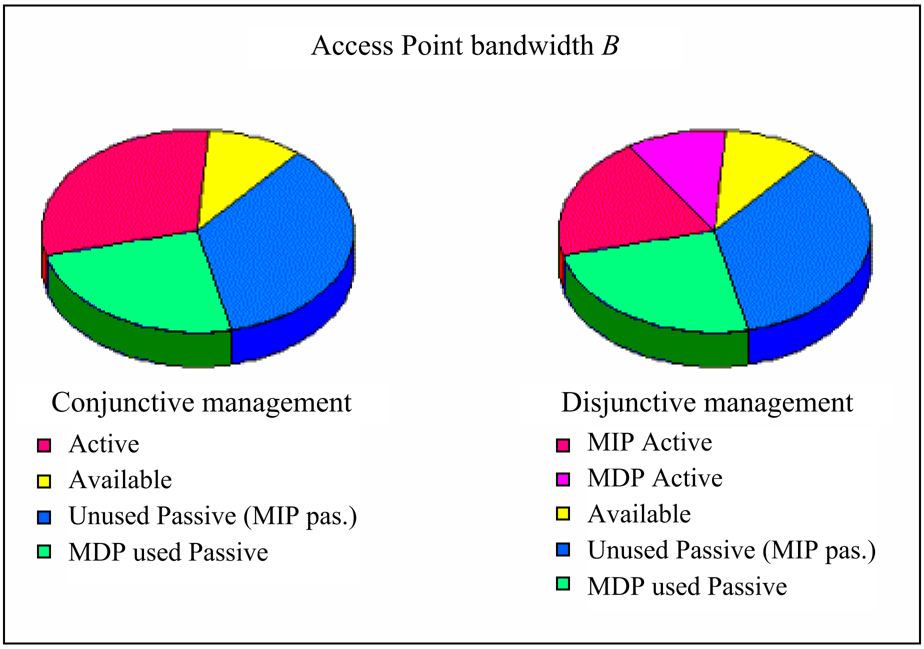


Figure 3.7. AP bandwidth B subdivision when “Passive MDP” flows are present.

Under these considerations, the values of L_a and n for the two types of management are:

Conjunctive Management

B=Active∪MIPpassive∪MDPpassive∪Available, with zero-intersection sets;

MDP request: L_a*=Active∪MIPpassive∪MDPpassive∪Available; n=MDPadmittedNumber+1*;

MIP request: L_a*=Active∪Available; n=MIPadmittedNumber+MDPadmittedNumber+1*.

Disjunctive Management

B=MIPactive∪MDPactive∪MIPpassive∪MDPpassive∪Available, with zero-intersection sets;

MDP request: L_a*=MDPactive∪MIPpassive∪MDPpassive∪Available; n=MDPadmittedNumber+1*;

MIP request: L_a*=MIPactive∪Available; n=MIPadmittedNumber+1*.

3.5.4 UB BAG and CAC

Now the proposed schemes are formally described with the use of pseudo-code. These are the employed variables:

- **Amdp**: current AP bandwidth assigned to MDP flows;
- **Amip**: current AP bandwidth assigned to "active" MIP flows;
- **S**: surplus of available bandwidth;
- **Pmdp**: passive bandwidth used by MDP flows;
- **Pmip**: passive bandwidth reserved to MIP flows (and not used by MDP flows);
- **numberMIP**: number of admitted MIP flows in the current AP;
- **numberMDPA**: number of admitted 'active' MDP flows in the current AP;
- **numberMDPP**: number of admitted 'passive' MDP flows in the current AP;

UB Call Admission Control

```
//Type of request and service class discovery
//$L_a$, $n$ and $p_{outage}$
long La=0; int n=0; double pOutage=0;
if ( request==MDP ) {
//it's a RESV message received by the current AP
        if ( management==Conjunctive ) {
                //there is only one threshold as input parameter
                pOutage=InputPoutage;
                //MIP bandwidth is excluded to give them a certain level of priority
                La=Amdp+Pmdp+Pmip+S;
                n=numberMDPA+numberMDPP+1;
        }
        else {
                //Disjunctive management
                pOutage=InputPoutageMDP;
                //MDP is admitted on the MIP passive unused bandwidth
                if (Pmip ≥ MinimumRequestedBandwidthLevel) {
                    La=Pmip+Pmdp;
                    n=numberMDPP+1;
                }
                //MDP is admitted on the active available bandwidth
                else {
                    La=Amdp+S;
                    n=numberMDPA+1;
                }
        }
}
//a MIP request has arrived with a RESV or SPEC message to the current AP
else {
        if ( management==Conjunctive ) {
                pOutage=InputPoutage;
                //MDP are also included to give higher priority to MIP flows
                La=Amdp+Amip+S;
                n=numberMDPA+numberMIP+1;
        }
        //Disjunctive Management
        else {
                pOutage=InputPoutageMIP;
                La=Amip+S;
                n=numberMIP+1;
        }
}
//the admission control can be made by the evaluation of eq. (3.17)
double p0 = evaluateP0 ( La, n);
//make the comparison
```

```
if ( p0 ≤ pOut )
        request admitted;
else
        request refused;
```

It must be outlined how the CAC, in the conjunctive case, gives priority to MIP flows. Now, the UB BAG is described.

UB Bandwidth Allocation alGorithm

```
//The bandwidth reallocation is needed in three particular situations: link state change, arrival of a new
//user (so, a degrade operation is needed) or a departure of an existing user (a flows upgrade is
//needed):
//a link state has changed
if ( link_changed) {
        //discover the identifier and the service class of the flow whose link has changed
        int id = identify_id( );
        Class class = identify_class ( id );
        User us  = identify_user ( id );
        bool degradation = ( Dm (current_status, id) > Dm (previous_status, id) );
        Vector vgap = new Vector<user_e_gap>;
        int pos=0;
        //fill the vgap list in increasing order of gaps
        while ( pos < usersInTheCell.size( ) ) {
                //in the disjunctive management only users of the same class are considered
                //in the conjunctive management all the users give their contributions to reallocations
                User ut1= usersInTheCell.elementAt( pos );
                if ( management==disjunctive ) {
                        if ( ut1.class == class ) {
                                vgap.insertOrdered ( evaluate_gap ( ut1 ) );
                        }
                else  vgap.insertOrdered ( calcola_gap ( ut1 ) );
                pos++;
        }
        //the gaps vector has been built up; now the outage event must be checked
        if ( ut.instantaneousUtility( ) ) < minimum_utility ) {
                long deficit = evaluateTheAmountOfNeededBandwidth( ut, class );
                //the needed bandwidth must be taken accounting for the belonging class
```

```
if ( ut.class == passiveMDP) {
        if ( deficit > PassiveFreeBandwidth( ) {
                dropFlowForOutageNotErased( ut );
        }
        else {
                incraeseBandwidthForOutageErasing ( ut );
        }
        //the flow is not a “passive” MDP
        else {
                bool outageErased=false;
                //checking starts from the user with higher gap
                int pos=vgap.size( ) - 1;
                while ( ( !outageErased ) && ( pos ≥ 0) ){
                        //looking in the gaps list for benefactor(s) in order to find
                        //the needed amount of bandwidth; the class management
                        //has been already taken into account while creating gaps list
                        User ben = findBenefactor ( vgap );
                        exchangeBandwidth ( ben, ut);
                        if ( ut.instantaneousUtility( ) ≥ minimum_utility )
                                outageErased = true;
                        pos--;
                }
                //the outage has not been erased
                if ( ut.instantaneousUtility( ) < minimum_utility ) {
                        dropFlowForOutageNotErased( ut );
                }
        }
}//end of outage management
//now the outage has been erased and if the user has not been dropped the
//reallocation must be made
if ( !ut.dropped ) {
        //the link has degraded and the position of the user in the gaps list must be
        //known
        int indexO = UserIndex ( vgap, ut );
        if ( degradation ) {
                //the user whose link has changed must give up, if possible, some
                //bandwidth
                long x=0; int j=0;
```

```
    //for all users with a lower gap
    while ( j < indexO  ){
        //the optimization problem of paragraph 3.4.1 must be
        //solved
        x = maximizeCombinedUtility ( vgap, indexO, j );
        if (x !=0) break;
        j++;
    }
    if ( x !=0 ) {
        //exchange the bandwidth among benefactor and
        //beneficiary
        User benefactor = vgap.elementAt (indexO);
        User beneficiary = vgap.elementAt ( j );
        exchangeTheBandwidth ( benefactor, beneficiary, x);
    }
    //if  x==0 there will not be any reallocation because neither the user
    //whose link has changed can give up some bandwidth nor other
    //users can receive some bandwidth because their level have reached
    //the maximum
}
//user link quality has upgraded
else {
    //l’utente, il cui link è variato,  deve ricevere banda se può,  perché il
    //link si è spostato in uno stato con degradazione più bassa per cui si
    //cerca un benefattore
    long x=0; int j=vgap.size( )-1;
    //for all users with higher gap
    while ( j > indexO  ) {
        //the optimization problem of paragraph 3.4.1 must be
        //solved
        x = maximizeCombinedUtility ( vgap, indexO, j );
        if (x !=0) break;
        j--;
    }
    if ( x !=0 ) {
        //exchange the bandwidth among benefactor and beneficiary
        User benefactor = vgap.elementAt ( j );
        User beneficiary = vgap.elementAt ( indexO );
```

```
                exchangeBandwidth ( benefactor, beneficiary, x);
            }
        }
    }
}//end of link change management
//a user has arrived/departed
else {
    //the arrived/departed user must be identified
    int id = identify_id( );
    Class class = identify_class ( id );
    User ut  = identify_user ( id );
    Vector vgap = new Vector<user_e_gap>;
    int pos=0;
    //create the gaps list in increasing order
    while ( pos < usersInTheCell.size( ) ) {
        User ut1= usersInTheCell.elementAt( pos );
        if ( ut1.class == class ) {
            vgap.insertOrdered ( evaluate_gap ( ut1 ) );
        }
    }
    //flow degradations
    long deficit = evaluateNeededBandwidth( );
    if ( ( new_flow_arrived ) && ( deficit > 0 ) ){
        //the reallocation starts from the user with the highest gap
        int j=vgap.size( )-1;
        //the first round of bandwidth collecting
        while ( ( deficit > 0 ) && ( j ≥ 0 ) ) {
            //evaluate the amount of bandwidth that the current user can give up
            //without creating outage
            long x = evaluate_available_bandwidth ( vgap.elementAt ( j ) );
                if ( x > 0 ) {
                    deficit -= x;
                    degradeBandwidth ( vgap.elementAt ( j ), x );
                }
            j--;
        }
    //if necessary, execute a second round
    if ( deficit > 0 ) {
            j=vgap.size()-1;
            while ( deficit > 0 ) {
                long relBandwidth = OccupiedBandwidth ( vgap.elementAt ( j ) );
                deficit -= relbandwidth;
                dropFlow ( vgap.elementAt ( j ) );
                j--;
            }
            //the deficit has surely eliminated
        }
    }//new flow arrival
    //the bandwidth must be reallocated for the departure of a flow
    else if ( departed_flow ) {
        //the procedure starts from the flow with the lower gap
        int j=0;
        long surplus = evaluateAvailableBandwidth( );
        while ( ( surplus > 0 ) && ( j < vgap.size( ) ) ) {
            //evaluate how much bandwidth the current user can receive without
            //exceeding the allowed maximum level
            long x = evaluateTheMaximumBandwidth ( vgap.elementAt ( j ) );
            if ( x > 0 ) {
                surplus -= x;
                increaseBandwidth ( vgap.elementAt( j ), x );
            }
            j++;
        }
    }
}//end of bandwidth reallocation
```

3.6 Why adaptivity in wireless networks: performance evaluation

Let us now investigate about the advantages of having a dynamic bandwidth allocation scheme that makes some decisions about the level of bandwidth that can be assigned to active flows, despite a static and less flexible scheme. Some simulations have been carried out in a simplified 1D scenario, only to show a comparison among the two policies. Only MDP users are considered, in order to avoid the presence of passive bandwidth, which may influence the real system utilization. The implemented dynamic scheme is the one proposed in the previous paragraph, while the static one manages the bandwidth as depicted in figure 3.8, where *B=Active+Available*.

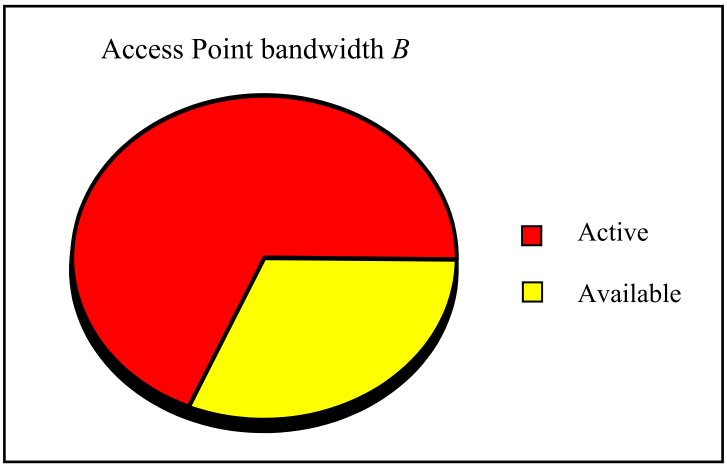


Figure 3.8. AP bandwidth *B* subdivision for a simple static bandwidth management.

Let us suppose that there are *n* admitted users in the current AP and the *(n+1)-th* makes its MDP service request; for a static bandwidth management no multiple bandwidth levels are allowed, so the occupied bandwidth by *n* MDP users will be: *Active=n·l*, where *l* is the only allowed bandwidth level. The (n+1)-th request is admitted only if *(n+1)·l<B* or, equivalently, if *Available>l*. If the new MDP flow is admitted, then *Active=(n+1)·l*.

Now, simulation results and comparisons will be shown, illustrating how the adaptivity is powerful in those systems where a certain grade of Soft-QoS is tolerated. For sake of simplicity, a simplified network is considered: it is the same of chapter 2, in figure 2.15. A generic mobile host follows the Random Way Point Model [68], adapted to a one dimensional space with a circular behavior: if the Call Holding Time (CHT) is long enough, a user that goes out from the coverage of the last cell will hand-in the first cell (from C5 to C1 in the specific case). Simulation parameters are the same of paragraph 2.8. For all the obtained curves, the static scheme is the one earlier described and the dynamic management is represented by the proposed idea of this PhD thesis. For the static scheme two campaigns of simulations have been carried out, changing the allowed level of bandwidth: in the "*Static Min*" policy only the lowest level of bandwidth is assigned to users (we chosen l_{MIN}=512Kbps), while in the "*Static Max*" policy only the highest level of bandwidth is assigned to users (we chosen

l_{MAX}=896Kbps). In the "*Dynamic 1*" scheme the effects of channel degradations are not considered, hypothesizing that the channel is always in the "best state"; in the "*Dynamic 2*" scheme the proposed model in chapter 1 is introduced. Four bandwidth levels are considered $B=\{512, 640, 768, 896\}$Kbps and the utility function is simply a piece-wise curve with values $U=\{1, 2, 3, 4\}$. The outage threshold p_{OUTAGE} has been fixed to 0.05.

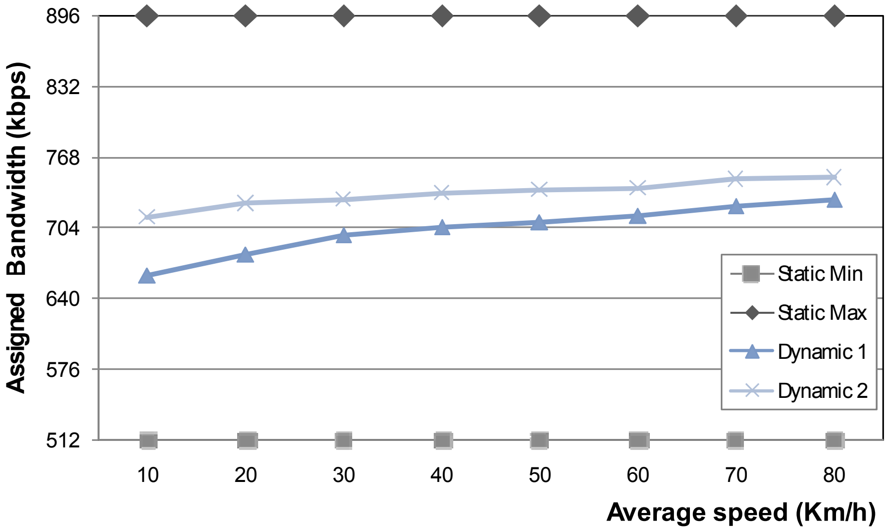


Figure 3.9. Average received bandwidth from MDP users.

Figure 3.9 shows the assigned bandwidth to MDP users versus their average speed: for the "*Static*" cases, neglecting the effects of the wireless link and packet errors, the average assigned bandwidth is the same of the input chosen level and no other considerations are needed (no reallocations are made, so each AP always gives to the admitted user the same bandwidth of its requests). For the dynamic cases, mobility effects are evident, due to the slight increasing trend for higher speeds: the time spent in a cell for each user goes decreasing and a higher bandwidth availability is offered. If link behavior is taken into account, there is an enhancement in the received bandwidth, because the system reacts to link quality variations. Nevertheless, it can be observed that the "*Dynamic 2*" scheme performs better in terms of bandwidth performance, because the results of the "*Static Max*" scheme can be disregarded, since it is only an ideal case. It must be outlined that only in the "*Dynamic 2*" case the depicted bandwidth is the one perceived by user.

Figure 3.10 illustrates the utility values for different average speeds. The same trends of the previous figure are obtained, so similar considerations can be made: obviously in the two "*Static*" schemes a constant trend is obtained, because the bandwidth maintains itself at the same (minimum or maximum) level and the considered utility function is directly proportional to its values. This time, in the "*Dynamic 2*" case, perceived utility is lower, because the received bandwidth is affected by channel degradation and the instant received utility follows the behavior of eq. (3.5), because the utility function is evaluated on the real bandwidth received by MDP users. In figure 3.10 the utility is evaluated on the assigned bandwidth level, since no channel degradations are considered; only in the "*Dynamic 2*" case, the considered utility is the real perceived utility (channel degradations are taken into account).

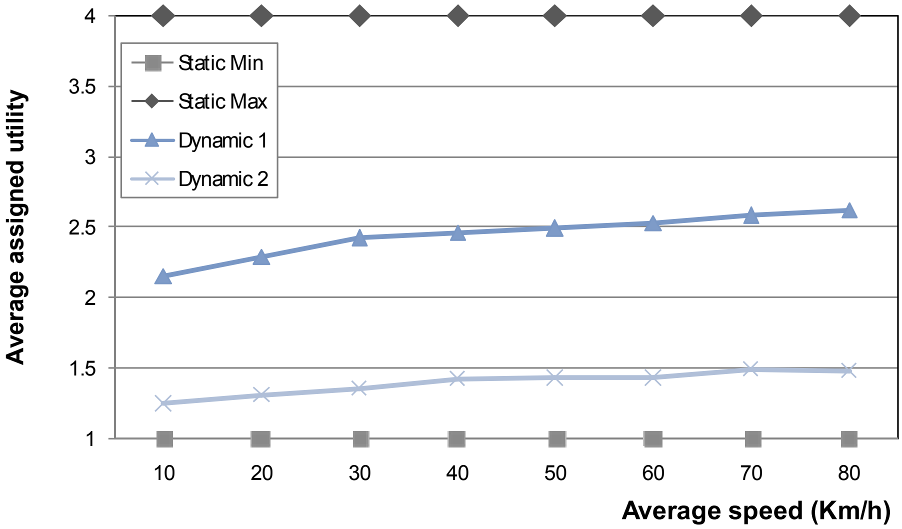


Figure 3.10. Average perceived utility from MDP users.

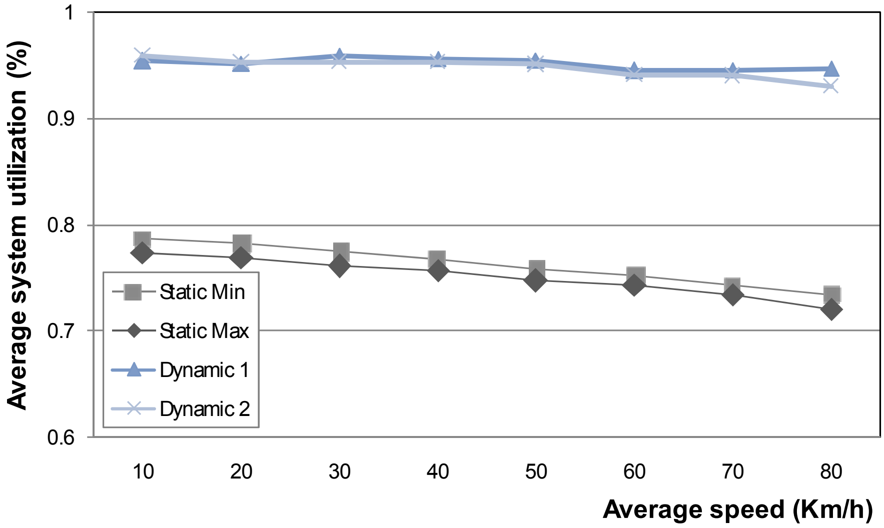


Figure 3.11. Average system utilization.

The main differences among static and dynamic schemes are evident when considering the average system utilization: figure 3.11 shows the enhancements introduced by the dynamic scheme. In the dynamic case (1 or 2) the number of admitted flows increases because the system is able to adapt itself to traffic conditions, dynamically reallocating the assigned bandwidth; in this way, respecting the chosen outage threshold, the system can admit a higher number of users by degrading the bandwidth of existing users to lower permitted levels; in addition, the bandwidth is upgraded when a user leaves the current cell (call termination or handover). For the static case, in addition to the lower obtained values, there is a decreasing trend for increasing speeds: this is due to the higher overhead introduced by the increased number of hand-over events. In the dynamic cases, this phenomenon is not so evident: system utilization cannot reach the maximum value of 100% because of the intrinsic protocol overhead, but there is not the decreasing trend for higher average speeds because, when a user leaves/enters a cell, the AP is able to react immediately to the new system conditions.

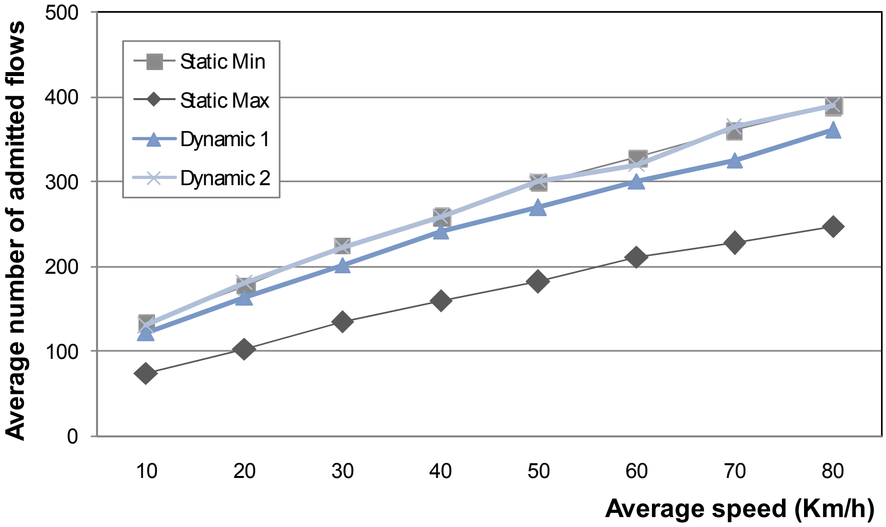


Figure 3.12. Average number of admitted MDP flows.

Figure 3.12 depicts the average number of admitted MDP flows for different average speeds values; as it can be expected, it increases for higher average speeds because of the higher bandwidth availability: increasing the average speed the cell sojourn time decreases so more users can enter into the system. The difference between "Static Min" and "Static Max" is due to the difference of the assigned bandwidth level: when the highest bandwidth level is assigned there are less available resources for admitting new flows (a similar treatment can be made when the lowest bandwidth level is assigned). For the dynamic cases, the number of admitted flows is higher than the "*Static Max*" case because a lower amount of resources are assigned to the users on the average; lower performance are obtained in comparison with the ideal

"*Static Min*" case, because no lower resources than the "*Static Min*" case can be assigned to users; the "*Dynamic 2*" case, concluding, offers some performance that are comparable with those of the ideal "*Static Min*" scheme, in terms of admitted flows.

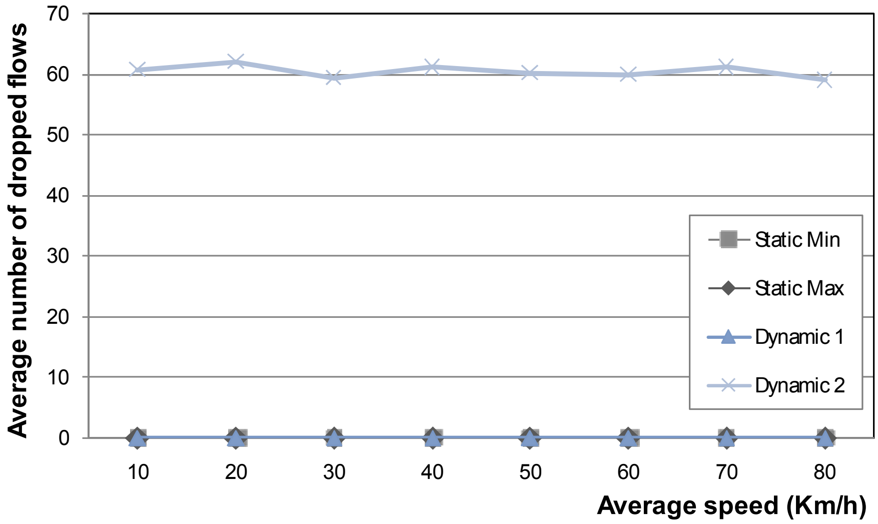


Figure 3.13. Average number of dropped MDP flows.

Figure 3.13 shows the average number of dropped MDP flows for different average speeds: in the first three cases ("*Static Min*", "*Static Max*" and "*Dynamic 1*"), there are no dropped flows because the schemes do not account for channel conditions (it can be said that the wireless link has no degradations); when the wireless link behavior is considered ("*Dynamic 2*" case), there are some dropped flows (from 41 to 43, with a mean of 42, for different average speeds) against an average number of service requests of 1080 and an admitted flows trend as in figure 3.12; for lower speeds the percentage of dropped flows (the ratio between dropped and admitted) is higher: the cell sojourn time of a single flows is higher, then there is a higher probability to be affected by bandwidth reallocations.

Another concept must be underlined: the wireless link modeling of chapter 1 needs some attention; it is necessary to capture the real nature (with some approximations, obviously) of wireless links, under multi-path fading conditions. Many works in literature face the problem of bandwidth reallocations and call admission control, maximizing some performance parameters and introducing some quality indexes, but most of them do not account for channel conditions. The following figure shows how outage events (the utility or the bandwidth below the minimum acceptable level) can happen if there are no high-level mechanisms that consider the degradation level of the channel are introduced. Figure 3.14 shows the same curves of figure 3.9 for the "*Static Min*" and "*Static Max*" cases, introducing the real bandwidth received by MDP users, with the degradations introduced by the wireless links.

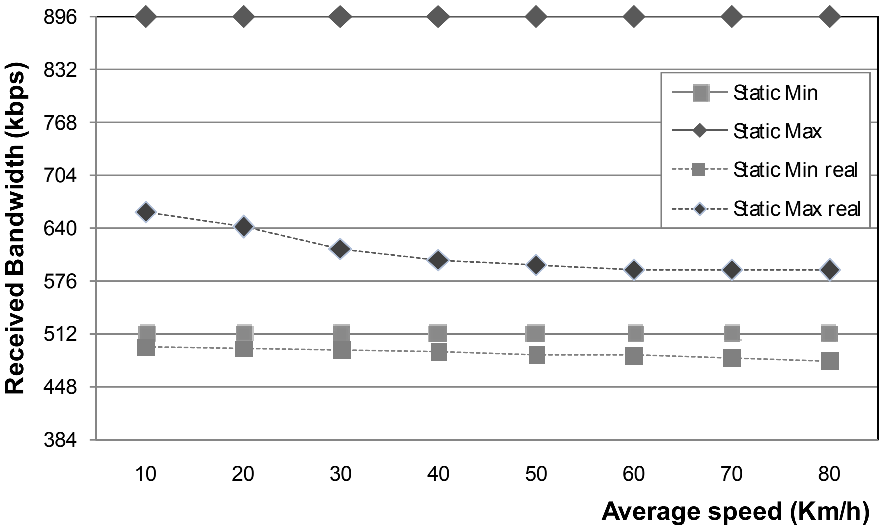


Figure 3.14. Average assigned/received bandwidth to/for MDP users.

Continuous curves are the same of figure 3.9, while dashed ones represent the effective bandwidth received by mobile hosts: in the "*Static Min*" scheme, the real received bandwidth is lower than the minimum admitted level, while in the "*Static Max*" case it is lower than the assigned level. The course is slightly decreasing because of the higher degradations for higher speeds (a higher Doppler shift is introduced). It is obvious that, due to the intrinsic BER of the wireless link, the received bandwidth level in never the same of the assigned one, so considering the absence of ideality in wireless connections is mandatory, to avoid mistakes in real system dimensioning (low-level correction mechanisms, such as FEC, Interleaving, Scrambling and/or Coding, are not always able to avoid/correct the signal corruption introduced by wireless links).

3.7 Conclusions on chapter 3

In this chapter some new considerations about the adaptivity in wireless systems have been made. The concept of utility function has been introduced, as an indicator of the user satisfaction level. If the perceived utility must be maximized, then there must be a way to describe how a user is satisfied when the received bandwidth level varies. Utility functions are useful to solve such kind of problems and it has been shown that there are many works in literature that describe the best trends of utility functions, appropriate for the specific application (tolerant, intolerant, etc.). After a description of some important concepts of utility functions, a new bandwidth allocation protocol has been introduced, with the aim of having a new scheme that can be applicable in the ISPNs systems. Since different applications can be introduced in an ISPN system, the proposed idea takes care of considering the specific utility function, as well as the wireless channel modelling. In this way some important goals can be reached: fairness among users belonging to the same class; high system utilization and QoS guarantees. A complexity analysis has been given for the UB CAC and BAG schemes and the importance of a dynamic bandwidth allocation scheme has been demonstrated, through an addicted campaign of simulations. The obtained results have shown that the introduction of a dynamic scheme for bandwidth management increases system performance, in terms of utilization and number of admitted flows. In addition, it has been shown that the introduction of a channel model is mandatory if channel degradations must be taken into account when dimensioning a wireless system or while serving MDP requests.

Chapter 4 – A new direction-based mobility prediction algorithm

4.1 Introduction

In this chapter, all the previous acquired concepts are used to introduce a novel prediction algorithm: by using the MRSVP (chapter 2, paragraph 2.5) the exchange of the MSPEC message is necessary to build up passive reservations from the current location of the generic mobile host on the remote proxy agent. As illustrated in chapter 2, the MSPEC message is used in the Proxy Discovery phase, when the current AP (where the call has originated) must communicate to mobile host the remote addresses of its proxy agents. In previous chapters, the assumption of knowing Remote Proxy Agents (RPAs) addresses has been made; now, a new algorithm to determine a possible set of RPAs is introduced, after a deep and accurate mobility analysis. This chapter is structured as follows: first of all, an illustration of the prediction problem in wireless environments is given with an overview of the existing works; then, a preliminary mobility study and analysis in a 1D environment is made, under a 1D Random WayPoint Mobility Model (RWPMM); it is extended to a 2D environment and another mobility model is also introduced [77], in order to give more effectiveness to users movements. The prediction algorithm is introduced and some performance evaluations are given. In the last chapter (fifth chapter) the overall system is evaluated (the MRSVP is integrated with the UB CAC and BAG as in chapter 3 and the predicting policy is also introduced).

4.2 State of the art about mobility management and prediction

In the last decade, mobile computing has been extensively studied in both the computer and communication communities. However, most of the recent studies focus on the network layer communication protocols and few of them on the mobility aspects caused by mobile hosts' behavior. The concept of mobility management has been used in cellular mobile networks for many years, but it was primarily designed to support mobile voice communication. Traditionally, mobility management includes functions to passively keep track of the location of users terminals and to maintain connections to the terminals belonging to the system. We argue that to maintain uninterrupted high-quality service for distributed applications in a mobile environment, data routing should not only passively reflect the change of terminal location, but also aggressively anticipate the behavior of mobile users. In addition, as discussed in chapter 2, mobility prediction is also important when passive reservations need to be made. In [78], an aggressive mobility management scheme is proposed: a set of Mobile Motion Prediction (MMP) algorithms is used to predict the "future" location of a mobile user according to user's movement history patterns. The data or services are pre-connected and pre-assigned at the new location before user moves into the new location. Thus, the user can immediately receive service or data with virtually the same efficiency as at the previous location. The proposed idea is based on the fact that every user has some degree of regularity in its movements, that is to say people movements consists of random movements and regular movements and the most of mobile users have some regular daily (hourly, weekly, etc.) movement-patterns. The authors of [78] analyze the goodness of the proposed scheme in terms of prediction accuracy rate versus mobility randomness factor: obviously the prediction is more accurate when the mobility randomness is lower.

In [79] a Mobility-Dependent Predictive Resource Reservation (MDPRR) scheme is proposed to provide flexible usage of limited resource in mobile multimedia wireless networks. An admission control scheme is also considered to further guarantee the QoS of real-time traffic. The coverage area of a cell is divided into non-hand-off, pre-hand-off and hand-off zones, so that bandwidth is reserved in the target/sub-target cell as mobile stations move into the pre-handoff zone and leave the serving base station. The amount of bandwidth to be reserved is dynamically adjusted, according to the location, the instantaneous variation of velocity and direction of mobile stations. Prediction performances are evaluated in terms of call dropping or blocking probability versus calls arrival rate, system utilization and bandwidth assignment.

The guard channel scheme is generally referred as the Fixed Bandwidth Reservation (FBR) scheme [80] which can improve the dropping probability of handoff connections by reserving a fixed number of channels exclusively for handoff connections. The drawback of this scheme is that the reserved bandwidth is often wasted in the hot spot area. Dynamically reserving bandwidth for handoff calls is an effective way to reduce hand-off dropping probability and increase bandwidth utilization.

Existing approaches ([81], [82]) of dynamic reservation for hand-off connections can be classified into probabilistic [81] and dynamic [82] reservation. The resource estimation algorithm using the shadow cluster concept [82] is a virtual message system where information about position or movement pattern are exchanged with neighboring cells. However, these schemes induce large amount of overheads between APs. In the Predictive Channel Reservation (PCR) scheme [82], reservation requests are sent to neighboring cells by extrapolating the motion of Mobile hosts. The PCR scheme can provide real-time hand-off predictions, however, multimedia traffic was not considered.

In [83] the authors propose a new prediction scheme based on User Mobility Profile (UMP, a combination of historic records and predictive patterns of mobile terminals), which serve as fundamental information for mobility management and enhancement of QoS in wireless multimedia networks. A UMP framework is developed for estimating service patterns and tracking mobile users, including descriptions of location, mobility and service requirements. For each mobile user, the service requirement is estimated using a mean-square error method. In particular, an adaptive algorithm is designed to predict future positions of mobile terminals in terms of location probabilities, based on moving directions and residence time in a cell. The performance of the new UMP proposed scheme are evaluated through some simulation results, which have shown that the proposed schemes correctly manages mobility and resources (blocking/dropping probabilities and location tracking costs are evaluated).

Through estimation of mobile users' trajectory and arrival/departure times in [84], a group of future cells is determined: it constitutes the most likely cluster into which a terminal will move.

In this PhD thesis a novel approach to predict the more suitable set of "future possibly visited" cells is proposed: the most of the previous illustrated schemes (as well as other works in literature) makes only a dynamic one step prediction, so passive reservations can be made step by step, during host movements; these schemes reduce the overhead of the system, but do not ensure a “dropping-free” system, because when the next cell is determined, there are no guarantees about the bandwidth availability in that cell. The algorithm of [83] shows the best results if compared with the other ones. Now the proposed idea is illustrated in different steps: first of all a generic 1D treatment is given, showing how a quantitative description of the mobility behavior can be given; then the prediction scheme is proposed and finally some extensions are added to it. The complete system is analyzed in chapter 5.

4.3 Random WayPoint Mobility Model (RWPMM)

The movement pattern of users plays an important role in performance analysis of mobile and wireless networks. In cellular networks, for example, a user's mobility behaviour directly affects the signalling traffic needed for hand-over and location management. The extra signalling messages over the air interface consume radio resources and increase the associated database query load. The modelling of movement is thus an essential building block in analytical and simulation-based studies of these systems. Mobility models are needed in the design of strategies for location updating and paging, radio resource management (e.g., dynamic channel allocation schemes) and technical network planning and design (e.g., cell and location area layout, network dimensioning). The choice of the mobility model has a significant effect on the obtained results. If the model is unrealistic, invalid conclusions may be drawn. With the increasing number of subscribers and the decreasing cell size in future cellular systems, the mobility pattern of users will even more influence the performance of the network. Models that proved to be a good choice in simulation of macro-cellular environments show some drawbacks when being applied in micro- and pico-cellular environments. Mobility modelling also plays an important role in analysis of algorithms and protocols in wireless local area networks (WLANs) and self-organizing wireless ad hoc networks. Whereas in cellular networks there exists a number of approaches that model the macroscopic movement behaviour of users (e.g., random walk from cell to cell, description of the cell residence time), in these cases we need a "microscopic" model.

Now, a brief overview on the RWPMM will be given (more details are contained in [68]). The process representing the movement of a node within a convex area $A \subseteq R^2$

according to the RWPMM can be described as follows. Initially, the node is placed at the point P_1, chosen from a uniform distribution over A. Then, a destination point (also called waypoint) P_2 is chosen from a uniform distribution over A and the node moves along a straight line from P_1 to P_2 with constant velocity V_1 drawn independently of the location from a velocity distribution with pdf $f_V(v)$. Once the node reaches P_2, a new destination point P_3 is drawn independently from a uniform distribution over A and velocity V_2 is drawn from $f_V(v)$ independently of the location and V_1. The node again moves at constant velocity V_2 to the point P_3, then the process repeats. Formally, the RWP process is defined as an infinite sequence of triples [92], $(P_0, P_1, V_1), (P_1, P_2, V_2), (P_2, P_3, V_3), \ldots$ as illustrated in figure 4.1.

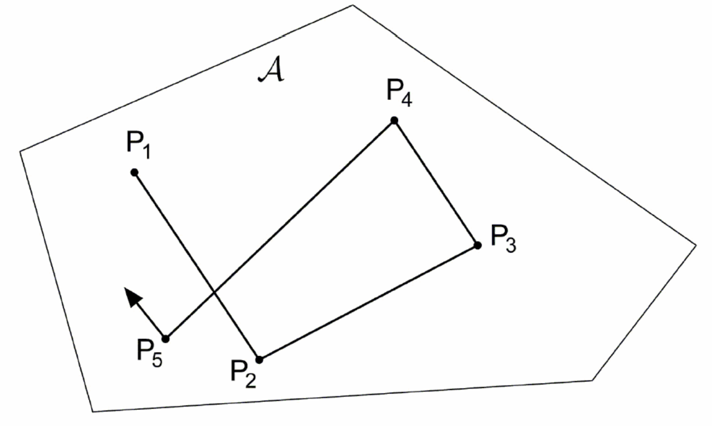


Figure 4.1. Typical RWPMM segments (legs).

Thus, the path of a node consists of straight line segments, called legs, defined by a sequence of independently and uniformly distributed waypoints $\{P_i\}$ in a convex set $A \subseteq R^2$. Furthermore, on each leg (P_{i-1}, P_i) the node velocity V_i is an i.i.d. random variable, independent of the node location having the pdf $f_V(v)$. It is also possible to extend the model by defining random pause times (i.i.d. random variables) at the waypoints. The influence of this generalization on the node distribution can be analyzed in a rather straight forward manner as the process consists of two independent and alternating modes, mobile and stagnant [90].

Let the random variable X denote the location of a waypoint P. The waypoints are uniformly and independently distributed over A, i.e. the probability density function (pdf) of X is:

$$g(r) = \begin{cases} \dfrac{1}{A}, & r \in A \\ 0, & else, \end{cases} \qquad (4.1)$$

where A denotes the area of the set $A \subseteq R^2$. This uniform distribution is denoted by U(A), so X ~ U(A). The random variable representing the location of the node at an arbitrary point of time is denoted by *R* and its pdf by f(r). Note that two consecutive legs in the RWP process share a common waypoint and thus are not independent. However, many properties of the RWP process can be analyzed by studying the corresponding independent leg process, where the legs are, as the name suggests, independent and identically distributed. Formally, for a given RWP process the corresponding independent leg process can be obtained by considering, e.g., every second leg (for details refer to [90]).

Velocities can be taken from a uniform distribution in the range $[v_{min}; v_{max}]$, or from any distribution (e.g., the beta distribution or a discrete distribution, as in [90]). Given the pdf $f_V(v)$ from which the velocities at the waypoints are drawn, the stationary distribution of the velocity for a node moving according to the RWP model is given by:

$$\frac{1}{v} f_V(v), \tag{4.2}$$

up to a normalization constant. This is because the time spent on a leg is proportional to $1/v$, and V and X are independent. From this, it is also obvious that for $f_V(v)=U[v_{min}; v_{max}]$ the stationary distribution is only defined for $v_{min} > 0$. Hence, letting $v_{min} = 0$ implies that stationarity is never reached or, more precisely, in the stationary state all the nodes are stopped, as pointed out by Yoon et al. in [91] and later by others in [90]. Finally, note that the stationary distribution of the location of a node and the stationary node velocity distribution are independent of each other, as has been formally shown in [92].

4.4 Cell Stay Time (CST) analysis and prediction in a simplified 1D scenario

Let us hypothesize for now that users can move only along one direction, following a given mobility model (RWPMM [68], [90], SRMM [77], etc.), adapted to one dimension: in this way, only quantitative analysis must be made about the number of cells that user will visit. Directional treatments will be added in next paragraphs.

In order to simplify the previous exposed mobility model, the convex region becomes a 1D space, so $A \subseteq R$. The same considerations can be made, regarding legs and waypoints, but now they belong to a single direction, that is to say cell diameter D, as depicted in figure 4.2. In eq. 4.1 the area of *A* is substituted by the diameter length *D*.

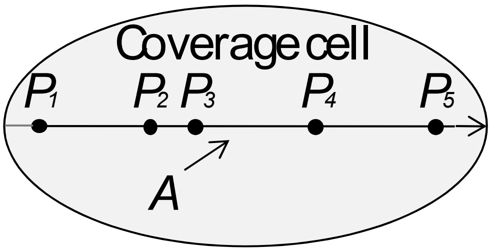


Figure 4.2. Typical RWPMM legs in a 1D environment.

As exposed in previous paragraphs, passive reservations represent a good way to guarantee and maintain QoS to MIP (see chapter 2 for the difference between MIP and MDP) users during hand-off events; this policy is based on "in-advance making" bandwidth reservations over all the cells in the network, without evaluating neither average host's speed, nor Call Holding Time (CHT). In this section we propose a new criterion for increase system utilization, while maintaining pre-reservation policy, improving the performances of WLANs system. This time, passive reservations are made after evaluating average hosts' speed and calls duration: if the host moves in the cells with extremely slow speed, pre-reserving over all system APs is not necessary, because the user will probably never visit all of them. Moreover, if two mobile hosts have the same speed, the number of visited cells may vary in function of calls duration.

4.4.1 "Monitor" simulations and CST distribution evaluation

Our proposed algorithm starts with the estimation of the average Cell Stay Time (CST) distribution: for this purpose, once the RWPMM has been implemented in the simulator (details about it will be given in chapter 5), lots of "monitor simulations" (simulations dedicated to the observation of some system parameters, without any performance evaluation) have been carried out: each run gives a sample point of the average CST; collecting many samples (around a thousand) the average CST distribution can be obtained (the plot of the probability density function has been made in MATLAB) and this information has been used to evaluate the number of cells visited by mobile hosts, in order to make passive reservations only on the cells that a host will effectively visit, leaving the bandwidth availability in the other ones. In our studies, a Poisson arrival time distribution [85] and an exponentially distributed CHT [86] have been considered for any mobile host (these assumptions are the classic ones that are made in the most of literature works). Coverages are represented by circular areas of radius *R* and the cell overlapping *s* has been fixed to 10%.

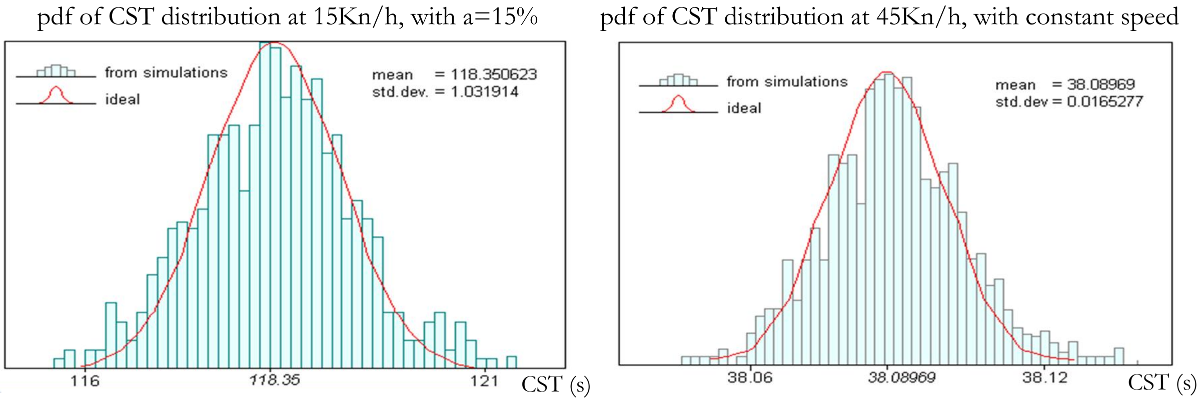
In the first campaign of simulations, users' speed has been fixed to a constant value v_{avg}; that is to say all users move along coverage areas with the same speed, so they only have different connection times; this kind of simulations has been employed in order to evaluate the model validity under simplified conditions (constant speed); as shown in next paragraphs, in this case, low prediction errors are suffered.

In a second campaign of simulations, average speed has been uniformly selected in a range of $[v_{avg} - a, v_{avg} + a]$, where *a* is a variable percentage of v_{avg}, (its values may go, for example, from 0.05 to 0.5); as shown in next paragraphs, prediction error increases for high values of *a*, but it maintains itself in acceptable bounds.

In both cases (constant or variable average speed), for any fixed value of v_{avg}, a Probability Density Function (pdf) of the average CST of mobile hosts has been derived in order to make a predictive evaluation of visited cells.

CST is an important parameter in wireless networks because it permits the evaluation of how long a user will stay in a cell during its call holding time and how many cells he will visit. This can be useful in resource reservations, for an environment where node mobility is supported. In this work a further contribution is given to CST evaluation, through an explicit math formula that binds the average speed of the mobile user, the variance around the average speed and the cell diameter to the CST estimation.

With the availability of a thousand of CST samples, the distribution can be well observed and, after a results analysis, a CST distribution has been obtained, like the ones depicted in figure 4.3, with a Gaussian approximation for fixed values of speed and variation.


pdf of CST distribution at 15Kn/h, with a=15%

pdf of CST distribution at 45Kn/h, with constant speed

Figure 4.3. Different pdfs for CST distribution, under 1D RWPMM.

So the general expression of the CST pdf is:

$$f_{X_{CST}}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad , \tag{4.3}$$

where $\mu = \mu_{CST}(v,\alpha,R)$ and $\sigma = \sigma_{CST}(v,\alpha,R)$ are respectively the average and standard deviation of the Gaussian distribution. It is possible to evaluate the error of considered CST and to make a cell stay time prediction based on confidence intervals and confidence levels, considering the worst case Cell Outage Probability (COP). It is possible to select a CST T_{CST} for a mobile host so that:

$$Prob(X< T_{CST}) <1\text{-}COP, \tag{4.4}$$

where X is normally distributed. T_{CST} is called a (1-COP)*100% upper confidence bound for X. If the average CHT T_{CHT} is known, it is possible to consider the term C_p (C partial) as:

$$C_p = \left\lceil \frac{T_{CHT}}{T_{CST}} \right\rceil . \tag{4.5}$$

So it is possible to use the C_p value to make the pre-reservation of MIP flows in order to leave more bandwidth availability in the not visited cells.

The assumption of the normally distributed CST, with different means and standard deviations depending on the fixed mobility and system parameters, has been verified through the Kolmogorov-Smirnov (KS) normality test [87]; different *p-values* (a *p-value* for a comparison test represents the likelihood, under the assumption that the null hypothesis is true, that the data would yield the obtained results) have been obtained, showing that there is a negligible error if a Gaussian approximation is employed for the CST distribution.

The Cumulative Distribution Function (cdf) of the CST from eq. (4.3) is:

$$F(x) = P(X_{CST} \le x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt \quad , \qquad (4.6)$$

and the probability that CST is lower than a value *x* with a fixed error threshold ε is given from eq. (4.7):

$$P(X_{CST} \le x) = P\left(Z \le \frac{x-\mu_{CST}}{\sigma_{CST}}\right) = \Phi\left(\frac{x-\mu_{CST}}{\sigma_{CST}}\right) = 1-\varepsilon , \qquad (4.7)$$

Where $z = \frac{X_{CST}-\mu_{CST}}{\sigma_{CST}}$, X_{CST} are random variables and $\Phi(\cdot)$ is the standard Gaussian distribution.

After the normalization of X_{CST}, through the tabular values of the Standard Normal distribution, it is possible to obtain the CST estimation for a given threshold ε such as referred in [88], [89]. Thus the knowledge of μ_{CST} and σ_{CST} is necessary to obtain a good estimation of CST, depending on *v*, α and *R* ([97] – [101], [105]).

A further contribution to the evaluation of CST is also given, through an explicit math formula that binds the average speed of the mobile user *v*, the variance around the average speed *a* and the cell diameter *D=2R*. Figure 4.4 shows the CST distribution parameters for different α values (*D* has been fixed to 500m and *v* to 15Km/h).

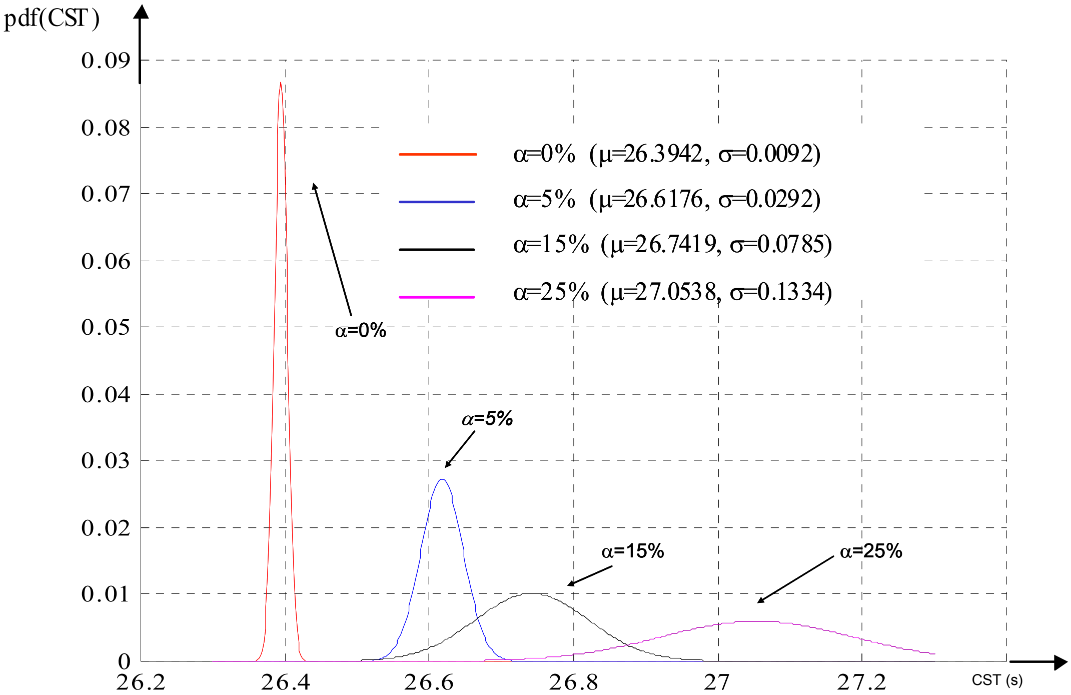


Figure 4.4. An example of CST Gaussian distribution and its approximation for ν = 15Km/h and R=250m.

From figure 4.4 it can be observed that both *μ* and *σ* increase for higher values of α, because of the higher fluctuations of chosen speeds during hosts movements. Table 4.5 summarizes the obtained p-values for different values of α, with *D*=300 and *v*=15Km/h.

α	μ_{CST}	σ_{CST}	**KS p-value**
0	68.39245	0.055022	0.5305
5	70.76407	0.128481	0.5087
15	70.9836	0.568966	0.6712
25	71.3662	0.829614	0.7012

Table 4.5. Values of *μ* and *σ* of CST distributions and K-S *p-values* for different mobility parameters.

4.4.2 CST polynomial regression in the 3D space

In order to have a unique formula to obtain a CST evaluation when system and mobility parameters are given, a regression analysis was performed under MATLAB application and the minimum observed value of the determination coefficient R² over all obtained polynomial functions is 0.9898 (a determination coefficient for a regression is R²=*1-e*, where *e* is the mean square error over all the sampled points). The considered values of the three variables are: [5, 15, 25, 35, 45, 55, 65, 75] for *v* (km/h), [5, 10, 15, 20, 25, 30, 35, 40, 45, 50] for α (% of variation around *v*) and [150, 175, 200, 225, 250, 275, 300] for *R* (m). In this way, the risk of a possible overfitting is avoided. The curves of the average μ_{CST} and σ_{CST} for different system parameters values are shown in the following figures. Details on polynomial regression technique can be found in [87].

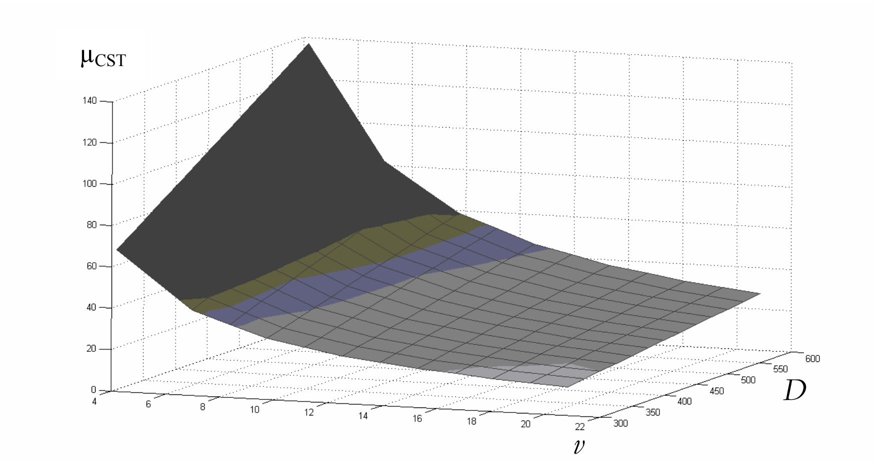


Figure 4.6. Values of μ_{CST} versus *v* (m/s) and *D* (m).

Figure 4.6 shows the relationship between μ_{CST}, users' average speed *v* and system cells' diameter *D*: maintaining *D* at a constant value, the curve has a decreasing course for increasing values of ν, because of the lower duration of users' average permanence in a cell; moreover, as can be expected, fixing a value for the users' average speed *v*, the average CST (μ_{CST}) always increases, for higher values of cell dimension; that is, a mobile host takes more time to fully cross a coverage area. In figure 4.6, fixing v, σ_{CST} increases for higher D values.

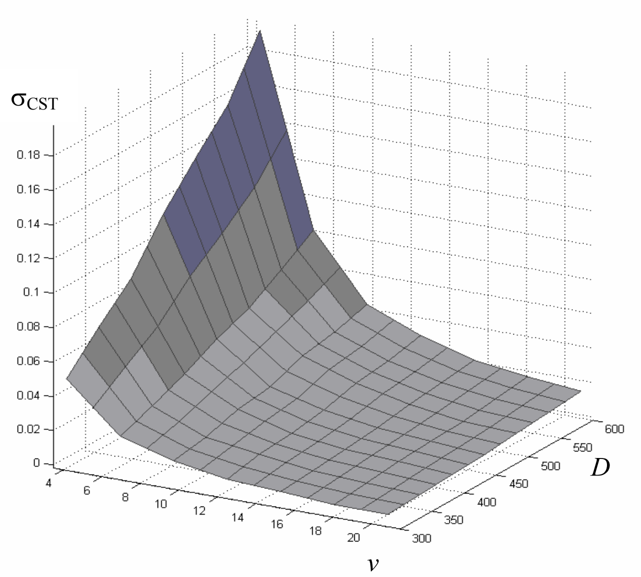


Figure 4.7. Values of σ_{CST} versus v (m/s) and D (m).

In figure 4.7, fixing D, σ_{CST} decreases for higher v values, while increases if v is fixed and D is increased. The standard deviation decreases for higher average speed values because the average CST decreases and there is less time to change speed around the mean value. In addition it increases if cell radius (so the diameter) is enlarged, because of the higher time spent in a cell.

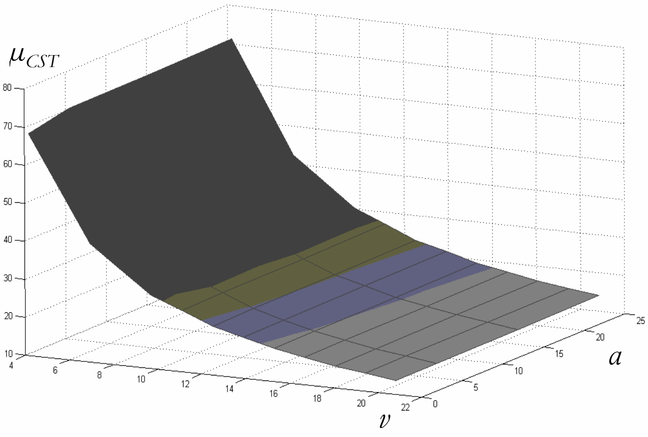


Figure 4.8. Values of μ_{CST} versus v (m/s) and α.

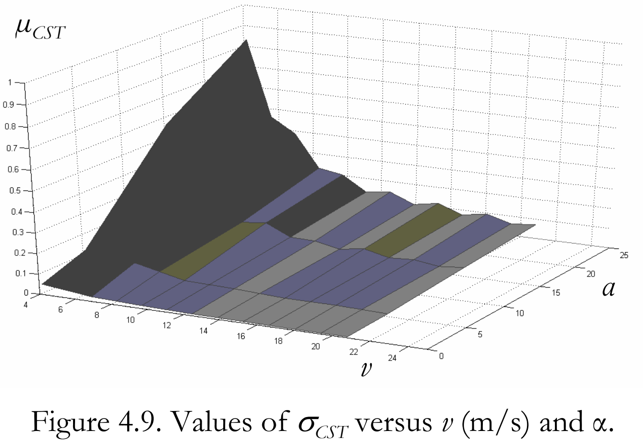


Figure 4.9. Values of σ_{CST} versus v (m/s) and α.

Figure 4.8 illustrates the course of μ_{CST} versus v and a. Maintaining a at a constant value, the curve decreases for increasing values of v, because of the same reason explained previously; in addition, fixing a value for users' average speed v, the average CST μ_{CST} increases slightly, for higher values of a; that is because the chosen speed values can vary in a larger range and the CST random variable has a higher standard deviation σ_{CST}. On the other hand, figure 4.9 shows that σ_{CST} also increases for higher values of a, because the higher probability of a speed change.

A complete analysis of how the CST distribution varies in function of system parameters has been given [105]; now, with the help of polynomial regression, two closed formulas for μ_{CST} and σ_{CST} are derived, valid on the admissible region composed by the considered values of D, v and a, as illustrated in figure 4.10.

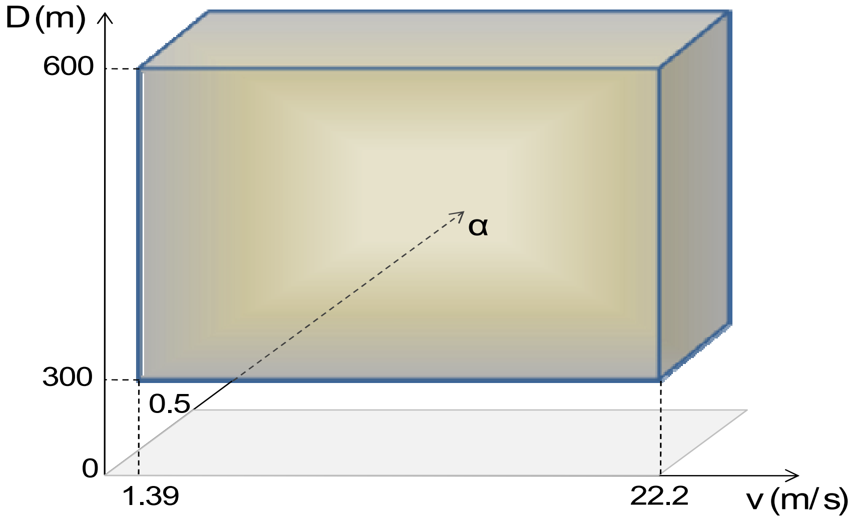


Figure 4.10. Admissible region of polynomial regression.

The general equation of the average of the CST is expressed in eq. (4.8) with a fourth order polynomial regression:

$$\mu_{CST}\ (v) = n_4v^4 + n_3v^3 + n_2v^2 + n_1v + n_0, \tag{4.8}$$

where $n_i = f_i(D,a)$ with i=0, 1, ..,5 and D=2R. Since the n_i terms follow a linear dependence on D, they can be expressed as:

$$n_i = a_iD + b_i \tag{4.9}$$

μCST can be represented in the following way:

$$\mu_{CST}(\bar{v}) = [n_0, n_1, K, n_4] \cdot [1 \quad \bar{v} \quad \bar{v}^2 \quad \bar{v}^3 \quad \bar{v}^4]^T = \langle n \rangle^T \cdot \langle \bar{v} \rangle^{n=4}, \tag{4.10}$$

where the notation $\langle \cdot \rangle$ is used to represent a column vector and $\langle \cdot \rangle^{\mathrm{T}}$ is the transpose operator applied to the vector. In eq. (4.10) $\langle v \rangle^i = [1 \quad \bar{v} \quad \ldots \quad \bar{v}^i]$ is a $(i+1)x1$ vector. In order to calculate the coefficient n_i it is important to evaluate coefficients a_i and b_i. After a second order regression between a_i and a, the following expression can be obtained:

$$a_i = m_{i1}\alpha^2 + m_{i2}\alpha + m_{i3}, \tag{4.11}$$

with i= 0, . . . , 5.

Considering another polynomial regression analysis (5-th order in this case) on the b_i coefficients for different α values around the average speed of the mobile nodes, the expression of the *b* terms can be represented as follows:

$$b_i = c_{i1}\alpha^5 + c_{i2}\alpha^4 + c_{i3}\alpha^3 + c_{i4}\alpha^2 + c_{i5}\alpha + c_{i6}, \qquad (4.12)$$

with i= 0, . . . , 5.

Using a matrix notation, the terms a_i can be calculated as follows:

$$A = M \cdot \langle \alpha \rangle^{n=2}, \qquad (4.13)$$

where A is a (5x1) vector, M is a (5x3) matrix and $\langle \alpha \rangle^{n=2}$ is a (3x1) vector. Terms b_i, instead, are represented in the following form:

$$B = C \cdot \langle \alpha \rangle^{n=5}, \qquad (4.14)$$

where B is a (5x1) vector, C is a (5x6) matrix and $\langle \alpha \rangle^{n=5}$ is a (6x1) vector. Thus, the final matrix expression of the average CST is the following:

$$A = \left[\left(M \cdot \langle \alpha \rangle^{n=2}\right) \cdot D + C \cdot \langle \alpha \rangle^{n=5}\right] \cdot \langle v \rangle^{n=4}. \qquad (4.15)$$

The coefficients of the matrixes M and C are expressed in figure 4.11.

$$M = \begin{pmatrix} -2.4 \cdot 10^{-9} & 5.1 \cdot 10^{-8} & 5.4 \cdot 10^{-6} \\ 1.5 \cdot 10^{-7} & -3.2 \cdot 10^{-6} & -3.5 \cdot 10^{-4} \\ -3.3 \cdot 10^{-6} & 7.4 \cdot 10^{-5} & 8.5 \cdot 10^{-3} \\ 3.3 \cdot 10^{-5} & -7.6 \cdot 10^{-4} & -9.7 \cdot 10^{-2} \\ -1.2 \cdot 10^{-4} & 3.1 \cdot 10^{-3} & 5.1 \cdot 10^{-1} \end{pmatrix} \qquad C = \begin{pmatrix} 4.1 \cdot 10^{-10} & -2.2 \cdot 10^{-8} & 3.3 \cdot 10^{-7} & -7.2 \cdot 10^{-7} & -3.7 \cdot 10^{-6} & -3.4 \cdot 10^{-6} \\ -2.2 \cdot 10^{-8} & 1.2 \cdot 10^{-6} & -1.8 \cdot 10^{-5} & 3.5 \cdot 10^{-5} & 2.0 \cdot 10^{-4} & 2.4 \cdot 10^{-4} \\ 4.1 \cdot 10^{-7} & -2.2 \cdot 10^{-5} & 3.3 \cdot 10^{-4} & -5.9 \cdot 10^{-4} & -3.5 \cdot 10^{-3} & -6.4 \cdot 10^{-3} \\ -3.3 \cdot 10^{-6} & 1.8 \cdot 10^{-4} & -2.6 \cdot 10^{-3} & 4.3 \cdot 10^{-3} & 2.4 \cdot 10^{-2} & 7.9 \cdot 10^{-2} \\ 1.0 \cdot 10^{-5} & -5.6 \cdot 10^{-4} & 8.5 \cdot 10^{-3} & -1.8 \cdot 10^{-2} & -3.2 \cdot 10^{-2} & -3.9 \cdot 10^{-1} \end{pmatrix}$$

Figure 4.11. Elements of M and C for polynomial regression on μ_{CST}.

The same regression analysis was carried out for the standard deviation course, on varying mobility parameter α. A first polynomial regression of fourth order of σ_{CST} as function of v has been obtained:

$$\sigma_{CST}(\bar{v}) = m_4\bar{v}^4 + m_3\bar{v}^3 + m_2\bar{v}^2 + m\bar{v} + m_0, \qquad (4.16)$$

thus:

$$\sigma_{CST}(\bar{v}) = [m_0, m_1, \mathrm{K}, m_4] \cdot \left[1 \quad \bar{v} \quad \bar{v}^2 \quad \bar{v}^3 \quad \bar{v}^4\right]^T = \langle m \rangle^T \cdot \langle \bar{v} \rangle^{n=4}. \qquad (4.17)$$

Matrix M' in figure 4.12 summarizes m_i values (on the columns) for different D values (on the rows) with i=0, 1, ..., 4. A second regression of fifth order for different D values for each m_i term with i=1, 2, ..., 4 has been performed:

$$m_i(D) = \beta_{i5}D^5 + \beta_{i4}D^4 + \beta_{i3}D^3 + \beta_{i2}D^2 + \beta_{i1}D + \beta_{i0}, \qquad (4.18)$$

where β_{ij} is the polynomial coefficient with j=0..5 and i=0..4.

$$m_i(D) = [\beta_{i0}, \beta_{i1}, \mathrm{K}, \beta_{i5}] \cdot \left[1 \quad D \quad D^2 \quad D^3 \quad D^4 \quad D^5\right]^T = \langle \beta_i \rangle^T \cdot \langle D \rangle^{n=5}. \qquad (4.19)$$

A third regression of 4th order is associated to the α variable such shown below:

$$\beta_{ij}(\alpha)=\gamma_{ij4}\alpha^4+\gamma_{ij3}\alpha^3+\gamma_{ij2}\alpha^2+\gamma_{ij1}\alpha+\gamma_{ij0}. \qquad (4.20)$$

Thus β is a 5x6 matrix (as depicted in figure 4.12), where each column is given by the following product:

$$\beta_i=[\gamma_{ij}]\cdot[\alpha_0,\alpha_1,\alpha_2,\alpha_3,\alpha_4,]^T \qquad (4.21)$$

$$M'=\begin{pmatrix}
2.20\cdot10^{-6} & -1.36\cdot10^{-4} & 3.15\cdot10^{-3} & -3.32\cdot10^{-2} & 1.44\cdot10^{-1}\\
2.91\cdot10^{-6} & -1.78\cdot10^{-4} & 4.08\cdot10^{-3} & -4.23\cdot10^{-2} & 1.78\cdot10^{-1}\\
3.61\cdot10^{-6} & -2.21\cdot10^{-4} & 5.02\cdot10^{-3} & -5.14\cdot10^{-2} & 2.13\cdot10^{-1}\\
3.62\cdot10^{-6} & -2.25\cdot10^{-4} & 5.23\cdot10^{-3} & -5.47\cdot10^{-2} & 2.32\cdot10^{-1}\\
3.63\cdot10^{-6} & -2.29\cdot10^{-4} & 5.43\cdot10^{-3} & -5.80\cdot10^{-2} & 2.52\cdot10^{-1}\\
4.39\cdot10^{-6} & -2.78\cdot10^{-4} & 6.57\cdot10^{-3} & -6.98\cdot10^{-2} & 2.98\cdot10^{-1}\\
5.15\cdot10^{-6} & -3.26\cdot10^{-4} & 7.72\cdot10^{-3} & -8.16\cdot10^{-2} & 3.45\cdot10^{-1}\\
6.55\cdot10^{-6} & -4.02\cdot10^{-4} & 9.18\cdot10^{-3} & -9.40\cdot10^{-2} & 3.87\cdot10^{-1}\\
7.94\cdot10^{-6} & -4.77\cdot10^{-4} & 1.06\cdot10^{-2} & -1.06\cdot10^{-1} & 4.29\cdot10^{-1}\\
7.53\cdot10^{-6} & -4.61\cdot10^{-4} & 1.05\cdot10^{-2} & -1.08\cdot10^{-1} & 4.48\cdot10^{-1}\\
7.11\cdot10^{-6} & -4.46\cdot10^{-4} & 1.04\cdot10^{-2} & -1.10\cdot10^{-1} & 4.66\cdot10^{-1}\\
7.94\cdot10^{-6} & -4.98\cdot10^{-4} & 1.16\cdot10^{-2} & -1.23\cdot10^{-1} & 5.19\cdot10^{-1}\\
8.77\cdot10^{-6} & -5.51\cdot10^{-4} & 1.29\cdot10^{-2} & -1.35\cdot10^{-1} & 5.71\cdot10^{-1}
\end{pmatrix}$$

$$\beta=\begin{pmatrix}
1.09\cdot10^{-16} & -2.46\cdot10^{-13} & 2.18\cdot10^{-10} & -9.52\cdot10^{-8} & 2.04\cdot10^{-5} & -1.72\cdot10^{-3}\\
-5.81\cdot10^{-15} & 1.30\cdot10^{-11} & -1.15\cdot10^{-8} & 5.02\cdot10^{-6} & -1.07\cdot10^{-3} & 9.08\cdot10^{-2}\\
1.08610^{-13} & 2.43\cdot10^{-10} & 2.14\cdot10^{-7} & -9.32\cdot10^{-5} & 1.99\cdot10^{-2} & -1.68\\
-8.41\cdot10^{-13} & 1.87\cdot10^{-9} & -1.65\cdot10^{-6} & 7.14\cdot10^{-4} & -1.52\cdot10^{-1} & 12.8\\
2.28\cdot10^{-12} & -5.06\cdot10^{-9} & 4.43\cdot10^{-6} & -1.90\cdot10^{-3} & 4.04\cdot10^{-1} & -33.8
\end{pmatrix}$$

Figure 4.12. Elements of M' and β for polynomial regression on σ_{CST}.

In this way, two closed forms for μ_{CST} and σ_{CST} are derived; they can be used in eq. (4.5) to determine the C_p parameter, that is to say the number of predicted cells that user will visit. As already discussed, this kind of technique can be used only by MIP users in a 1D environment, as depicted in figure 4.13.

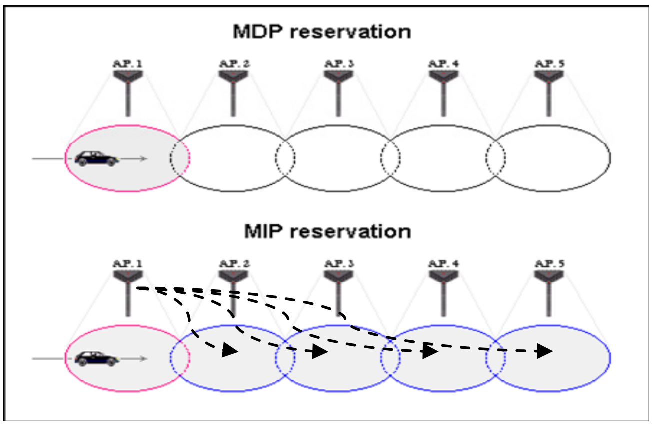


Figure 4.13. Different kind of reservations for MIP and MDP users.

Simulation results on the predictive policy in a 1D scenario will be presented in the next (and last) chapter of this thesis. Let us now introduce the same analysis in a 2D scenario, after the description of the considered mobility models.

4.5 Smooth Random Mobility Model (SRMM)

Now, the prediction algorithm must be extended for a 2D environment. Before the description of the extension of the previous 1D scheme, a new mobility model, called Smooth Random Mobility Model (SRMM [77]), is introduced, because it enhances the previous one (see paragraph 4.3) in terms of relationship between speed and direction of hosts' movements. In this thesis, the same studies that have been led out on the 1D RWPMM have been made on the SRMM, then the 2D extension have been introduced for the prediction algorithm.

In [77], Bettstetter introduces a new mobility model that can be used in simulations of mobile and wireless networks, in which the individual movement behaviour of users is reflected: a combination of principles for direction and speed control that make the movement of users (e.g., pedestrians and cars) smoother and more realistic than in previously known random models is employed. The general term "node" is used to denote any kind of network-enabled device (a pedestrian with his or her mobile terminal or a user or device inside a vehicle).

The SRMM is a random mobility model for movement in two dimensions on a microscopic scale. A new destination is chosen by the direction φ. The speed and direction changes are both probabilistic. The movement of nodes is not bounded by physical structures (such as streets, buildings, etc.) but nodes are allowed to move anywhere in the simulation plane. Furthermore, there is no correlation between different nodes, i.e., effects like "node following" or "group movement" are not modelled. Two stochastic processes are used: one process determines at what time a mobile station changes its speed and the other one determines when the direction is changed. In other words, the random direction model is enhanced with some new features, which make the simulated movement of nodes (cars and pedestrians) more realistic. Many mobility models in the literature consider that the new choice for speed v and direction φ is not correlated to previous values (such as in the RWPMM). This may cause unrealistic movement behaviour with sudden speed changes ($\frac{\partial v(t)}{\partial t} \to \infty$) or sharp turnings (high $\frac{\partial \varphi(t)}{\partial t}$ when v is high). The SRMM includes both autocorrelation features. The speed is changed incrementally by the current acceleration of the mobile user and also the direction change is smooth: once a station is intended to turn, the direction is (in general) changed in several time steps until the new target direction is achieved. This creates a smooth curve rather than a sharp turning.

The modelling of speed behaviour of nodes is based on the use of target speeds (the speed a node intends to achieve) and linear acceleration. A node goes with constant speed v until a new target speed is decided by a random process. The node then accelerates (or decelerates) until this desired speed is achieved (or again a new target speed is chosen in the meantime).

The speed behaviour of a node at time t can therefore be described by three parameters: its current speed $v(t)$, its current acceleration $a(t)$ and its current target speed $v^*(t)$. In addition, three static speed parameters that characterize a certain node class are defined: a maximum speed v_{max}, a set of preferred speeds $\{v_{pref0}, v_{pref1}, ..., v_{prefn}\}$ and the maximum values for acceleration/deceleration.

The maximum speed v_{max} reflects the maximum speed of a node class or the maximum allowed speed in the given scenario; the relationship $0 \leq v(t) \leq v_{max}$ must be verified at any time t. The set of preferred speeds models the fact that the speed distribution of vehicles and pedestrians over time is not uniformly distributed on $[0, v_{max}]$, but both user classes tend to move with certain "travel speeds", most of the time. For example, a car in the city intends to move with the maximum allowed speed v_{max} and also frequently stops at crossings and traffic lights (v=0). The maximum values for acceleration and deceleration reflect the physical speed up and slow down capabilities of a node class. When a simulation starts, at the beginning, all nodes are created with an initial speed $v(t=0)$, which is chosen from a certain speed distribution $p(v)$: it is correctly defined in a such way that the preferred speed values have a high probability and a uniform distribution is assumed on the entire interval $[0, v_{max}]$. For example, if there are three preferred velocities $v_{pref0}=0$, $v_{pref1}=4/7v_{max}$ and $v_{pref2}=v_{max}$, then the distribution will be the one expressed in eq. (4.22).

$$p(v)=\begin{cases} p(v=0)\delta(v) & v=0 \\ p(v=\frac{4}{7}v_{\max})\delta(v-\frac{4}{7}v_{\max}) & v=\frac{4}{7}v_{\max} \\ p(v=v_{\max})\delta(v-v_{\max}) & v=v_{\max} \\ \dfrac{1-p(v_{pref})}{v_{\max}} & 0<v<v_{\max} \\ 0 & else \end{cases} \quad . \tag{4.22}$$

In this way, a non-uniform speeds distribution is obtained, like the one depicted in figure 4.14.

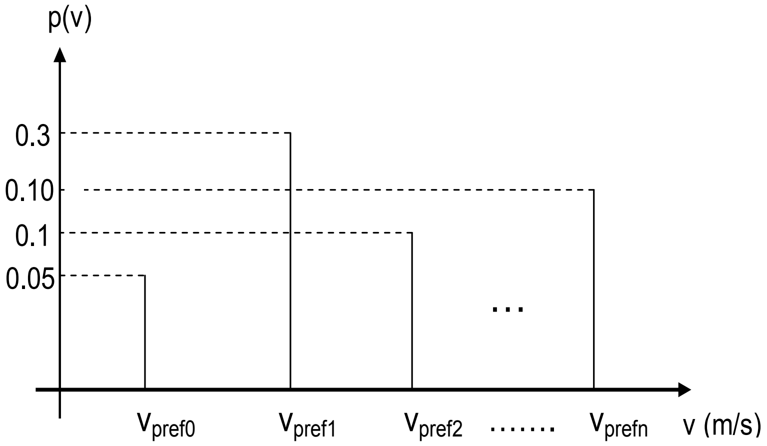


Figure 4.14. Example of probability distribution for *n* preferred speeds.

Remember that $p(v_{pref})=p(v_{pref0})+p(v_{pref1})+ \dots +p(v_{prefn})<1$, $v_{pref0}<v_{pref1}<\dots<v_{prefn}$ and v_{max} is a fixed threshold.

Now the changes over time of speed are described. As mentioned above, a node goes with constant speed v until a speed change event occurs. Upon this event, a new target speed v^* is chosen from the general form of eq. (4.22). The author of [77] modelled the frequency of speed change events according to a Poisson process: in a discrete-time simulation with normalized time $t/\Delta t$, a speed change event occurs with a certain probability p_{v^*} each time step, where $p_{v^*}<<1$. Using continuous time t, the time between two speed change events can be chosen from an exponential distribution with $\lambda=p_{v^*}/\Delta t$:

$$p(t)=\lambda e^{-\lambda t} \; . \tag{4.23}$$

The value of p_{v^*} determines the time between two speed change events. The mean time between two events is $\mu_{v^*}=1/\lambda$. Let t* denote the time at which a speed change event occurs and a new target speed $v^*=v^*(t^*)$ is chosen. Now, an acceleration $a(t^*)\neq 0$ must be set. It is taken from:

$$p(a) = \begin{cases} \dfrac{1}{a_{\max}} & for \quad 0 < a \le a_{\max} \\ 0 & else \end{cases} \tag{4.24}$$

if *v*(t*)>v(t*)*, or from:

$$p(a) = \begin{cases} \dfrac{1}{a_{\min}} & for \quad a_{\min} \le a < 0 \\ 0 & else \end{cases} \tag{4.25}$$

if *v*(t*)<v(t*).*

Clearly, a is set to 0 if *v*(t*)=v(t*).* The term a_{max} is the maximum possible acceleration, and a_{min} is the maximum possible deceleration of this node class. In the following time steps, the speed continuously increases or decreases. Each step, a new speed $v(t)$ is calculated according to $v(t)=v(t-\Delta t)+a(t)\Delta t$ until $v(t)$ achieves $v^*(t)$.

The principle for direction control is similar to the speed control principle. Each node has an initial direction $\varphi(t=0)$ which is chosen from a uniform distribution:

$$p(\varphi) = \frac{1}{2\pi}; \quad 0 \le \varphi < 2\pi. \tag{4.26}$$

A stochastic process decides when to change direction. A node moves in a straight line until a direction change event occurs. This happens with a probability $p_{\varphi^*}<<1$ each time step. With continuous time, the time between two direction changes follows an exponential distribution with a mean time between two direction changes of $\mu_{\varphi^*}=\Delta t/p_{\varphi^*}$. Once a node is intended to change its direction, a new target direction φ^* chosen from eq. (4.26). The direction difference between the new target direction chosen at time t^*, $\varphi^*(t^*)$, and the old direction $\varphi(t^*)$ is $|\Delta\varphi(t^*)|=|\varphi^*(t^*)-\varphi(t^*)|$. Note that $\Delta\varphi(t^*)$ is uniformly distributed between $-\pi$ and π.

Figure 4.15 shows some examples of mobility traces of the SRMM. For more details about SRMM refer to [77].

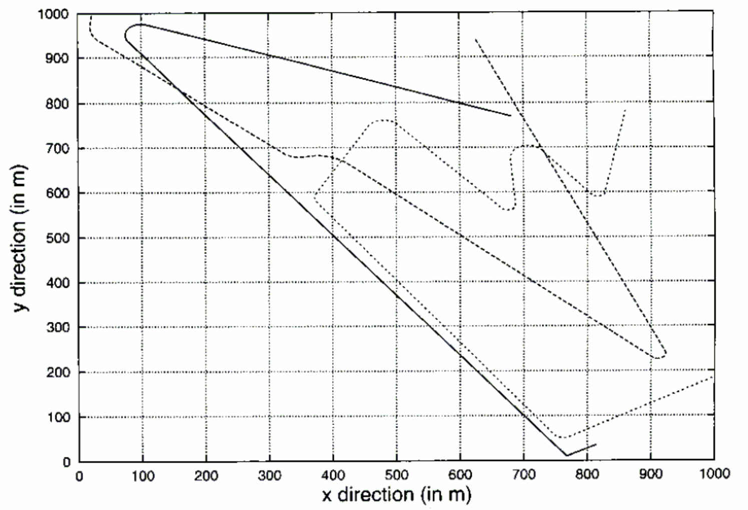
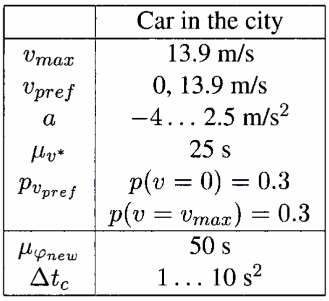


Figure 4.15. Three mobility traces under the SRMM in a 1000m x 1000m map.

The proposed model of [77] has been implemented in our simulation tool (see next chapter for details), using the concept of "stop-and-go" behaviour (with high probability for v_{pref0}=0 and v_{pref1}=v_{max}=13.9m/s=50Km/h), with the same parameters of [77] for an urban environment:



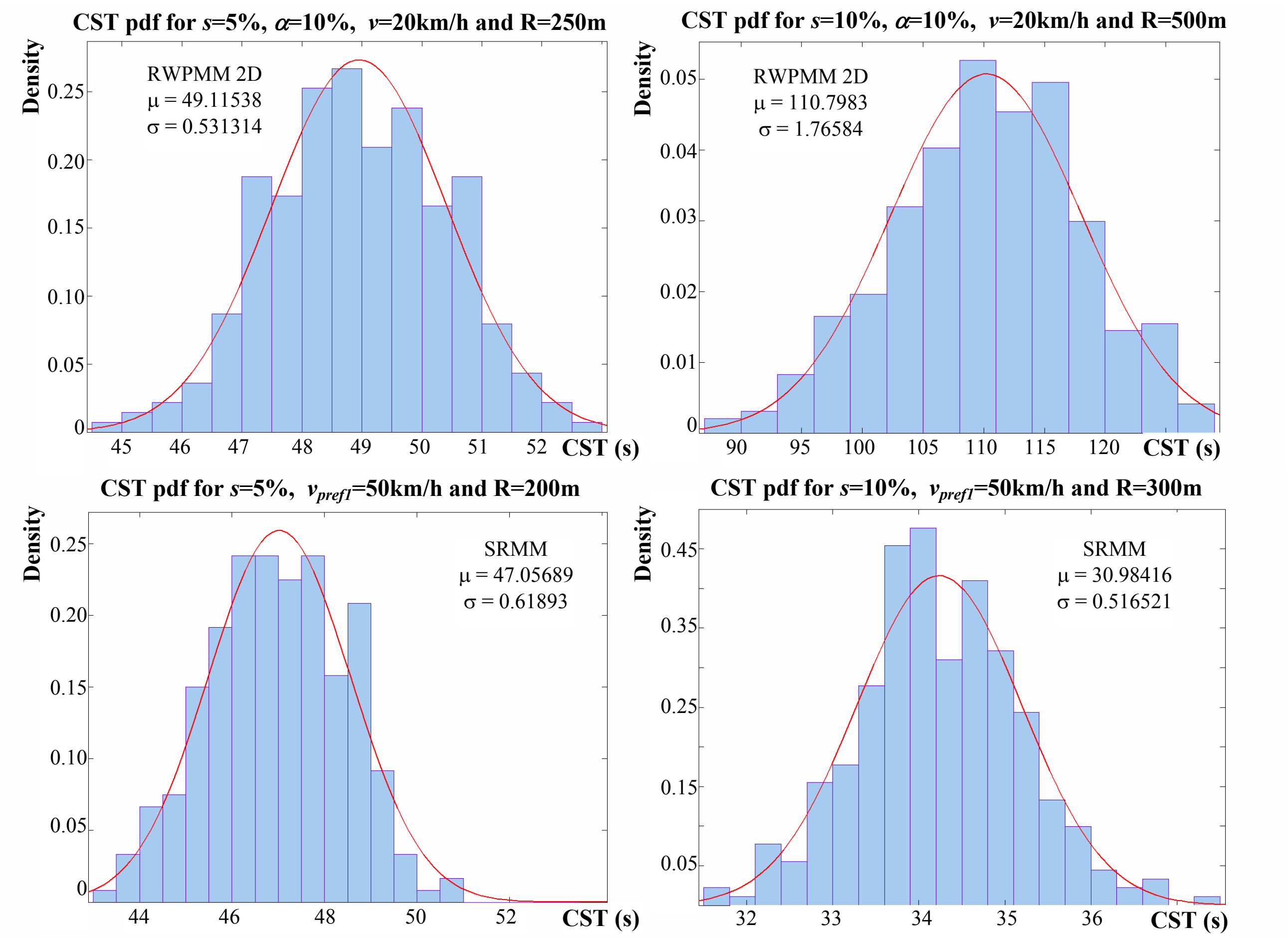
	Car in the city
v_{max}	13.9 m/s
v_{pref}	0, 13.9 m/s
a	$-4\ldots 2.5$ m/s^2
μ_{v^*}	25 s
$p_{v_{pref}}$	$p(v=0)=0.3$ $p(v=v_{max})=0.3$
$\mu_{\varphi_{new}}$	50 s
Δt_c	$1\ldots 10$ s^2

Figure 4.16. Mobility parameters for an urban environment.

In addition the concept of "wrap-around" movements has been also considered, in order to obtain a toroidal topology. In the next chapter all the details will be given when simulation input parameters will be specified.

4.6 CST analysis and prediction in a 2D scenario

The same "monitor" simulations of paragraph 4.4.1 have been carried out for the RWPMM and the SRMM, in order to have a first complete step for the 2D extension of the proposed idea and of the eq. (4.5) [102]. Some obtained results are shown in figure 4.17: it gives a description of how the CST is distributed under the RWPMM and the SRMM for different coverage radius R; in the RWPMM α represents the variation around the average speed *v*, while for the SRMM, v_{pref1} represents the second preferred speed of users, as explained in paragraphs 4.3 and 4.5 [103]. In both cases *s* represents the overlapping of adjacent cells and the red line is the obtained Gaussian approximation of the CST pdf (all the diagrams are obtained by the execution of 1000 simulation runs, as earlier discussed).


CST pdf for *s*=5%, *α*=10%, *v*=20km/h and R=250m

CST pdf for *s*=10%, *α*=10%, *v*=20km/h and R=500m

CST pdf for *s*=5%, v_{pref1}=50km/h and R=200m

CST pdf for *s*=10%, v_{pref1}=50km/h and R=300m

Figure 4.17. CST pdfs for some mobility and system parameters of RWPMM and SRMM.

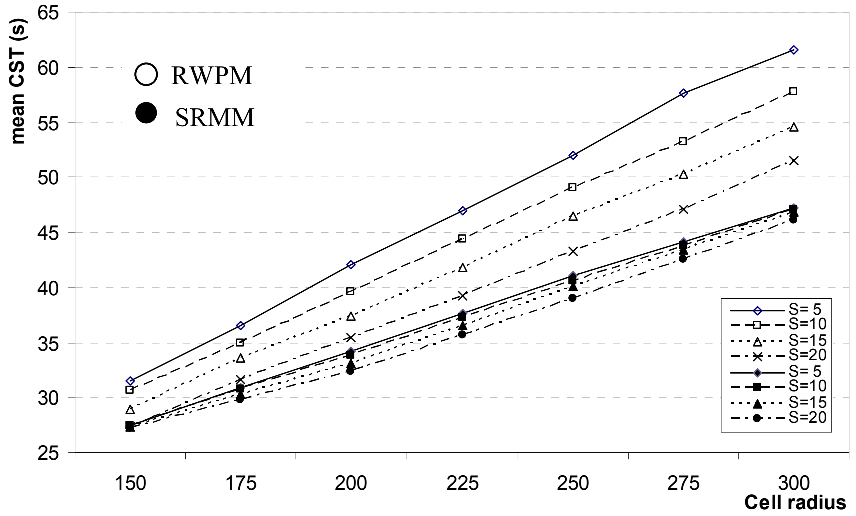


Figure 4.18. μ_{CST} vs R for RWPMM and SRMM.

Figure 4.18 shows the mean of CST distribution for different coverage area dimensions: as it can be seen, it increases for higher radius values (because of the higher space to be crossed) and decreases for higher values of *s* (a higher value of cell overlapping reduces the single cell area). No big differences in the trend are observed between the considered mobility models [104].

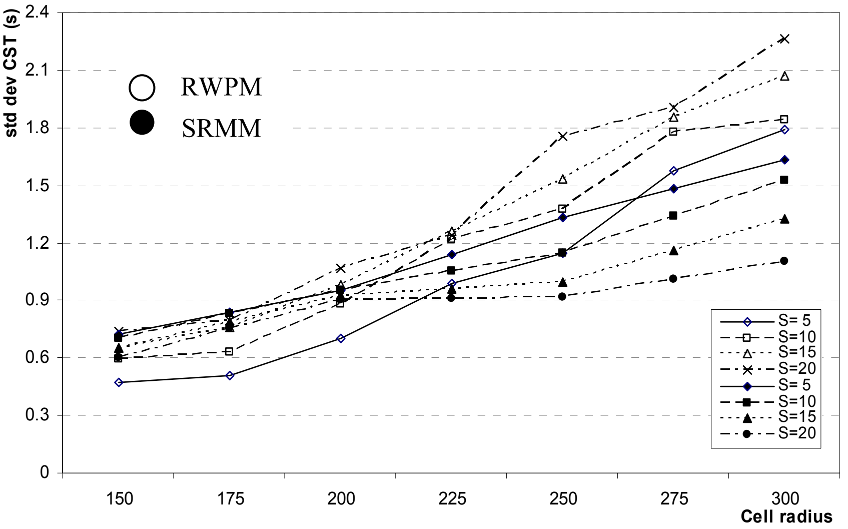


Figure 4.19. σ_{CST} vs *R* for RWPMM and SRMM.

Figure 4.19, on the other hand, depicts the trend of the standard deviation of the CST distribution: the same course of the mean is obtained, due the higher space to be crossed for higher radius (R) values; users have higher probability of changing direction and, so, a higher probability of remaining for a longer time in the same cell coverage (users speed has been fixed to 45km/h).

As example, other important results of "monitor simulations" are the statistical prediction model parameters (eq. (4.8) and eq. (4.16)), in function of the average hosts' speed. The obtained results are shown in tables 4.20 and 4.21, under the 2D RWPMM, for *R*=250m and *s*=10%.

	α=0 (constant speed)		α=5% (var. speed)	
Speed (km/h)	μ_{CST}	σ_{CST}	μ_{CST}	σ_{CST}
5	323.6451	1.046545	354.2467	8.9789
15	105.07847	0.0834	119.3434	0.1954
25	60.5112	0.03567	66.4554	0.1688
35	40.83243	0.03148	52.4518	0.0597
45	30.23874	0.02545	40.4576	0.05018
55	26.2372	0.02045	30.3212	0.04045
65	22.323445	0.01344	25.3545	0.0354
75	19.37627	0.00884	21.6966	0.0256

Table 4.20. Statistical parameters for RWPMM CCα=0 AND α=5% .

	α=15% (var. speed)		α=25% (var. speed)	
Speed (km/h)	μ	σ	μ	σ
5	366.3587	10.9894	360.5451	24.5588
15	116.22484	1.0545	116.7855	2.4333
25	72.123	0.45641	71.7243	0.7535
35	54.9474	0.23645	55.8553	0.3780
45	42.8802	0.1356	41.4881	0.2563
55	31.8897	0.1126	30.4345	0.1945
65	28.7852	0.0845	27.0556	0.1511
75	23.1533	0.05945	22.6667	0.13121

Table 4.21. Statistical parameters for α=15% AND α=25%.

From the tables above, it can be observed that the standard deviation σ_{CST} (and, so, the variance σ^2_{CST}) increases for higher values of α; this reflects bigger variations in the considered speed interval, that becomes wider, for high values of α.

4.6.1 Direction-aware prediction scheme in a 2D environment

Let us now explain the next step necessary to extend the previous treatment for a 2D simulation scenario. Figure 4.22 depicts the simulated topology (more details will be given in chapter 5); if only quantitative information (CHT and CST distributions) is available, the circular reservation is mandatory, because there is no way to dynamically choose different cells.

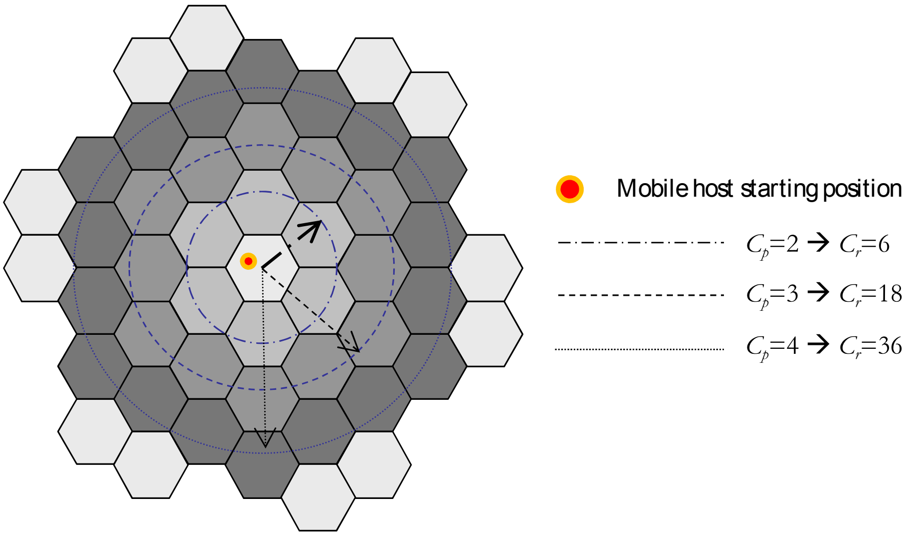


Figure 4.22. Simulated 2D network topology.

The predicted value of C_p from eq. (4.5) can be only used to make passive reservations in a circular way, around the current cell (where the call has been admitted): e.g., if C_p=2 (the user will make only one hand-over) a prediction must be made over the six adjacent cells (a circle of cells of radius C_p), because the hand-over direction is unknown; so, the number of required passive circular reservations C_r for MIP services increases with polynomial trend, such as:

$$C_r=3\cdot C_p\cdot(C_p-1), \tag{4.27}$$

where C_p is evaluated as in eq. (4.5) and $C_p\neq 0$. Obviously, the number of total reservations is $C_t=C_r+1$ (the active cell must be also considered). Clearly, the amount of bandwidth wastage is reduced if compared with a full reservation scheme over all system cells, but it is not sub-optimized. When the average host's speed is high, a higher value of C_p is obtained and the partial reservation may affect all the cells in the system (for C_p=4, C_r becomes 60!). The main goal of this paragraph is to reduce C_r, taking into account the directional behavior of mobile hosts. The C_r sub-set of predicted cells, obviously, must ensure an acceptable prediction error ([102], [103]).

So, if the system has no knowledge about the possible mobile hosts' directional movements, then there will be a lot of resources wastage, due to the enormous amount of passive pre-reserved bandwidth over C_r cells, which increases for longer calls or for higher values of the average speed, for fixed values of CHT and *v* respectively.

If additional information about the directional behavior of users is employed, above problems can be avoided and the value of C_r can be decreased, making it near or equal to C_p.

In this paragraph a novel algorithm is introduced, based on some additional information about users' directional behavior. The proposed idea is now illustrated.

A generic coverage area, generally with a circular shape, can be approximated with an *n*-edge regular polygon as depicted in figure 4.23 (*n* can be considered as an input control parameter):

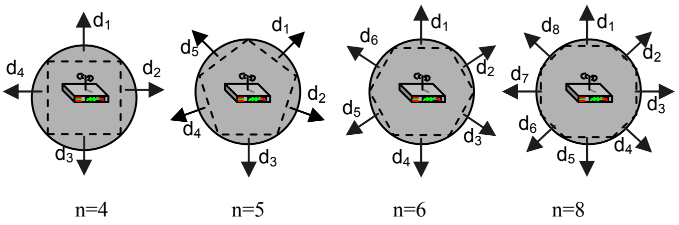


Figure 4.23. Possible (AP) coverage area approximations with regular polygons.

As it can be seen, for higher values of *n* better approximations can be reached. A set S_{ho} (hand-off directions set) of *n* possible movement directions (i.e. hand-off directions) can be then obtained: let indicate them with $d_1..d_n$, where $d_j=\underline{\theta}\cdot(2\cdot j-1)/2$ rad., $\underline{\theta}=2\pi/n$ rad. and $j=1..n$, so $S_{ho}=\{d_1, .., d_n\}$ and $|S_{ho}|=n$.

Choosing $n=6$ (hexagonal approximation), the finite set S_{ho} will be $\{d_1, d_2, d_3, d_4, d_5, d_6\}$ (in this case, each direction univocally identifies the next adjacent coverage cell). With the CST evaluation model in paragraph 4.4, the value of C_{pi} (C_p for mobile host i) can be obtained for a generic MIP call c_{MIP}, so the predicted number of hand-off events for c_{MIP} is $n_{ho\text{-}i}=C_{pi}\text{-}1$.

The conditional probability that a mobile host will be handed-out to direction $y\in S_{ho}$ after CST (a normally distributed value) amount of time, if it was handed into current wireless cell from direction $x\in S_{ho}$ can be defined as $p_{x,y}$:

$$p_{x,y}=p_{cMIP}(x,y)=p(out\ to\ y\in S_{ho}\ t=t_0+CST/\ in\ from\ x\in S_{ho}\ t=t0), \qquad (4.28)$$

where t_0 is the time instant at which the mobile host enters in the considered cell.

Once n and S_{ho} have been chosen ($n=|S_{ho}|$), a square nxn Hand-off Direction Probabilities (HDP) matrix M can be defined with the elements $M(x,y)=p_{x,y}=p_{CMIP}(x,y)$; note that CST~$N(\mu_{CST}\ ,\ \sigma^2_{CST})$. Matrix M depends only on the adopted mobility model and network cells subdivision and it is the same for all users in the system [104]. The matrix M has the hand-in directions on the rows and the hand-out ones on the columns and it can be filled out through a first addicted campaign of monitor simulations, while acquiring the CST distribution. Generally, the $M(x,y)$ elements are statistically distributed, so they have to be represented in the right way. As example, the probability to enter a cell from direction 6 and to go out to direction 3 under the SRMM is depicted in figure 4.24, for a chosen $n=6$.

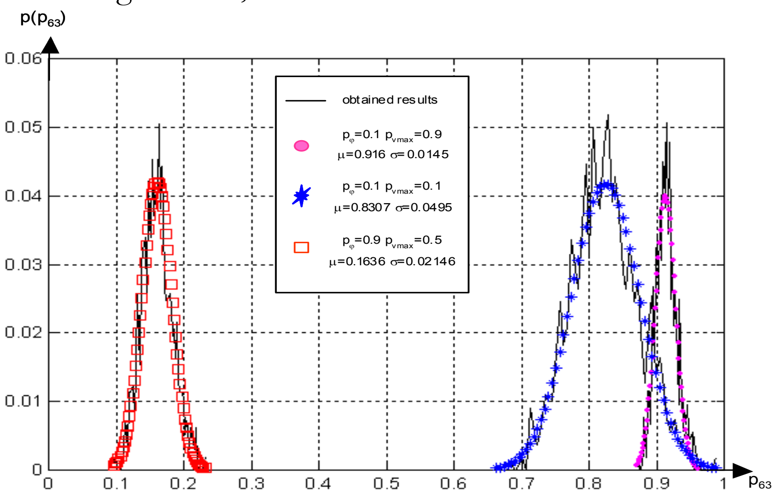


Figure 4.24. $p_{6,3}$ in SRMM model with different direction probabilities and preferential speeds.

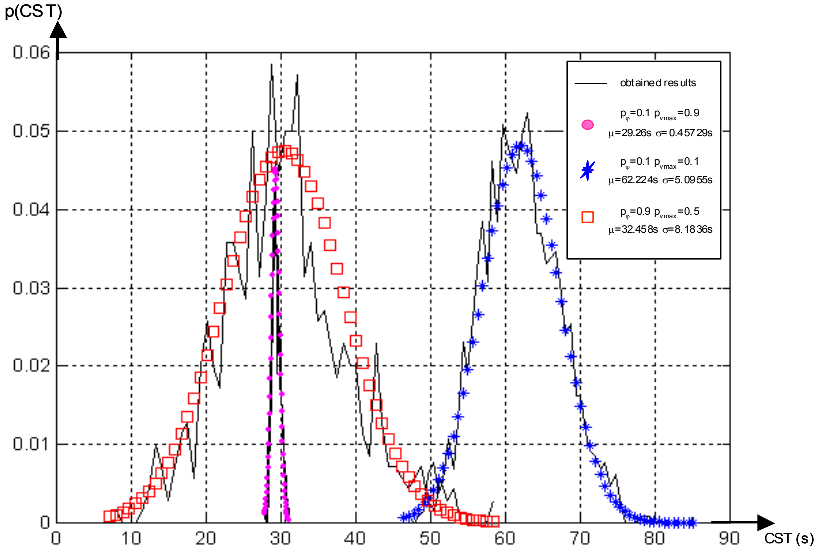


Figure 4.25. Average CST trend for different mobility parameters of the SRMM;

KS normality test [87] has been also made on the elements of *M(x,y)* for each combination of mobility parameters and the Gaussian distribution hypothesis can be made. Different values of p_φ have been considered, in order to obtain a complete set of *M* matrixes for different grades of randomness in hosts' movements. Two examples of *M(x,y)* for p_φ=0.6 and p_φ=0.1 with the mobility parameters of figure 4.16 are illustrated in figure 4.26 and 4.27. Each row contains the mean value and the standard deviation.

$$
M(x,y)=\begin{vmatrix}
0.271072 & 0.149047 & 0.155588 & 0.119348 & 0.154518 & 0.150427 \\
0.029653 & 0.02497 & 0.025759 & 0.022814 & 0.025041 & 0.024207 \\
 & & & & & \\
0.19932 & 0.237766 & 0.208252 & 0.124466 & 0.122317 & 0.10788 \\
0.024955 & 0.024024 & 0.023964 & 0.020378 & 0.019933 & 0.018556 \\
 & & & & & \\
0.124915 & 0.206507 & 0.238845 & 0.198155 & 0.109082 & 0.122496 \\
0.019344 & 0.024139 & 0.02836 & 0.02494 & 0.019902 & 0.022357 \\
 & & & & & \\
0.121575 & 0.153779 & 0.151786 & 0.26855 & 0.150536 & 0.153774 \\
0.022873 & 0.025686 & 0.023252 & 0.030941 & 0.025564 & 0.025546 \\
 & & & & & \\
0.124522 & 0.1254 & 0.108004 & 0.195957 & 0.240405 & 0.205712 \\
0.020919 & 0.02049 & 0.020234 & 0.024791 & 0.025924 & 0.025183 \\
 & & & & & \\
0.198397 & 0.108277 & 0.123875 & 0.123982 & 0.205851 & 0.239619 \\
0.024348 & 0.020101 & 0.021004 & 0.020568 & 0.025094 & 0.027464
\end{vmatrix}
$$

Figure 4.26. HDP matrix for SRMM with p_φ=0.6.

$$M(x,y)=\left|\begin{array}{cccccc}
0.0137 & 0.0244 & 0.2779 & 0.3663 & 0.3034 & 0.0256 \\
0.0061 & 0.0128 & 0.0429 & 0.0497 & 0.0476 & 0.0132 \\
 & & & & & \\
0.0325 & 0.0132 & 0.0399 & 0.3700 & 0.5056 & 0.0549 \\
0.0166 & 0.0044 & 0.0198 & 0.0525 & 0.0554 & 0.0251 \\
 & & & & & \\
0.3708 & 0.0430 & 0.0125 & 0.0316 & 0.0521 & 0.5054 \\
0.0545 & 0.0203 & 0.0030 & 0.0174 & 0.0227 & 0.0552 \\
 & & & & & \\
0.3692 & 0.2798 & 0.0249 & 0.0129 & 0.0248 & 0.2994 \\
0.0462 & 0.0464 & 0.0133 & 0.0060 & 0.0138 & 0.0461 \\
 & & & & & \\
0.3743 & 0.5094 & 0.0440 & 0.0328 & 0.0127 & 0.0437 \\
0.0554 & 0.0581 & 0.0210 & 0.0173 & 0.0031 & 0.0213 \\
 & & & & & \\
0.0318 & 0.0426 & 0.5094 & 0.3769 & 0.0427 & 0.0145 \\
0.0170 & 0.0223 & 0.0579 & 0.0556 & 0.0212 & 0.0071
\end{array}\right|$$

Figure 4.27. HDP matrix for SRMM with p_φ=0.1.

In order to obtain an analytical expression for CST mean *μCST* and standard deviation σ_{CST} in function of p_φ, a polynomial regression on the obtained discrete values has been carried out [87]; figure 4.28 shows the polynomial approximations for both curves while eq. (4.27) and eq. (4.28) are their analytical expressions (of 3rd and 1st order respectively).

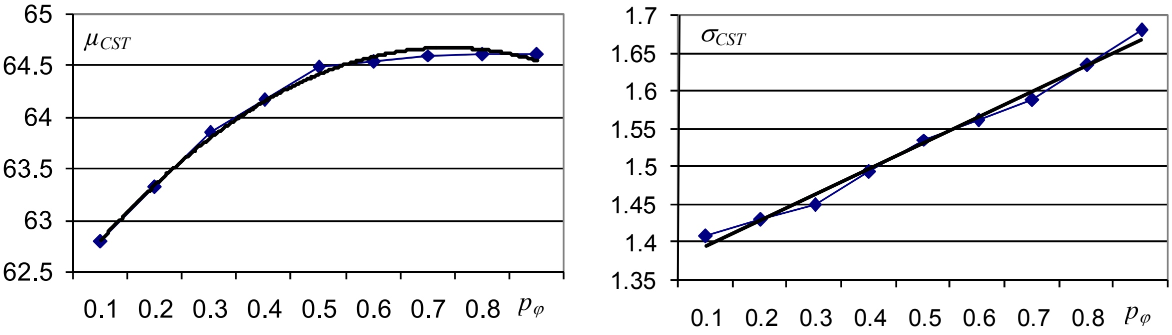


Figure 4.28. CST parameters regression in function of p_φ.

$$\mu_{CST}(p_\varphi)=2.9061\cdot p_\varphi^3-8.9476\cdot p_\varphi^2+8.5898\cdot p_\varphi+61.987. \quad (4.27)$$

$$\sigma_{CST}(p_\varphi)= 0.3406\cdot p_\varphi + 1.3595. \quad (4.28)$$

The expressions above come in handy for the introduced resource allocation schemes, because they give the values of μ_{CST} and σ_{CST} on a continuous space.

4.6.2 Static predictor for a 2D scenario

The first proposed policy consists of a static prediction scheme that selects the number of cells where to make in-advance passive reservations according to some parameters, such as maximum speed, average speed, CST and directionality. If, for a mobile host i, $C_{pi}\geq 2$ (i.e. at least one predicted hand-off event), let $j=1, ..., (C_{pi}-1)$ be the index associated to the *j-th* hand-off event, where C_{pi} is derived as in eq. (4.5); let indicate the number of desired predicted cells for *j-th* hand-off of user i with C_{ij}, where $C_{ij} \in \{1, \ldots, n\}$ $\forall j$. Three schemes are considered in the static scenario:

1) ***non-decreasing-trend reservation***: user i reserves on an increasing number of cells for increasing hand-off number ($C_{i1}\leq C_{i2}\leq\ldots\leq C_{iCpi-1}$);
2) ***non-increasing-trend reservation***: user i reserves on a decreasing number of cells for increasing hand-off number ($C_{i1}\geq C_{i2}\geq\ldots\geq C_{iCpi-1}$);
3) ***constant-trend reservation (special case of the previous ones)***: user i reserves on the same number of cells for every hand-off number ($C_{i1}=C_{i2}=\ldots=C_{iCpi-1}$).

If no preferential directions are obtained from the input mobility parameters, pre-reserving over only one future cell for every hand-off (that is C_{ij}=1 $\forall j$) may lead to an high error in predicting next visited cells, as illustrated in next chapter [106]. This problem can be solved by making next reservations not only over one cell, but pre-reserving over multiple hand-out directions, so the values of C_{ij} must be chosen in a right way. Once the C_{ij} values are chosen, the proposed scheme uses M to predict the next cell directions y for every *j-th* hand-off event of user i: if the current hand-in direction is x, then $y=index\{max[M(x)]\}$, where $M(x)$ is the *x-th* row of M and $x,y\in S_{ho}$; this is repeated C_{ij} times for every *j-th* hand-off; for every iteration, previous chosen values are not considered yet when picking up the current maximum. The following pseudo-code resumes the main steps of the algorithm; it receives n and $\mathbf{C_i}$=[C_{i1},..., C_{ho-max}] as input control parameters: since C_{pi} cannot be a-priori known because it depends on the CHT (that is assumed to be exponentially distributed) a maximum number of hand-off events C_{ho-max} for every call must be considered, under the assumption of $C_{pi}\geq 2$:

DIRECTION AWARE STATIC PREDICTOR FOR USER *i*

```
  take $n$ and $C_i$ as input parameters;
  evaluate $C_{pi}$ as in eq. (4.5);
  //make the prediction for a maximum value of $C_{ho-max}$ hand-off events
  if ($C_{pi}>C_{ho-max}$) $C_{pi}=C_{ho-max}$;
  //create the set of predicted handoff directions (the first cell is the one where the
  //call has been originated)
  Pcells={active_cell};
  //for every predicted hand-off event
  for (int j=1; j≤$C_{pi}$-1; j++) {
      //take from Pcells  the set of the predicted cells for the j-th hand-off
      current_Pcells=take_the_j-th_set_of _predicted_cells(Pcells);
      //make a prediction of $C_j$ next cells for all the candidate current predicted cell
      for (every cell in current_Pcells) {
          //if not predicting for the first hand-off event
          if (j>1) {
              current_cell=pick_the_next_cell(current_Pcells);
              //evaluate the hand-in direction for the current_cell
              current_x=take_the_current_hand-in_direction(current_cell);
              //remember the elements of M(current_x) already visited
              current_max_set=∅;
              //make the right prediction
              for (int i=0; i<$C_{ij}$; i++) {
                  //evaluate the most probable hand-out direction
                  current_direction_y= index{max[M(current_x)\current_max_set]};
                  //evaluate the next cell on current direction y adjacent to current cell
                  Pcells.append_distinct(adjacent_cell on direction current_direction_y) ;
                  //mark the current direction y as already visited
                  current_max_set.append(current_direction_y);
          }
          else {
              //evaluate the current moving direction of user i
              $d_{ik}$=evaluate_the_current_direction_of_user_i;
              //calculate the most probable $C_{i1}$ cells for first hand-off event on direction $d_k$
              first_set=determine_the_set_of_candidate_cells($C_{i1}$,$d_k$);
              Pcells.append(first_set);
          }

      }
}
return Pcells;
```

The algorithm starts the prediction from the active cell, where the call has originated; since no hand-in direction is available when a flow is admitted in a cell, matrix *M* cannot be used for predicting the next cell after the first hand-off, so the algorithm evaluates the current mobility direction $d_{ik} \in S_{ho}$ of user *i* in the active cell with one of the approaches of [93], [94] and, hypothesizing that user *i* will probably follow direction d_k until the first hand-out event, the identifier of the most C_{i1} probable cells that user *i* will visit following direction d_{ik} from the current position can be discovered and inserted into the prediction set *Pcells*. From j=2 to $j=C_{pi}$-1, the algorithm creates a temporary set called *current_Pcells* with the predicted cells belonging to *j-th* hand-off event; for each of them, it determines the hand-in direction called *current_x*, then it evaluates the maximum value in the vector *M(current_x)\current_max_set*; the last discovered maximum value is appended into the "*current_max_set*" vector, that is substracted form *M(current_x)* in the next iteration (in this way the new maximum value is always calculated, without considering the previous ones). The algorithm has a time complexity of $O((C_{pi}-1)\cdot n^2)$.

Table 4.29 summarizes the obtained *p*-values for different values of p_φ, with p_{v0}=0.4 and p_{vmax}=0.5; remembering that a *p-value* for a statistical test is a measure of how much evidence there is against the null hypothesis, different *p-values* have been obtained, showing the goodness of the Gaussian distribution hypothesis.

Table 4.30 shows the obtained values of CST parameters and the predicted number of cells on which MIP users make passive requests with a CHT exponentially distributed (λ_{CHT}=180 seconds) for different values of p_φ and p_{vmax}: C_r values are obtained through eq. (4.27) and they belong to the circular reservation policy, while $C_r(P)$ values are obtained following the approach previously proposed, for different reservation policies *P* (non-decreasing, non-increasing or constant); however, it can be seen that there is a resource gain if a directional treatment is introduced. For instance and without loss of generality, $C_{ho\text{-}max}$ has been fixed to 3 (under the assumption that the generic call *i* is long enough in order to suffer at least 3 hand-over events) and the notation $P(C_1\text{-}C_2\text{-}C_3)$ indicates that the reservation policy *P* makes passive reservations on C_1, C_2 and C_3 cells for *1-st*, *2-nd* and *3-rd* hand-off respectively, that is to say the input vector $\boldsymbol{C}$ is $[C_1,C_2,C_3]$, as previously illustrated (the subscript *i* is not used for sake of simplicity).

p_φ	μ_{CST}	σ_{CST}	*KS p-value*
0.1	62.7861	1.4075	0.5305
0.3	63.8491	1.4487	0.5087
0.5	64.4721	1.5342	0.6712
0.7	64.5944	1.586	0.6147
0.9	64.6055	1.6792	0.4994

Table 4.29. Values of μ and σ of CST distributions and KS *p-values* for different mobility

	P(1-1-1)		*P(1-2-3)*		*P(3-3-3)*	
	$C_r(P)$	C_r	$C_r(P)$	C_r	$C_r(P)$	C_r
$p_{vmax} = 0.9, p_{\varphi} = 0.1$ μ=29.26s, σ=0.45729s	5	60	22	60	52	60
$p_{vmax} = 0.1, p_{\varphi} = 0.1$ μ=62.22s σ=5.09547s	2	6	3	6	4	6
$p_{vmax} = 0.9, p_{\varphi} = 0.5$ μ=32.46s σ=8.18356s	4	36	16	36	21	36

Table 4.30. Number of cells involved in the bandwidth reservation phase, for different policies *P*.

4.6.3 Dynamic predictor for a 2D scenario

Differently from the previous idea, now the number of predicted cells for the *j-th* hand-off event are chosen dynamically, through an input control threshold δ. Also in this case C_{pi} is evaluated with the approach of eq. 4.5. Let us indicate the number of hand-off events of user *i* with h_i=C_{pi}-1; let vh_i be an information-support array (like the one illustrated in figure 4.31), where $vh_i[k]$ with k=*1..*h_i indicates the information (cells identifiers, hand-in directions, etc.) about the *k-th* future hand-off of user *i*; i.e. each entry in the array vh_i, $vh_i[k]$, can be a pointer to a list of tuples {*cell_id, from, to,* p_{cell_id}} for the *k-th* hand-off event, where:

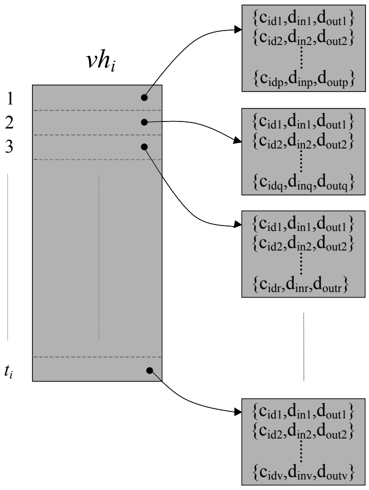

vh_i
1
2
3
t_i
$\{c_{id1},d_{in1},d_{out1}\}$
$\{c_{id2},d_{in2},d_{out2}\}$
$\{c_{idp},d_{inp},d_{outp}\}$
$\{c_{id1},d_{in1},d_{out1}\}$
$\{c_{id2},d_{in2},d_{out2}\}$
$\{c_{idq},d_{inq},d_{outq}\}$
$\{c_{id1},d_{in1},d_{out1}\}$
$\{c_{id2},d_{in2},d_{out2}\}$
$\{c_{idr},d_{inr},d_{outr}\}$
$\{c_{id1},d_{in1},d_{out1}\}$
$\{c_{id2},d_{in2},d_{out2}\}$
$\{c_{idv},d_{inv},d_{outv}\}$

Figure 4.31. An example of support structure for predicted cells storing.

- *cell_id* is a cell identifier;
- *from* $\in S_{ho}$, *to* $\in S_{ho}$ are respectively the hand-in and hand-out directions for the *cell_id* cell;
- p_{cell_id} is the probability that user will be under the coverage of the *cell_id* cell after the *k-th* hand-off.

The algorithm predicts *to* directions for each tuple, starting from *cell_ids*, *from* directions and p_{cell_id} values. Let δ be an input threshold for the cell estimation phase (as for *n*, δ is an input control parameter that affects system performances, as will be illustrated in section IV).

If the knowledge of the first hand-off cell is approached, for example, with one of the policies proposed in [93], [94], then the following threshold-based predictor algorithm can be performed in order to obtain the complete set of predicted cells that

MIP user *i* will probably visit, starting from the *2-nd* hand-off event and going on until the h_i*-th* one ($h_i \geq 2$):

DIRECTION AWARE THRESHOLD-BASED PREDICTOR FOR USER *i*

```
//for every predicted hand-off event of user i
for (int k=2; k ≤ $h_i$; k++) {
  //index on the cells of the k-th hand-off event
  int l=1;
  //for each cell of the current k-th hand-off event
  while (l ≤ $vh_i[k]$.size()) {
    //let us analyze the current l-th tuple in the k-th element of $vh_i$
    current_tuple= $vh_i[k]$.elementAt[l];
    //the hand-in direction is known
    curr_hand_in_dir= From(current_tuple.to);
    //probability of user i of being in current cell after the
    //(k-1)-th hand-off
    $p_{curr}$=current_tuple.$p_{cell\_id}$;
    //find the "more suitable" hand-out candidate cells over
    //the n possible hand-out directions
    for (int p=1; p ≤ n; p++) {
     //the probability of hand-out on direction p after having
     //handed in on direction curr_hand_in_dir is evaluated
     curr_prob=M(curr_hand_in_dir, p)*$p_{curr}$;
     //threshold based comparison
     if  (curr_prob ≥ $\delta^{f(k)}$){
       //the current cell can be considered a valid candidate
       id=Cell_id( current_tuple.cell_id,  p);
       //the $v_{hi}$ vector must be updated
       create_a_tuple{id, curr_hand_in_dir, p, curr_prob};
       append the tuple in $vh_i[k+1]$;
     }
    }//for p
    l++;
  }//while l
   clean $vh_i[k+1]$ from duplicates;
 }//for k
create an empty cell identifiers list p_cells;
//extract cell ids from tuples and append them to p_cells
for (int k=1; k<=$t_i$; k++) {
  for (int l=0; l<$vh_i[k]$.size(); l++)  {
    current_tuple=$vh_i[k]$.elementAt(l);
    append current_tuple.cell_id to p_cells;
  }
}
return p_cells.
```

As earlier discussed, the candidate cell for the first hand-off must be discovered, because no hand-in direction is available when a flow is admitted in a cell. With one of the approaches of [93], [94] the current mobility direction $d_j \in S_{ho}$ of user *i* is obtained and the term *first_id=first_Cell_id(current_id,* d_j*)* can be evaluated, by an appropriate function *first_Cell_id* that determines the identifier of the cell that user *i* will visit (*current_id* is the identifier of the current cell). As it will be shown in next chapter, this approach leads to a negligible amount of error for the first prediction (around 3%-4%). At this point, a tuple {*first_id,_ , d_j, 1*} can be created and appended in *$vh_i[1]$*; the *from* direction cannot be discovered because user *i* has started its flow in the current *first_id* cell, without handin-in it from any direction while p_{first_id}=1 because the probability of hand-out from *first_id* cell during the first hand-off is 1. Let us hypothesize, only for now, that the elements of *M* are constant values, so the main aim is now the prediction making for all the cells contained in the list of *$vh_i[k]$*, with *k=2..h_i*. Each tuple in *$vh_i[k]$* contains the hand-in direction, the cell identifier and the probability of user ***i*** of being in the cell after the *(k-1)-th* hand-off; through a threshold-based comparison the algorithm must decide what are the cells that user *i* will visit with higher probability when handing-out the cell of the *l-th* tuple of *vh_i [k]*, *l=1…vh_i[k].size()* with a well known hand-in direction; the hand-in direction *curr_hand_in_dir* belongs to S_{ho} and it specifies a unique row of *M*; the algorithm calculates the probability of hand-out from the current cell on direction *p* after having handed-in from direction *curr_hand_in_dir* when the probability of being in the previous cell before the current hand-off is p_{curr}: if the obtained value is higher than $\delta^{f(k)}$, then the cell that is adjacent to the current one on direction *p* must be considered as a possible future cell and a tuple {*adjacent_p_cell, from*, *p*, *curr_prob*} is appended in *$vh_i[k+1]$*. The exponent *f(k)* is a function of *k* and its expression is discussed in a little while. The power operation is necessary in order to take into account the increase of prediction error for higher values of *k*. The function "***cell_id Cell_id(cell_id current_id, direction to)***" returns the identifier of the cell adjacent to ***current_id*** cell on ***to*** direction; the function "***direction From(direction to)***" translates the hand-out direction ***to*** of the previous cell in the hand-in direction of the next cell. When repeating all the steps *h_i-1* times, a cleaning routine must be executed after finishing appending elements in *$vh_i[k]$* position, because of possible duplications of cell identifiers; the same results can be obtained if the "append" function avoids duplicates. Also in this case, the algorithm has a time complexity of $O((C_{pi}-1)\cdot n^2)$.

The prediction result is the set of cell identifiers of the tuples for each *vh_i* list. The hypothesis of *M* composed by constant values is not suitable: *M(x,y)* consists of a couple of values, the mean and the standard deviation of the obtained distribution, as depicted in figure 4.26 and 4.27. So in the proposed pseudo-code $M(x,y)=N(\mu_{x,y},\sigma_{x,y})$.

In our simulations we considered 4 different expressions for the exponent function $f(k)$: a) $f(k)=1$; b) $f(k)=\alpha k$; c) $f(k)=\alpha/k$ and d) $f(k)=(\alpha k)^{-1}$, with $\alpha>0$, in order to appreciate the different behaviours of the algorithm by varying $\Delta=\delta^{f(k)}$ structure and how δ is weighted for consecutive values of k (i.e. consecutive hand-offs).

A different approach has been followed in the previous static scheme, because a static reservation policy has been adopted and the HDP matrix has been applied through the selection of a prefixed number of columns without considering the gap in direction probabilities. The static scheme does not account for M structure and a prediction sequence C_i must be specified as an input parameter ([107], [108]).

4.6.4 Statistical Bandwidth Multiplexing (SBM)

Another important issue of the proposed work is the statistical passive bandwidth multiplexing: when a MIP user pre-reserves a certain amount of passive bandwidth in the remote Access Points (APs - statistically identified by one of the proposed algorithms in the previous paragraphs) it may be considered as available resource when other incoming calls make requests to enter into the system. So, the CAC module of the whole Access Points constellation has to implement a time-based bandwidth multiplexing. Let us suppose that MIP user j (MIP_j) has been admitted into the net, after the making of its active reservation and its passive reservations and another MIP user k (MIP_k) is making a new service request ($j\neq k$). Let us suppose that MIP_j has also made a passive reservation in the remote AP h (AP_h); predicting that $MIPj$ will probably reach AP_h at the enter time instant $t\text{-}in_{jh}$, then its passive bandwidth PBW_{jh} (Passive BandWidth of MIP_j reserved in APh) can be re-used by MIP_k until the time $t\text{-}in_{jh}$; so if MIP_k will probably leave AP_h at the exit time $t\text{-}out_{kh}$, then it can re-use PBW_{jh} for the period $T_{kh}=[t\text{-}in_{kh} \div t\text{-}out_{kh}]$ only if $t\text{-}out_{kh} < t\text{-}in_{jh}$. That is to say MIP_k will leave AP_h before MIP_j enters it with the request of the availability of PBW_{jh}, that will be switched into active bandwidth. In this way, the wastage of PBW_{jh} is avoided at least for the T_{kh} duration.

The overall AP bandwidth capacity C, without multiplexing, can be so considered as composed by three main contributions:

a) the bandwidth used by MIP active reservations (A_{MIP});

b) the free bandwidth, available for new active-passive requests (S);

c) the passive and unused bandwidth (P_{MIP}) that can be multiplexed.

It is always verified that $C\text{-}S=A_{MIP}+P_{MIP}$. The AP utilization U is defined as $U=A_{MIP}/C$. When the time-based multiplexing is utilized, the P_{MIP} contribution is further subdivided in: $ActiveP_{MIP}$, the active MIP bandwidth multiplexed on P_{MIP}, $PassiveP_{MIP}$, the passive MIP bandwidth multiplexed on P_{MIP} and $FreeP_{MIP}$, the amount of P_{MIP} not yet multiplexed. In this case, the AP utilization increases to $\mathrm{U}=(A_{MIP}+ActiveP_{MIP})/C$. The CAC module must ensure that $A_{MIP}+ActiveP_{MIP}+FreeP_{MIP}+S \leq C$ is always verified.

The CAC algorithm can be resumed as the following pseudo-code when a new active or passive service request arrives to the AP*h*; the request is accepted if a true value is returned:

CAC ALGORITHM WITH STATISTICAL MULTIPLEXING

```
//let BWreq be the bandwidth level requested by MIPk
//the request can be accepted with no multiplexing
if (BWreq ≤ S) {
        //the available free bandwidth must be decreased
        S-=BWreq;
        //APh is an active Access Point
        if (MIPk request == active) AMIP+=BWreq;
        //APh is a remote Access Point
        else PMIP+=BWreq;
        return true;
//the CAC tries to accept the request on the passive bandwidth
else {
        //determine the amount of bandwidth that can be multiplexed
        availableBW=determinePMUX(MIPk);
        if (BWreq ≤ availableBW) {
                if (MIPk request == active) ActivePMIP+=BWreq;
                //APh is a remote Access Point
                else PassivePMIP+=BWreq;
                return true;
        }
}
//the request must be refused
return false;
```

The key function of the multiplexing CAC in the AP_h is “BWlevel determinePMUX(BW request)”; when it receives the argument MIP_k, it evaluates the time period T_{kh} for the bandwidth request *k*; then it compares the obtained time period with all the occupancy periods of the *n* non-multiplexed accepted passive reservations MIP_j as previously exposed, so $j=1..n$. The returned value BWlevel takes into account all the *j* non-multiplexed reservations for which $t\text{-}out_{kh} < t\text{-}in_{jh}$ is verified.

4.7 Conclusions on chapter 4

In this chapter, a deep overview on the mobility analysis for wireless environments has been given; the importance of mobility prediction for wireless systems (such as WLANs) has been outlined, introducing some schemes for passive reservations enhancements. Simulation results of the proposed approaches will be given in next chapter, where simulation results are well collected and explained in a right way. Two mobility models have been considered, RWPMM and SRMM, but the proposed schemes are completely uncorrelated with the employed mobility model; only the matrix *M* resumes the behavior of mobile hosts and, if the model is changed, only a single round of "monitor simulations" is needed to correctly fill-out the *M(i,j)* elements. The introduction of the proposed schemes has started with the analysis of a simplified 1D scenario, in order to give some knowledge about the CST evaluation and analysis, in terms of statistical distribution. After the quantitative analysis of CST distribution, an extension has been introduced, by considering the complete 2D space and the directional behavior of mobile hosts. So after a circular reservation, a directional one has been proposed. The obtained CST distributions have been shown, in terms of mean and standard deviation parameters. In addition, a polynomial regression has been shown to be a good way to obtain a CST evaluation by simply introducing system and mobility parameters. The static scheme and the dynamic one have been proposed and formally described by pseudo-code. Obviously, the performance of the proposed schemes must be evaluated; this will be made in the next chapter.

Chapter 5 – Simulation tool and performance evaluation

5.1 Introduction

In previous chapters a deep and complete description of the proposed idea has been given, in order to well understand the problems that have been faced during the PhD research period. In this chapter, the implemented simulation tool, based on the *actors programming paradigm*, will be briefly described (the attention is not focused on it) and simulation results will be shown, in order to appreciate the effectiveness of the proposed bandwidth management scheme and prediction algorithm. A first simulation campaigns regard the 1D environment with the integrated channel model based on the Markov chain; then the 2D extension is introduced, through the simulation of a complete wireless network.

5.2 The simulation tool: a brief overview

A discrete-events [95] *C++* network simulator based on the *actors programming paradigm* [96] has been implemented, taking into account the concepts that have been exposed from chapter 1 to chapter 4. In particular, four main actors have been implemented:

- **FiniteStateMarkovChain** (FSMC): it is associated to each mobile host and it evolves according to the equations of Chapter 1;
- **MobileHostMRSVP**: it models the behaviour of a single user, in terms of reservation protocol (MRSVP) and mobility model (RWPMM or SRMM);
- **AccessPointMRSVP**: it models the behaviour of a single AP, in terms of CAC scheme, bandwidth management and wireless link monitoring;
- **SenderMRSVP**: it models the sender node; it sends data packets to each mobile node, respecting the MRSVP constraints.

Figure 5.1 illustrates, in a simplified manner, the architecture of the implemented simulation tool: each wireless link is univocally associated to a single mobile host; it assumes a particular degradation level, accounting for mobile host mobility profile. A generic mobile host can exchange messages with its current AP, while passive reservations are admitted only with the message exchange between AP; the sender can communicate with a mobile host by the intermediation of an AP.

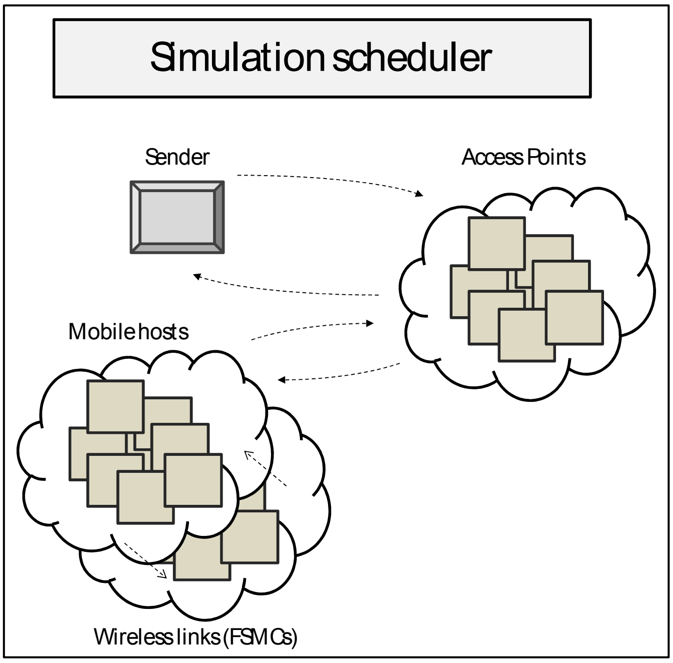


Figure 5.1. Actor based simulation tool.

As in every discrete-events simulator, the scheduler enqueues the messages of each actor, then it dispatches them at the right time. Now the simulation campaigns are described, following the same steps of the previous chapters.

Concluding this brief overview, it can be said that the main features of the implemented tool are: complete implementation of the MRSVP protocol, implementation of the described mobility models (RWPMM, SRMM) and introduction of the Finite State Markov Chain for the channel modelling. Obviously, also the CAC schemes and the static-dynamic algorithms of chapter 4 have been introduced in the AP handler. Simulation results are now shown.

5.3 Simulation results for a 1D environment

In a first campaign of simulations the 1D scenario of figure 5.2 has been considered; As illustrated, there are 5 wireless cells, each one covered by an AP; a MRSVP sender is connected to the APs through an "infinite-bandwidth" wired switching-subnet (the wired bandwidth availability is not comparable with the wireless one). The total bandwidth of each AP is 5.5Mbps. Each mobile host starts its flow (after the CAC) in a certain current cell (e.g. mobile host 1, MH1, in cell C1), then it moves straight in a circular way following the 1D RWPMM (e.g. if it starts in the cell C4, it will visit C4, C5 then C1, C2, etc.), until it has visited all the cells or the connection has finished. The proposed Call Admission Control and Bandwidth Reallocation schemes of chapter 4 are implemented in each AP. Some important simulation parameters are:

- mean of requests arrival rate (Poisson process) λ_a: 3 flows/s;
- exponentially distributed call duration with mean μ = 180s;
- admissible bandwidth levels for each flow (Kbps): 512, 640, 768, 896;
- token bucket size (bit): 896000;
- token bucket rate (bit): 512000;
- token bucket peak-rate (bit): 896000;
- packet size (bit): 512;
- percentage of variations around the average speed α=10%
- cell radius *R*=250m; cell overlapping percentage *s*=10%.

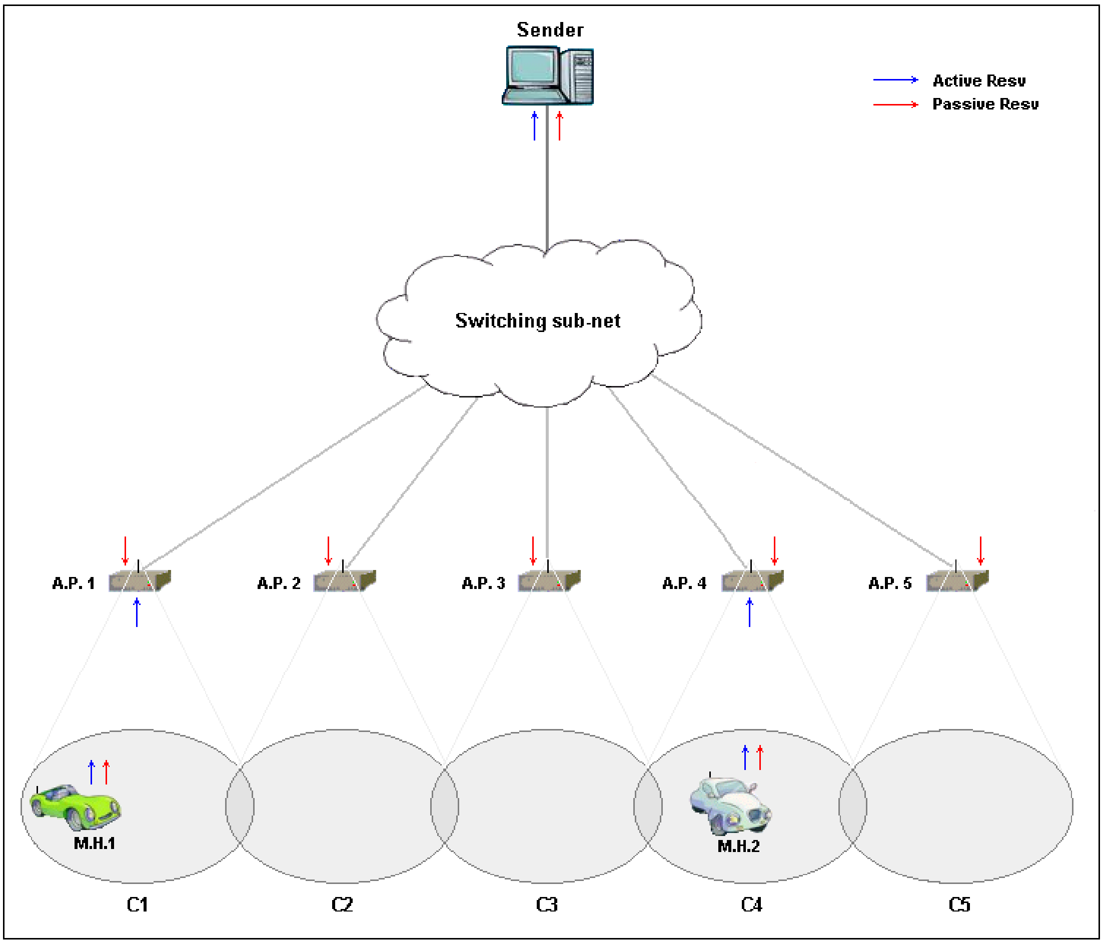


Figure 5.2. 1D simulated network.

The FSMC has been tuned for the CCK modulation with an average SNR ρ=4dB (4-state FSMC). The obtained parameters are (see tables 1.12 and 1.13):

- Number of states K=4, N=K-1=3;
- Steady state probabilities (p_i): 0.09253, 0.33614, 0.40704, 0.1643;
- SNR ranges (dB): [0, 0.47502), [0.47502, 2.37764), [2.37764, 6.46132), [6.46132, ∞);
- State degradations (%): 41.83, 24.587, 9.513, 0.

The performance of the system are investigated in terms of received bandwidth, user satisfaction level, average system utilization and average number of admitted and dropped flows. These are the typical parameters that are usually observed in order to evaluate the correctness of the proposed policies. Final statistics are obtained following the "independent replies with termination" simulation method: each run simulates 800 seconds of discrete-events time and the number of runs for each point has been fixed to 20; in this way the confidence interval of 95% has been respected for every considered parameter.

The utility function that has been introduced for MIP and MDP traffic is illustrated in figure 5.3:

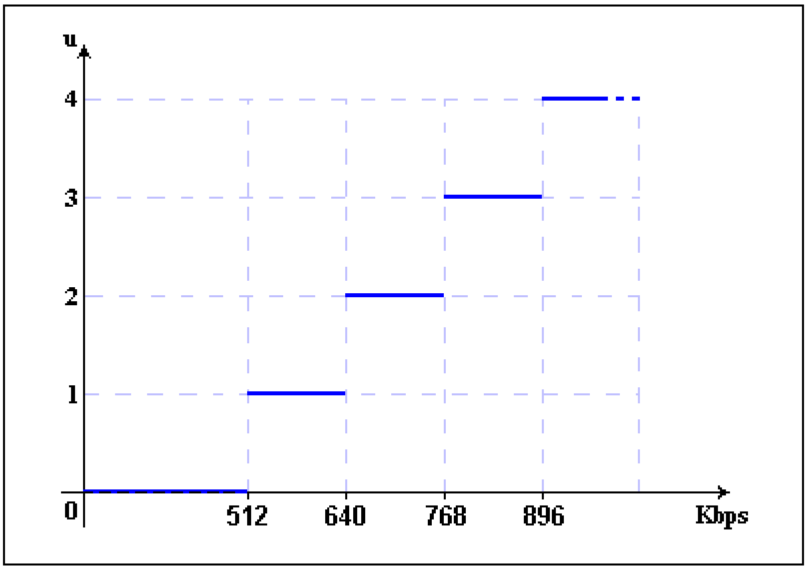


Figure 5.3. 1D simulated network.

The considered utility function is a piece-wise non-decreasing function, since there is only a finite set of bandwidth levels. So, the allowed utility values are: {*1, 2, 3, 4*}.

As example of the complexity reduction of the Call Admission Control proposed in chapter 3, the CAC matrix that contains the elements of eq. (3.17) is shown in figure 5.13bis for every couple of (available bandwidth, admitted users).

Av. Bandwidth				NUMBER OF ADMITTED USERS								
(kbps)	0	1	2	3	4	5	6	7	8	9	10	11
5632	0	0	0	0	0	0	0	0.01623	0.31279	0.83346	1	1
5504	0	0	0	0	0	0	0	0.03456	0.41601	1	1	1
5376	0	0	0	0	0	0	0	0.06521	0.52638	1	1	1
5248	0	0	0	0	0	0	0.00076	0.11156	0.63655	1	1	1
5120	0	0	0	0	0	0	0.0022	0.17962	0.73675	1	1	1
4992	0	0	0	0	0	0	0.00716	0.26652	0.82233	1	1	1
4864	0	0	0	0	0	0	0.01896	0.37031	0.88963	1	1	1
4736	0	0	0	0	0	0	0.03919	0.48754	1	1	1	1
4608	0	0	0	0	0	0	0.07785	0.60478	1	1	1	1
4480	0	0	0	0	0	0	0.13654	0.71489	1	1	1	1
4352	0	0	0	0	0	0.00253	0.21569	0.811	1	1	1	1
4224	0	0	0	0	0	0.00649	0.3221	0.88448	1	1	1	1
4096	0	0	0	0	0	0.01954	0.44202	1	1	1	1	1
3968	0	0	0	0	0	0.04821	0.56707	1	1	1	1	1
3840	0	0	0	0	0	0.09061	0.69211	1	1	1	1	1
3712	0	0	0	0	0	0.16609	0.79728	1	1	1	1	1
3584	0	0	0	0	0	0.26791	0.87916	1	1	1	1	1
3456	0	0	0	0	0.00836	0.38552	0.93867	1	1	1	1	1
3328	0	0	0	0	0.01886	0.52807	1	1	1	1	1	1
3200	0	0	0	0	0.05173	0.66377	1	1	1	1	1	1
3072	0	0	0	0	0.11779	0.77947	1	1	1	1	1	1
2944	0	0	0	0	0.19762	0.87826	1	1	1	1	1	1
2816	0	0	0	0	0.32895	0.94129	1	1	1	1	1	1
2688	0	0	0	0	0.48126	1	1	1	1	1	1	1
2560	0	0	0	0.02763	0.62025	1	1	1	1	1	1	1
2432	0	0	0	0.05368	0.76873	1	1	1	1	1	1	1
2304	0	0	0	0.13115	0.87844	1	1	1	1	1	1	1
2176	0	0	0	0.27153	0.94166	1	1	1	1	1	1	1
2048	0	0	0	0.39659	0.98499	1	1	1	1	1	1	1
1920	0	0	0	0.58468	1	1	1	1	1	1	1	1
1792	0	0	0	0.76232	1	1	1	1	1	1	1	1
1664	0	0	0.0914	0.86427	1	1	1	1	1	1	1	1
1536	0	0	0.14885	0.95713	1	1	1	1	1	1	1	1
1408	0	0	0.31065	1	1	1	1	1	1	1	1	1
1280	0	0	0.57029	1	1	1	1	1	1	1	1	1
1152	0	0	0.70063	1	1	1	1	1	1	1	1	1
1024	0	0	0.8775	1	1	1	1	1	1	1	1	1
896	0	0	1	1	1	1	1	1	1	1	1	1
768	0	0.30233	1	1	1	1	1	1	1	1	1	1
640	0	0.39733	1	1	1	1	1	1	1	1	1	1
512	0	0.65	1	1	1	1	1	1	1	1	1	1
384	0	1	1	1	1	1	1	1	1	1	1	1
256	0	1	1	1	1	1	1	1	1	1	1	1
128	0	1	1	1	1	1	1	1	1	1	1	1

Figure 5.3bis. The CAC matrix for the chosen simulation parameters.

5.3.1 Conjunctive and Disjunctive non-predictive 1D bandwidth management

Let us now analyze the obtained results for those simulations where MIP flows reserve passive bandwidth over all the AP and varying the outage threshold, as well as traffic percentages (20%MIP-80%MDP, 40%MIP-60%MDP, etc.). When a disjunctive policy in used, in each figure there are shown the values of $p_{outageMIP}$ and $p_{outageMDP}$ (in this order), while for the conjunctive management only one value of p_{outage} is shown. For the disjunctive management the MIP threshold has been fixed to $4\cdot10^{-3}$ (a very low value, to prevent MIP outage events), while for MDP it varies from $4\cdot10^{-3}$ and $4\cdot10^{-1}$. In this way MIP users might not be influenced by thresholds variations, while the MDP statistics might vary. For the conjunctive management the threshold varies from $4\cdot10^{-3}$ and $4\cdot10^{-1}$. Now some results for non-predictive management without MDP bandwidth reuse are shown. Simulation results for the 1D case are given only for sake of completeness; the main attention will be focused on the proposed predictive algorithms in a 2D environment.

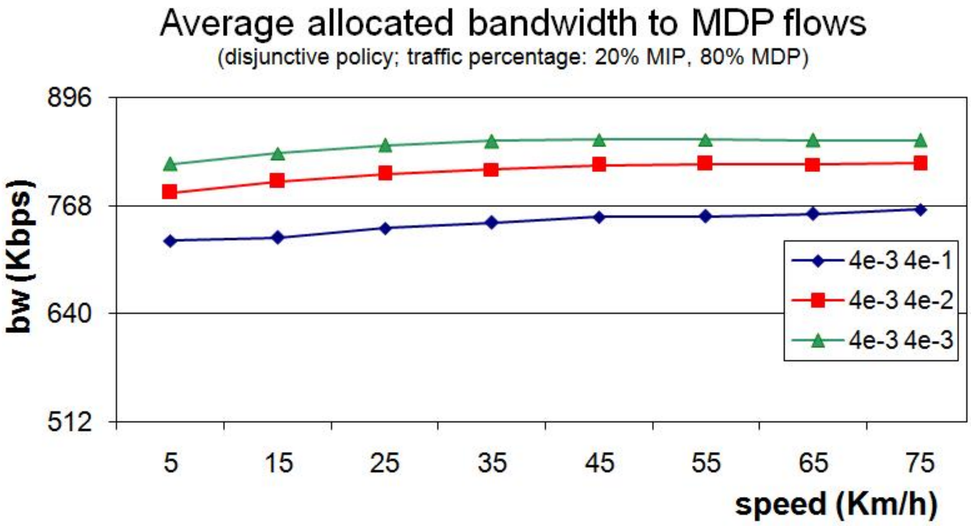


Figure 5.4. Average allocated bandwidth for MDP flows.

The bandwidth level associated to MDP flows is affected by the threshold variations: when diminishing the threshold, the CAC module limits the number of admitted MDP, so there is more bandwidth availability. Also the MDP received utility is increased when the threshold is lowered: it can be concluded that maintaining the outage threshold to low levels, a lower number of flows will be admitted, receiving a better service, in terms of bandwidth and service satisfaction (some curves are omitted only for space limitations).

The following figures 5.5 and 5.6 show how the system is occupied from different class users: the low utilization of MIP flows is justified by and their low traffic percentage (20%) or the low chosen threshold; in addition the CAC must be made over all the system cells, while the MDP flows must complete their CAC only on the current cell.

Each AP is heavy utilized by MDP flows and, often, MIP requests cannot find available passive resources. In figure 5.7, an optimal system utilization is shown.

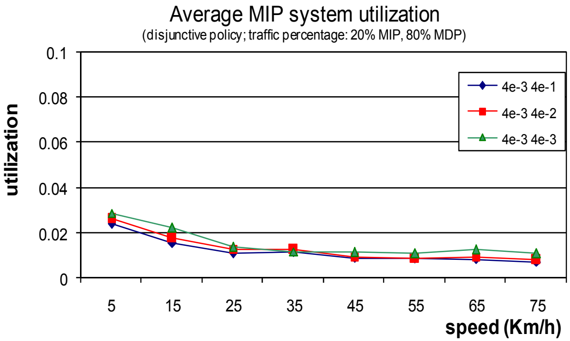


Figure 5.5. Average MIP utilization.

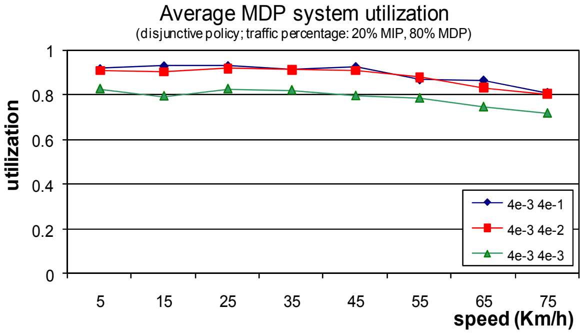


Figure 5.6. Average MDP utilization.

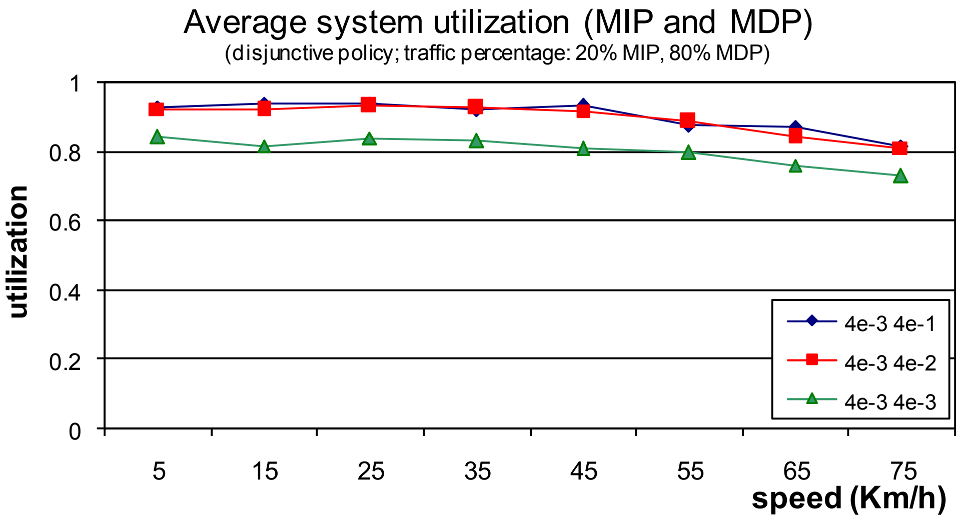


Figure 5.7. Total average system utilization.

The decreasing trend of all the utilization curves is due to the faster evolution of wireless link for higher speeds (the probability of remaining in the same state becomes low), thus the number of bandwidth reallocations becomes higher. In addition, if the outage threshold is decreased, then the utilization decreases too, because of the lower number of amditted flows.

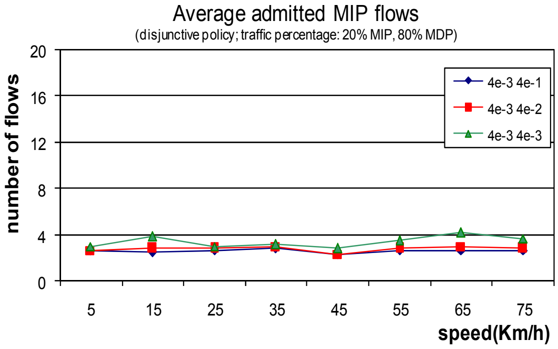


Figure 5.8. Average admitted MIP.

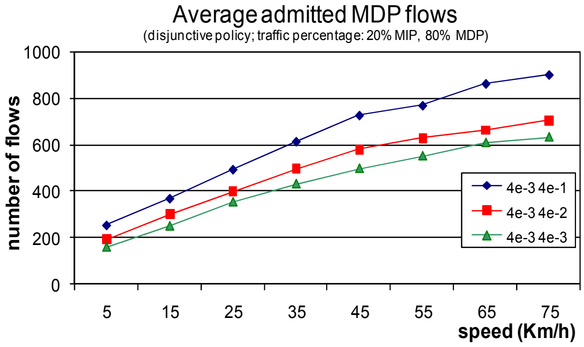


Figure 5.9. Average admitted MDP.

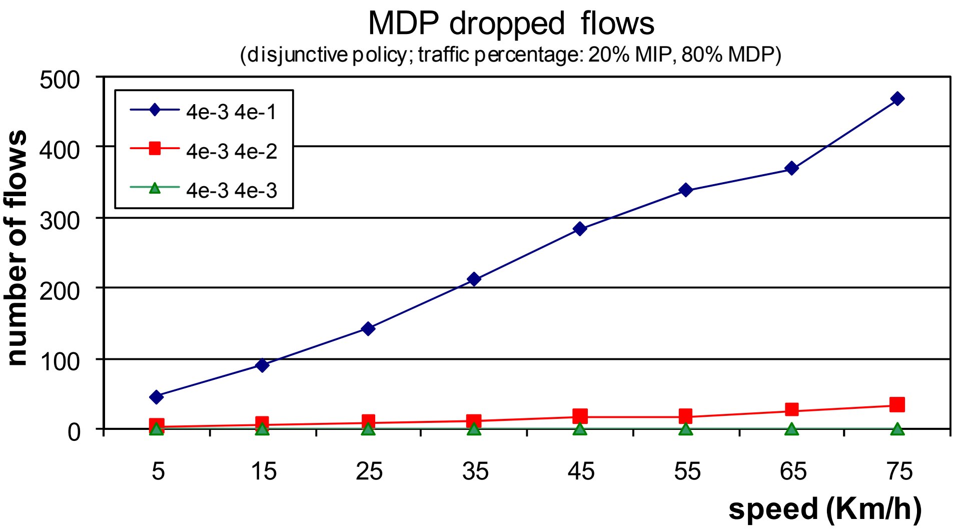


Figure 5.10. Average admitted MDP.

In order to confirm the low MIP system utilization, figure 5.8 shows that the number of admitted MIP flows is low if compared with the one of the MDP case (figure 5.9). But, on the contrary, the few admitted MIP flows (3 or 4) does not have any dropping; it is not true for the MDP. So MDP flows do not have any influence on the MIP ones (disjunctive management): the trend is not affected by the variation of the input MDP threshold. In addition, figure 5.10, illustrates how the chosen threshold 0.4 results to be too much high, because the system cannot manage all the outage events, so many MDP flows must be dropped.

When the outage threshold is decreased, the admitted MDP decrease too, verifying the policy used by the admission control (the frequency of outage events decreases too). When the average speed is increased, there is an higher number of hand-off events, because the average CST decreases; in this way, more MDP users can enter into the system. From other obtained simulation results it can be seen that, when the traffic percentages vary (to 80% MIP and 20% MDP), there is an enhancement in the allocated MIP bandwidth, but the system utilization decreases drastically: the amount of passive bandwidth increases and the MDP users, in this scenario, cannot reuse it, so it is wasted.

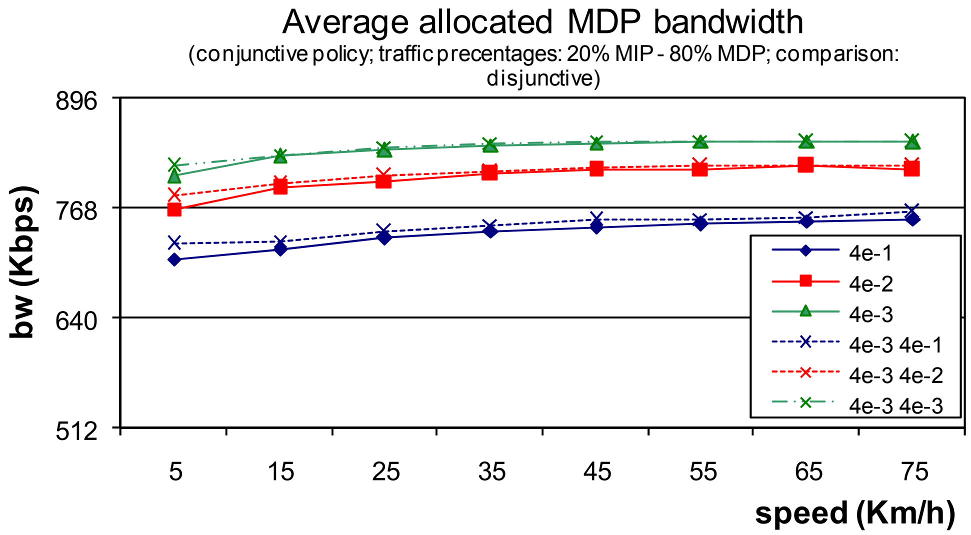


Figure 5.11. Average allocated MDP bandwidth: comparison between disjunctive and conjunctive management.

In figure 5.11 a comparison in terms of allocated MDP bandwidth for disjunctive and conjunctive management is made: continuous curves represent the obtained results for disjunctive management and it can be seen that there are no visible differences with the disjunctive case. Although in the conjunctive management the threshold varies also for MIP users, there are too few admitted MIP flows so their impact on the allocated bandwidth is not visible. Since MDP users cannot multiplex the passive bandwidth in this simulation scenario, the only priority of MIP users in bandwidth reallocations is the pre-emption of MDP users when some flows must be dropped.

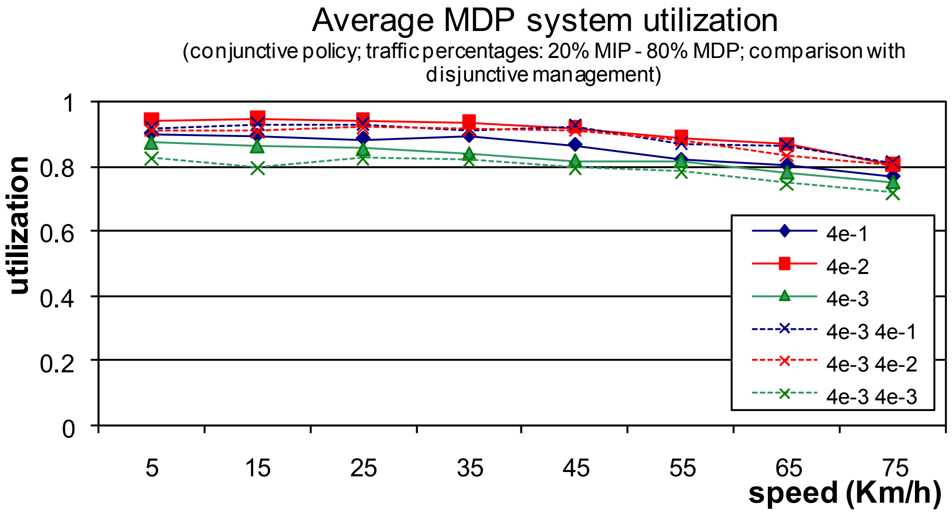


Figure 5.12. Average MDP system utilization: comparison between disjunctive and conjunctive management.

Figure 5.12 shows that the system, as in the previous case, is utilized only by MDP flows (MIP system utilization does not exceed the value of 3%) and the lower value obtained for higher threshold is due to the slightly higher presence of admitted MIP into the network.

When traffic percentages are changed to 80% MIP and 20% MDP, system utilization goes down, because of the higher presence of passive reservations.

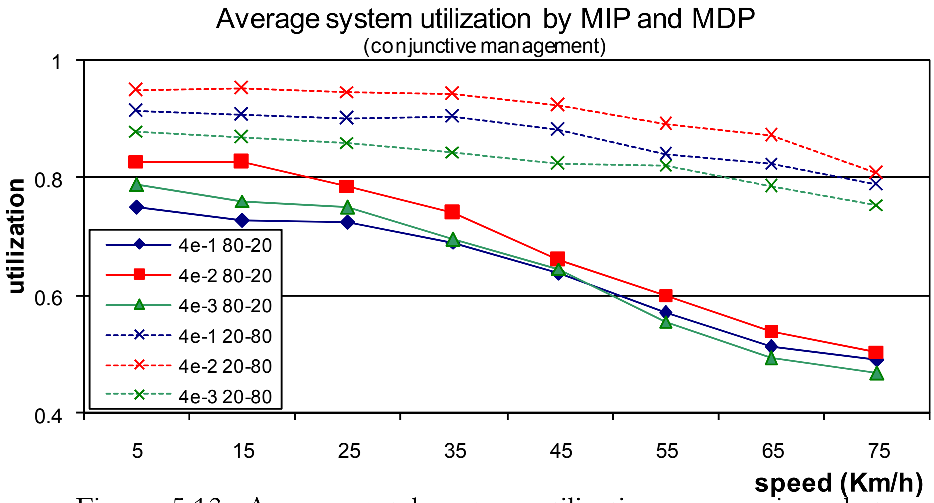


Figure 5.13. Average total system utilization: comparison between different traffic percentages.

Figure 5.13 illustrates the decreasing of system utilization when MIP traffic percentage is increased. The decreasing trend for higher speeds is due to the increased overhead for passive reservations (lower CST → higher number of MIP admitted flows).

5.3.2 Conjunctive and Disjunctive non-predictive 1D bandwidth management with passive bandwidth MDP multiplexing

The main aim now is the 1D performance evaluation with the introduction of MDP passive bandwidth reuse as described in chapter 3. For space limitations problems, only some figures will be shown. The attention will be mainly focused on the 2D predictive management in next paragraphs.

When the "MDP multiplexing" is introduced, there are no differences with the previous case if the MIP percentage is low (from 0 to 25%), because of the absence of enough passive bandwidth to multiplex. When enough passive bandwidth is present into the system, introducing the "MDP multiplexing" leads to higher system utilization in both conjunctive and disjunctive cases, as illustrated in figures 5.14 and 5.15. It must be outlined that the introduced multiplexing scheme can be applied only to MDP users, because no QoS guarantees are given. The passive multiplexing for MIP users introduced in the previous chapter takes into account the predicted hand-off time instants, in order to avoid MIP degradations or droppings.

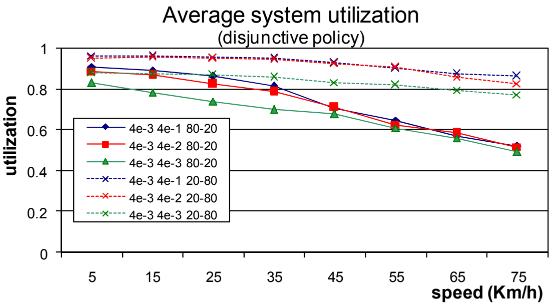


Figure 5.14. Average system utilization with MDP multiplexing (disjunctive management).

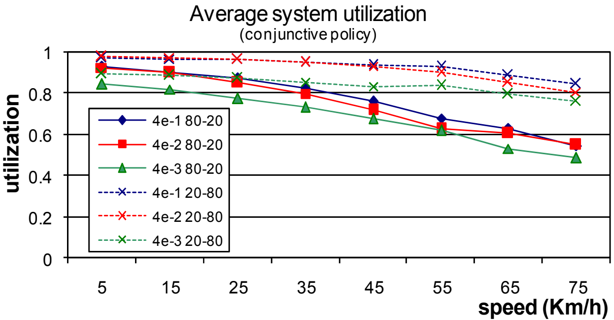


Figure 5.15. Average system utilization with MDP multiplexing (conjunctive management).

5.3.3 Predictive passive bandwidth management

As explained in previous chapters, now the data collection of the "monitor simulations" is of primary importance because it must be applied in the prediction algorithm based on eq. (4.5). The same simulation experiments of previous paragraphs have been led out and a comparison will be now shown in order to appreciate the benefits of the proposed 1D prediction scheme; as earlier discussed, now the passive reservations are made by sending SPEC messages only to the AP that are considered in the possible visited cells set. So, not all the APs of the system will be affected by passive reservations.

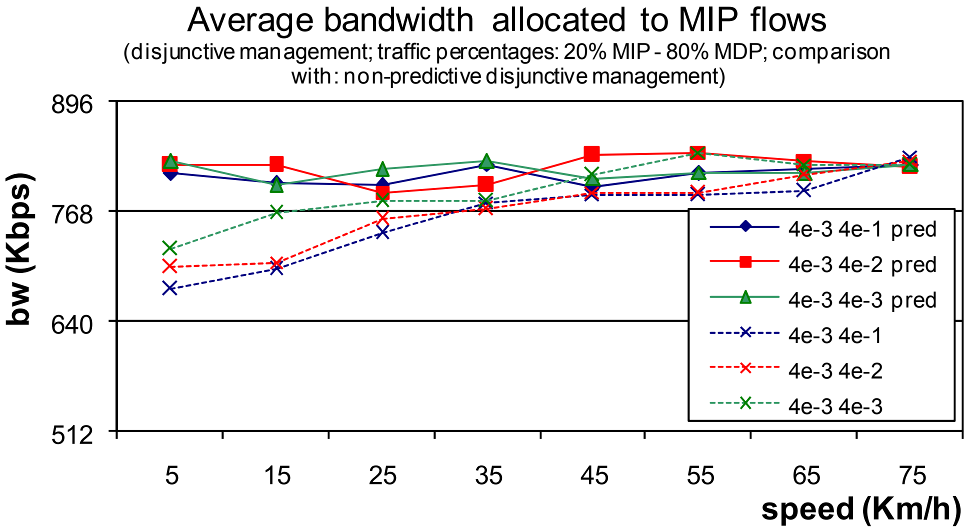


Figure 5.16. Average allocated bandwidth to MIP flows.

In figure 5.16 it is shown that there is a slight enhancement in the allocated MIP bandwidth: reducing passive reservations increases free system bandwidth, which can be better allocated to MIP flows, especially for lower speeds, when the system stationarity is more evident, because of the lower amount of hand-over events.

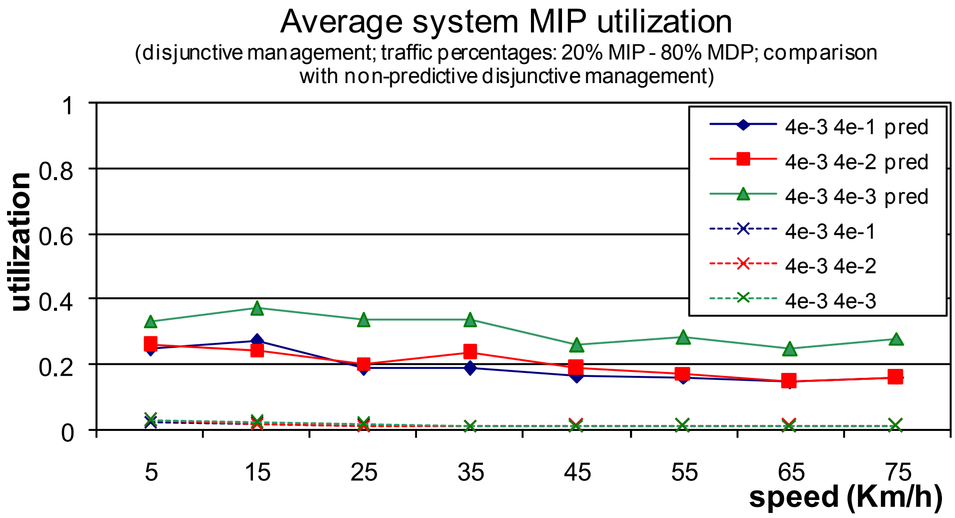


Figure 5.17. Average system MIP utilization.

Better enhancements can be observed when considering MIP system utilization: from very low values (around 3%), higher utilization can be reached (between 16% and 37%), as depicted in figure 5.17, with an average gain of 20%-25%.

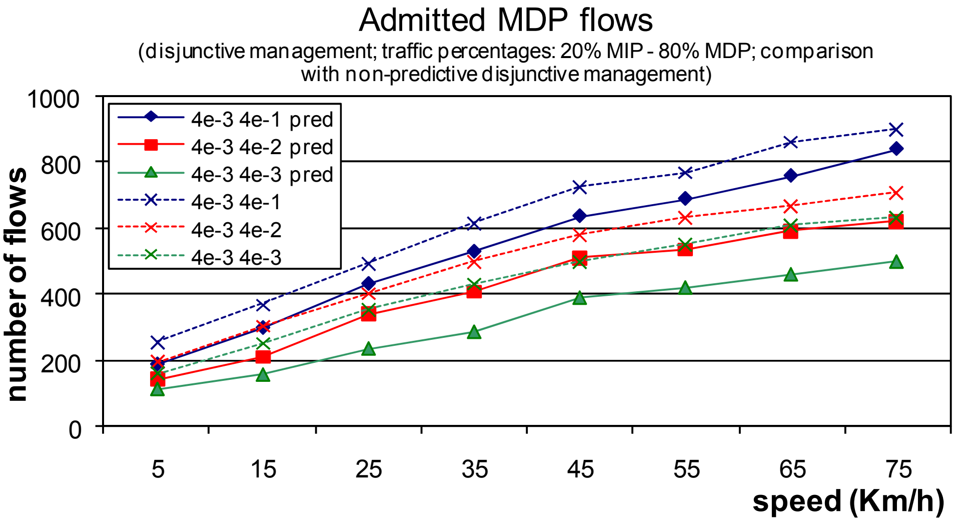


Figure 5.18. Average admitted MDP flows.

If a lower number of passive reservations is obtained, more MIP users can enter into the system, so the amount of free MDP available bandwidth becomes smaller, then a lower number of MDP will be admitted. A higher number of MIP connections is allowed into the system because the term *C* of eq. (3.18) is reduced.

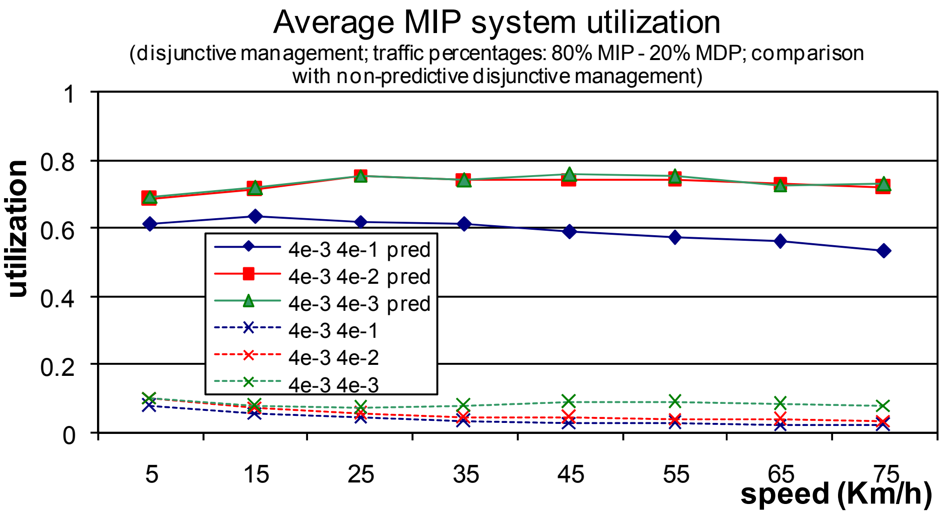


Figure 5.19. Average MIP system utilization.

Figure 5.19 illustrates the enhancements in system utilization when MIP percentage is high: the number of gained passive reservations is considerable and MIP system utilization goes from 10% to 75%; MDP utilization decreases from 75% to 30%. As we expected, the introduction of a passive predictive policy increases system performances, because many unused reservations are avoided.

When the predictive policy is not used, MIP flows are penalized because of the high number of CAC requests which must be made in order to enter into the system.

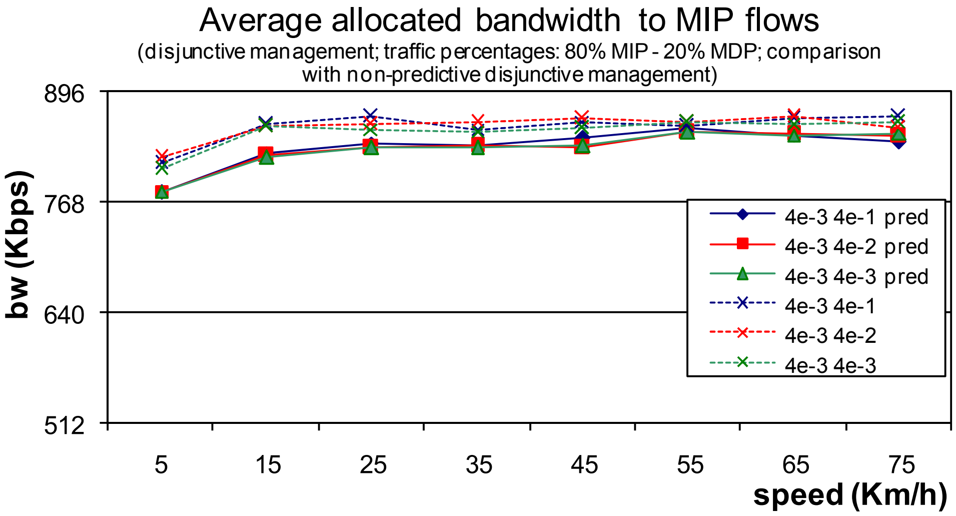


Figure 5.20. Average allocated bandwidth to MIP flows.

Figure 5.20 shows that, although there is a slightly lower amount of allocated bandwidth, it can be well considered acceptable if considering the enormous gain in system utilization; obviously the allocated bandwidth is decreased because of the higher presence of MIP admitted flows.

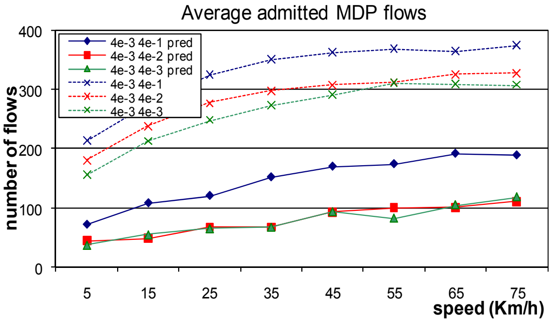


Figure 5.21. Average admitted MDP flows (disjunctive management).

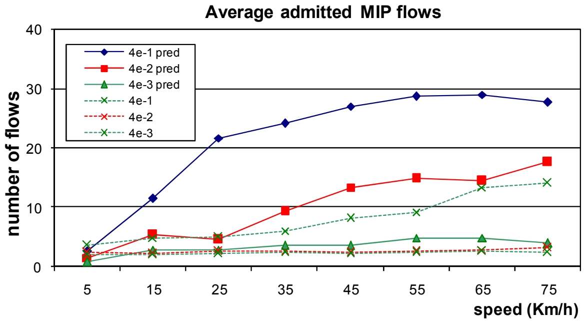


Figure 5.22. Average admitted MIP flows (conjunctive management).

Figure 5.21 and 5.22 show the effects of CST prediction on the average number of admitted flows for 80% MIP and 20% MDP traffic percentages: it is shown that for the disjunctive case there is a lower number of admitted MDP flows, because there is a higher value of admitted MIP users and the available resources are more scarce (from a maximum value near to 400 of the non-predictive case to a maximum value near to 200 of the predictive case). For the conjunctive case the trend of MIP flows is shown: when the predictive CST scheme is introduced, the number of admitted flows is doubled (from a maximum near to 15 to a maximum near to 30).

5.3.4 Conclusions on the 1D CST-prediction scheme

Independently from the adopted management policy (conjunctive or disjunctive), from the adopted traffic percentage and from the availability of MDP passive bandwidth multiplexing, it must be outlined that the introduction of the predictive 1D CST-based passive reservations leads to some system enhancements, as we expected:
with the knowledge of the CST distribution it is possible to apply eq. (4.5) in order to determine, when the mobile host makes its MIP service request, the number of future cells C_P that the host will probably visit during its active session. As illustrated into the figures, MIP flows can properly participate to system utilization (without prediction, their utilization maintained below 10%) with a better bandwidth management in terms of allocated resources for each flow, that becomes comparable or higher than the MDP allocated bandwidth; so, a certain amount of unused passive reservations are avoided.

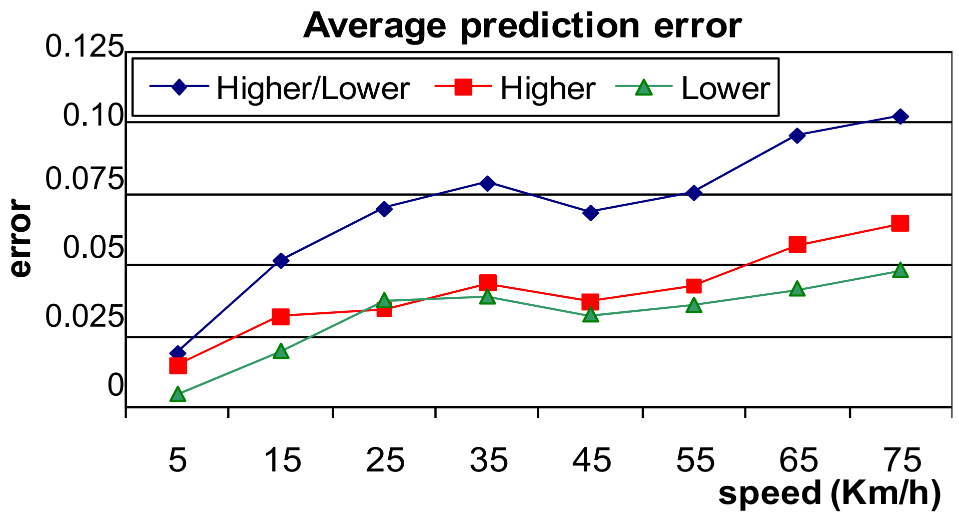


Figure 5.22bis. Average prediction error.

Figure 5.22bis illustrates the average prediction error of the 1D proposed reservation scheme for 100% MIP traffic (there are no differences if the conjunctive/disjunctive management is considered): the highest curve indicates the percentage of MIP flows that visit a number of cells that is higher or lower than C_p, while the other ones represents the percentage of flows that visit a higher (lower) number of cells if compared with C_p. It is evident that the error increases for higher average speed because the granularity of the prediction makes bigger, then higher accuracy is needed. Nevertheless, for the considered speed range, the error is often below than 10% and it is an acceptable result.

Now the attention will be focused on the 2D predictions management.

5.4 Simulation results of the 2D circular reservation policy

Before introducing the performance evaluations of the final algorithms proposed in chapter 4, the circular reservation of paragraph 4.6.1 and eq. (4.27) are considered. In order to evaluate the goodness estimation of the CST and the performance of rate adaptation scheme, call admission control for MIP and MDP and the conformance to QoS parameters (outage probability, minimum received utility and a high system utilization), different simulations have been led out also in a 2D scenario, under the RWPMM. This time, the simulated net consists of 49 wireless cells, each one covered by an access point (for sake of simplicity, figure 5.23 illustrates only one of the seven simulated 2D clusters). Other simulation parameters are the same of the previous paragraph.

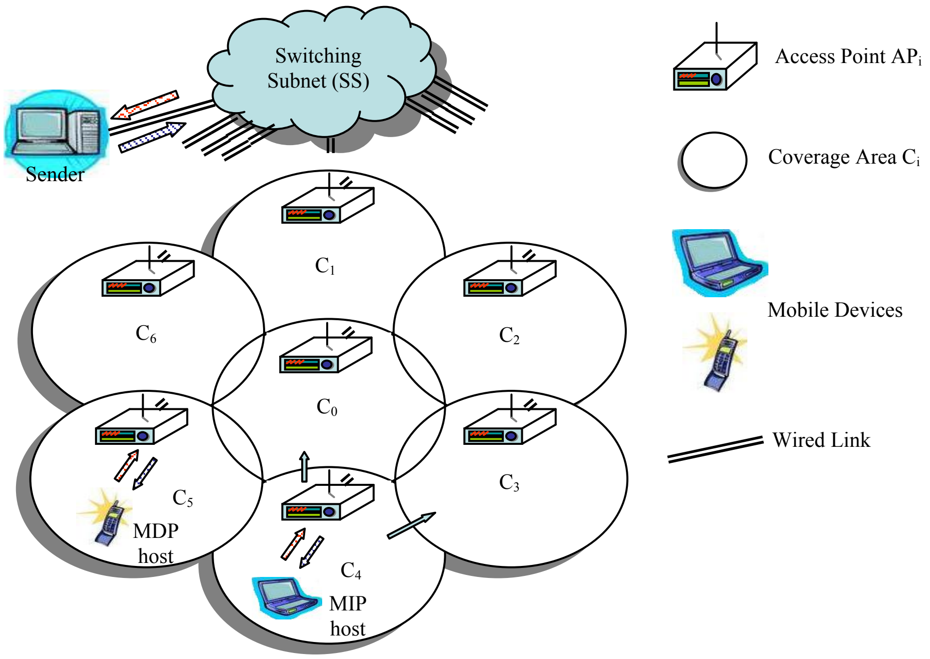


Figure 5.23. A cluster of the 2D simulated wireless network.

The following curves illustrate the performances of the utility-oriented algorithm for different values of outage threshold and mobile host speed, in absence of any predictive policy.

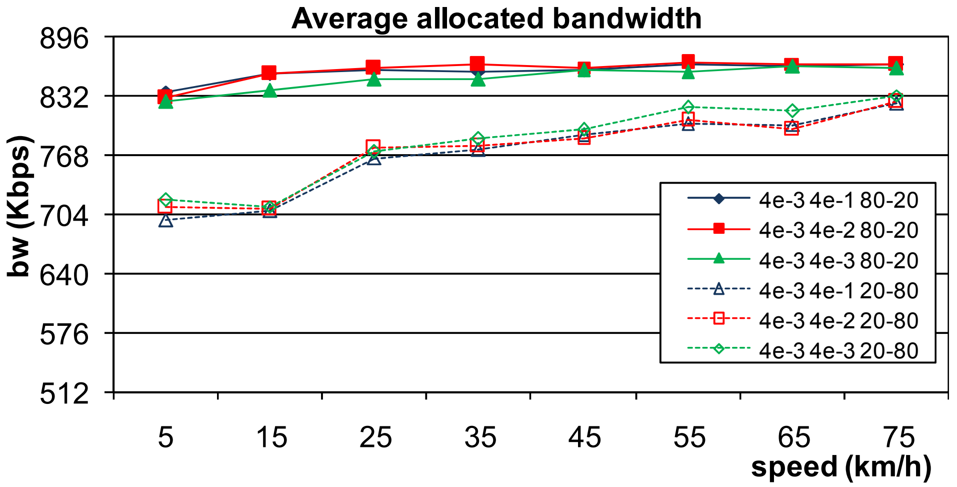


Figure 5.24. Average allocated bandwidth for MIP users.

In figure 5.24 it can be observed the improvement in resource allocation for MIP users, by increasing their traffic percentages from 20% to 80%; in this figure and in the following ones the first two columns indicate the fixed outage probability for MIP and MDP traffic, while the last column indicates MIP%-MDP% traffic percentages. For higher MIP traffic, more users, belonging to this class, can enter the system, pre-empting MDP flows and degrading MDP reservations more frequently; in addition, for high MIP traffic percentages, increasing outage threshold, a decreasing in allocated bandwidth can be observed: in this situation, there are more MIP users sharing the same cells capacities, so they must perceive a lower amount of bandwidth. In addition, when the MIP traffic percentage is low (around 20%) the effects of mobility of MDP users reflect on the MIP bandwidth, so it increases from 700kbps to 830kbps if the average speed is increased.

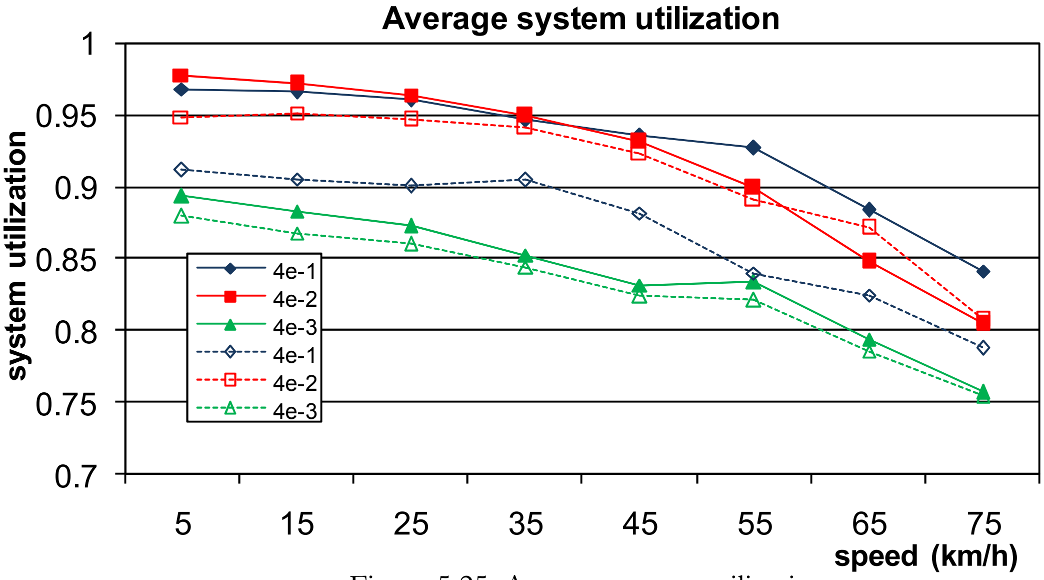


Figure 5.25. Average system utilization.

Figure 5.25 shows the average system utilization when a MIP-MDP traffic percentage of 20%-80% is set: the system is lower utilized by increasing hosts' speed. For high speed values there are more link variations and, consequently, system must handle a larger number of bandwidth reallocations; this causes a utilization wastage, which can reach a magnitude of 15%-20%. Varying outage threshold, there are different observed values of resource utilization, for the same reason early discussed, that is to say the admission control is less selective for higher threshold values, so more users can enter the network and a higher utilization can be reached. The bandwidth wastage is not evident because of the presence of MDP flows in the system; when the traffic is only composed by MIP flows, the system is under-utilized, because of the unused passive pre-reserved bandwidth. The difference between continuous curves and dashed ones is the chance for MDP to reutilize (continuous) or not (dashed) the passive bandwidth; MIP flows, after a hand-off, can pre-empt MDP users obviously. It is evident that there is a gain in the system with the multiplexing of passive bandwidth, especially for high values of p_{outage}: higher threshold values allow higher MDP in the system, which can reuse the available passive bandwidth.

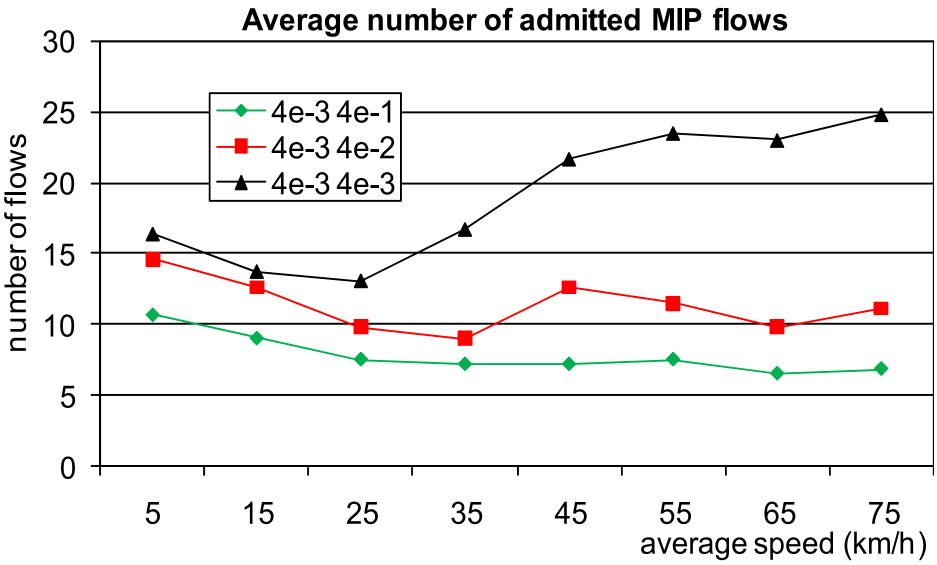


Figure 5.26. Average number of admitted MIP flows.

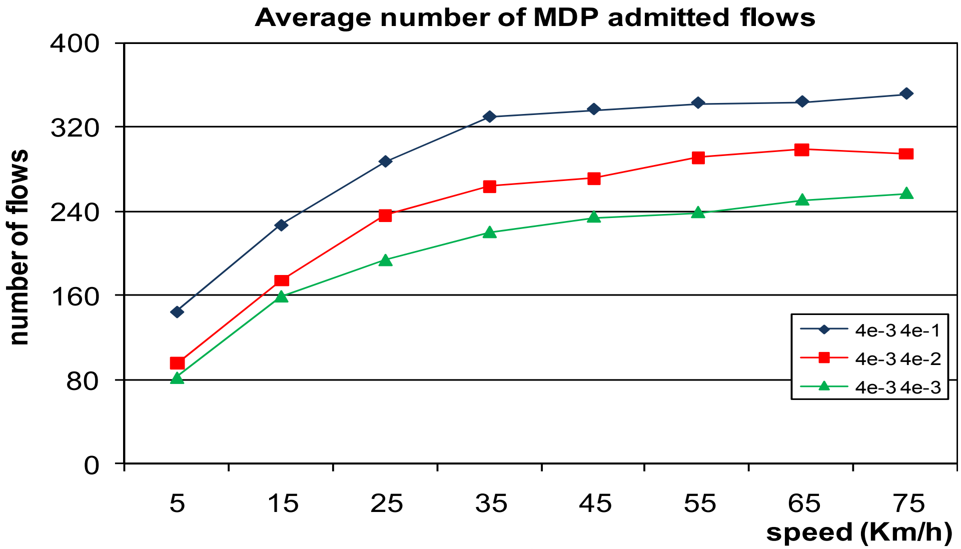


Figure 5.27. Average number of admitted MDP flows.

From figure 5.26 and 5.27 it can be observed that the minimum and maximum values of admitted flows (curves are obtained for the 20%-80% traffic percentages): for both cases they depend on the chosen $p_{threshold}$ value for MDP traffic (in those figures the threshold for MIP users has been fixed to $4*10^{-3}$, in order to guarantee a good level of outage avoidance). MDP users make requests only to current cells, while MIP users make reservations over all system cells, so they are subject to a more strictly CAC policy and the probability of a system admission is lower than MDP. For increasing speed, some observations must be made: for the MDP case, as the average speed increases, the cell-stay time of each user decreases, so more users can find bandwidth availability. For the MIP case, for middle-high ($4*10^{-2}$, $4*10^{-1}$) the effect of MDP does not impact on MIP admission numbers: obviously if the MDP threshold decreases, a higher number of MIP users can be admitted because of the higher bandwidth availability.

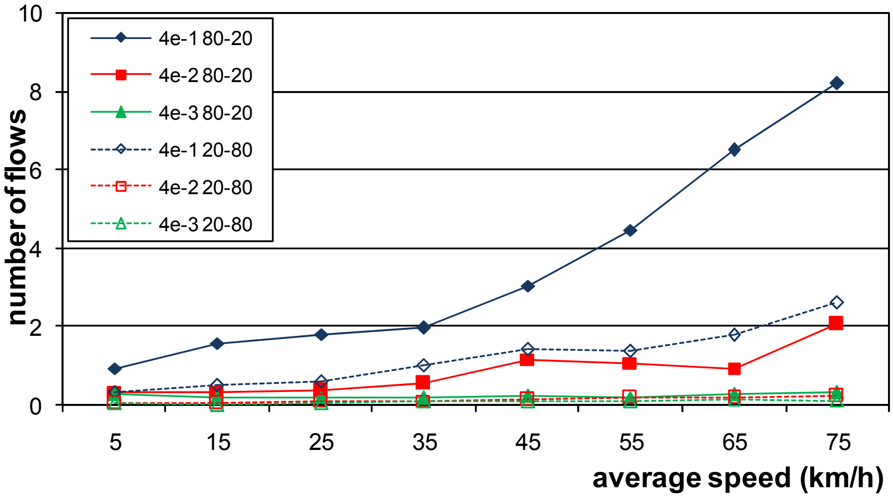


Figure 5.28. Average number of dropped MIP flows.

From figure 5.28, it can be observed that high values of outage threshold (like 0.4) cannot be suffered by MIP flows, because there are too many dropped flows (8 for an average speed of 75km/h), while for low values (like 0.004) the phenomenon can be disregarded. The second column represents the MIP-MDP traffic percentages, while the first one represents the chosen $p_{threshold}$ values.

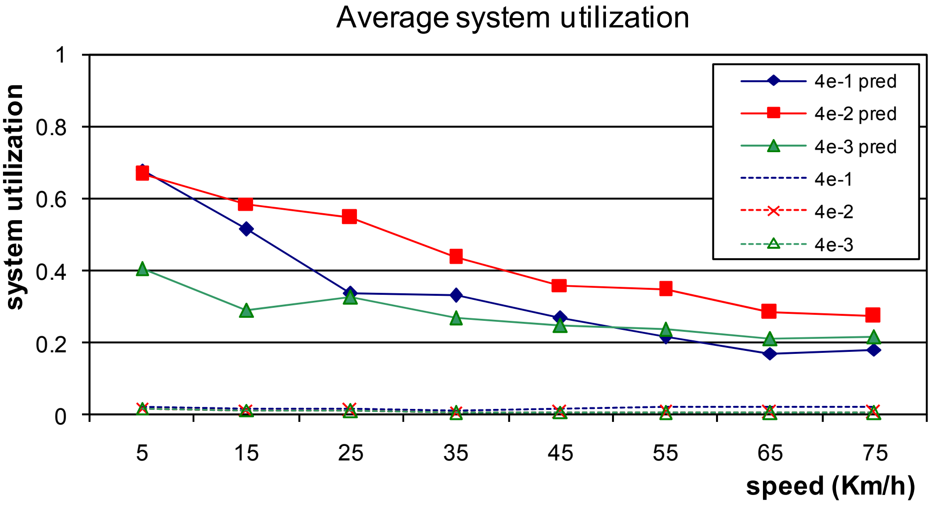


Figure 5.29. Average system utilization.

Figure 5.29 shows the obtained results for the average system utilization by MIP users, with different values of maximum desired outage probability: $4*10^{-1}$, $4*10^{-2}$, $4*10^{-3}$ to 10% and a MIP-MDP traffic percentage of 80%-20%. Full reservation (dotted lines with passive reservations over all network cells) and predictive reservation (continuous lines with passive reservation over a cells circle of radius C_r as in eq. (4.27)) are shown but this time the enhancements are not very exciting, especially for higher speeds: surely the gain of passive reservations here is more evident than the case of MIP-MDP traffic percentages of 20% and 80% illustrated in figure 5.25, because now there is a higher presence of MIP flows (so, passive requests which can be gained). For lower speeds a gain of about 60% in system utilization can be reached, but when the average speed is increased the value of C_p becomes higher so the maximum gain reduces to about 25%, because of the wasted passive bandwidth.

From the curves illustrated in this paragraph, it can be seen that high percentages of MIP traffic lead to a system under- utilization, although the MDP users can use passive resources; in addition high values of outage threshold cause a violation of QoS requirements for MIP users. Moreover, it is shown that a circular partial pre-reservation policy can ensure either wireless QoS during hand-off events or higher system utilization than a full pre-reservation policy; we also showed that our model presents a prediction error (more evident for low speed values and high values of α) that causes negligible effects on QoS guarantees that MIP users require for their connections. The only disadvantage of the proposed scheme is the high resource wastage when mobile hosts move with high speeds: the value of C_r increases with polynomial trend in function of C_p, so the directional treatment is mandatory, as illustrated in previous chapter.

5.5 Simulation results of the proposed 2D direction-aware reservation scheme

Remembering that there are no rules about choosing the value of n (paragraph 4.6.1), simulations results shown that n=6 is a good trade-off between accuracy and computational complexity of the proposed algorithm; higher values of n make better the approximation of the wireless cell coverage area but make worse the spent computational time. Simulation results of the dynamic-scheme (paragraph 4.6.3) are compared with those of the static-scheme (paragraph 4.6.2) and the obtained enhancements are shown. The same mobility parameters of figure 4.16 have been considered.

First of all the HDP matrix *M* have been filled up, so a certain number of monitor simulations have been launched; in particular we executed 1000 monitor runs with a single duration of T_{sim}=400s (the chosen duration ensures the satisfaction of the confidence interval of 95%). After a statistical analysis of the obtained values with MATLAB tool, the matrix *M* of figure 4.26 has been generated. The static algorithm have been tested with the following input parameters:

a) constant trend: $C_{i1}=C_{i2}=C_{i3}=1$;

b) increasing trend: C_{i1}=1, C_{i2}=2, C_{i3}=3;

c) decreasing trend: C_{i1}=3, C_{i2}=2, C_{i3}=1 with C_{ij}=1 $\forall j \in \{4..C_{ei-1}\}$.

As earlier discussed in previous chapter, for the dynamic case, some different expressions of the exponent function $f(k)$ can be considered; in particular we considered four different expressions: a) $f(k)=1$; b) $f(k)=\alpha k$; c) $f(k)=\alpha/k$ and d) $f(k)=(\alpha k)^{-1}$, with α>0, in order to appreciate how δ is weighted for consecutive values of k. In this paragraph, only the results for case b) with $\Delta=\delta^{\alpha k}$ are considered, because the employing of the αk exponent into the dynamic algorithm leads to better results. After a deep analysis of the possible values of α the value α=1.12 has been chosen, because it guarantees the optimal performances for the chosen exponent function. As for the monitor simulations, the duration have been fixed to T_{sim}=400s for each run. Different campaigns have been carried out, also varying the amount of MIP and MDP traffic percentages; in the following, if 60% is the MIP percentage then, obviously, 40% is the percentage of MDP traffic.

Figure 5.30 depicts the trend of the assigned bandwidth versus the percentage of MIP service requests: in both cases (static and dynamic) the curves have an increasing behaviour if the MIP traffic percentage increases; that is the overall number of admitted flows decreases drastically because higher MIP requests lead the system to make a high amount of passive reservations, so a lower number of users have the chance to enter the system.

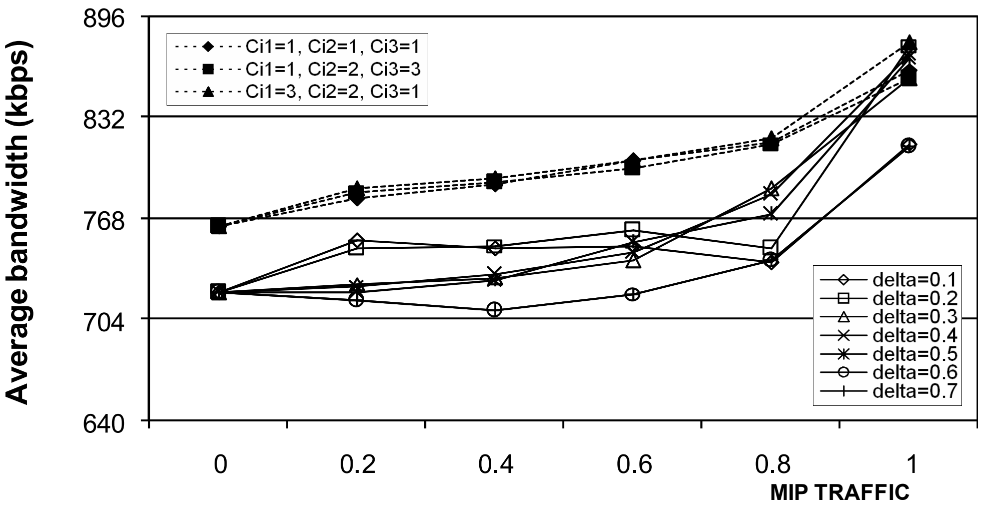


Figure 5.30. Average allocated bandwidth.

In this way, the bandwidth can be shared among a lower number of users and they can receive a better QoS treatment: as it can be seen the average bandwidth goes from about 720kbps to about 880kbps (near the maximum level). It can be also observed that, as it will be explained later, the choice of the input in the static case does not affect the performances in terms of bandwidth, while in the dynamic case lower values of δ offer better performances (the maximum gap is about 40kbps for δ=0.4). In every case, in terms of assigned bandwidth, the static algorithm performs slightly better than the dynamic one.

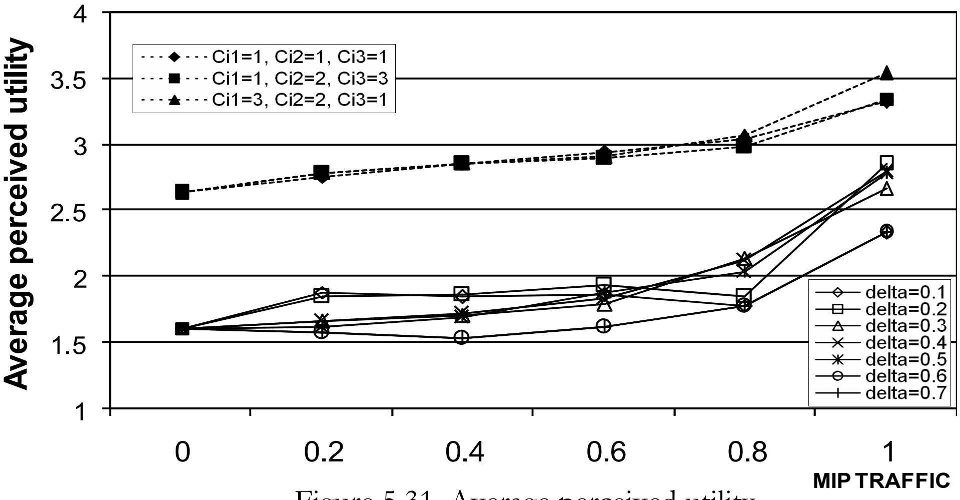


Figure 5.31. Average perceived utility.

Figure 5.31 illustrates how the utility is perceived from users by varying the MIP traffic percentages and some prediction input parameters. As illustrated later, it must be outlined that the wireless channel evolution does not depend on the adopted prediction policy, so it has a similar evolution for both static and dynamic case. In fact, the perceived utility follows the trend of the assigned bandwidth and the same considerations can be made: in both cases (static and dynamic) if the percentage of MIP flows increases the perceived utility is higher, due to the lower number of admitted MDP flows, which makes the bandwidth availability to be higher. Also in this case the static prediction policy offers better performances, while low values of delta are preferred if a dynamic policy is pursued.

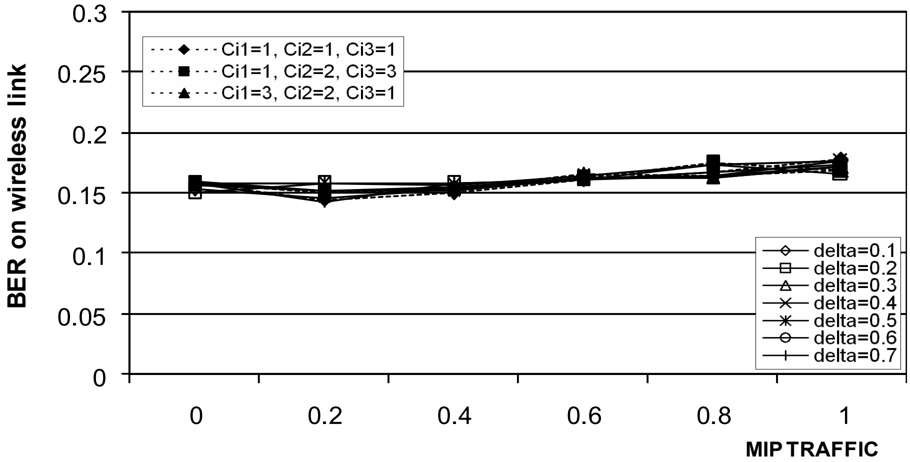


Figure 5.32. Average BER on wireless link.

Figure 5.32 shows the trend of the average BER for different traffic percentages: the wireless link condition does not depend on the amount of MIP or MDP requests or on the adopted prediction policy, so the trend in the figure is nearly constant (it fluctuates from a minimum value of 14.9% to a maximum value of 17.5%); the little variations that can be observed in figure 5.32 are introduced by the stochastic properties of the Markov chain that describes a single wireless link between a mobile host and its coverage Access Point. It can be concluded, as expected, that there are no sensible variations on the channel performances for different traffic conditions or different employed algorithms.

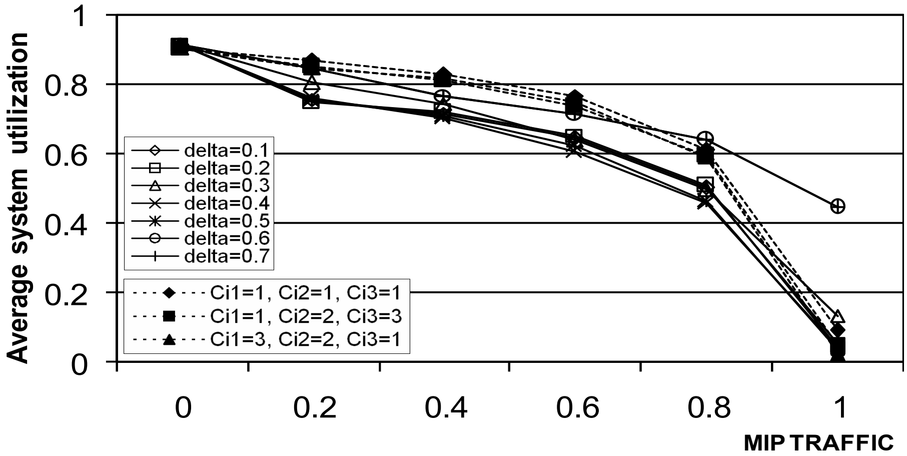


Figure 5.33. Average system utilization vs MIP traffic percentage.

Figure 5.33 shows the average system utilization for increasing values of MIP requests: in both cases the trend is decreasing because of the higher number of passive reservations, with a consequent reduction of MDP admitted flows. When no MIP users make service requests the utilization reach its maximum value, around 92%-93%, because no passive reservations are present in the system; on the contrary, the system is very under-utilized if only MIP users make service requests. The maximum gap between static and dynamic schemes is observed for a MIP percentage of 60% and it is around 10%-12% (the case of 100% MIP is excluded because no MDP are present into the system). As it can be seen the static scheme performs slightly better than the dynamic one in terms of system utilization, except for high percentages of MIP flows, because high values of δ lead to a system utilization of about 46%.

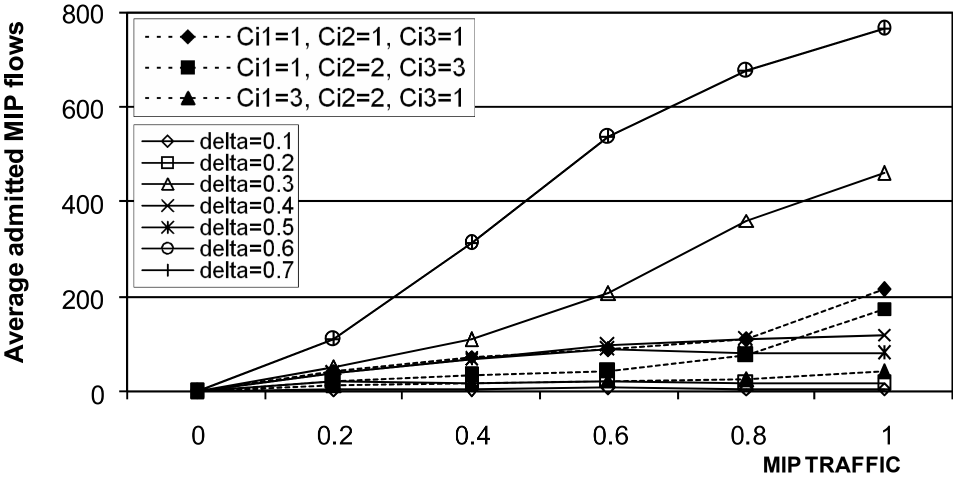


Figure 5.34. Average admitted MIP flows vs MIP traffic percentage.

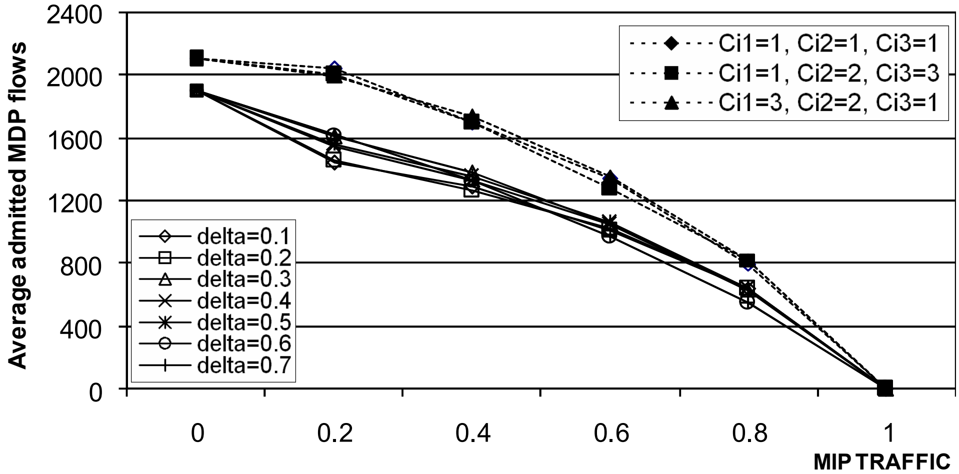


Figure 5.35. Average admitted MDP flows vs MIP traffic percentage.

Figures 5.34 and 5.35 depict the average trend of the number of MIP and MDP admitted flows respectively; either in the static case or in the dynamic one there is an obvious increasing behaviour of MIP admissions for a higher number of MIP requests, while the MDP ones drastically decrease. For the static prediction case the number of admitted MIP flows is often lower than 200 while, in the dynamic case, for δ=0.3, δ=0.6 and δ=0.7 the number of admitted MIP flows is not comparable, because it increases until about 780. Both in the static or dynamic cases the MDP admission is not affected by the chosen policy or the chosen input parameters. The dynamic algorithm ensures higher admission chances for MIP flows, in spite of MDP flows that find a lower amount of available bandwidth if more MIP requests are accepted.

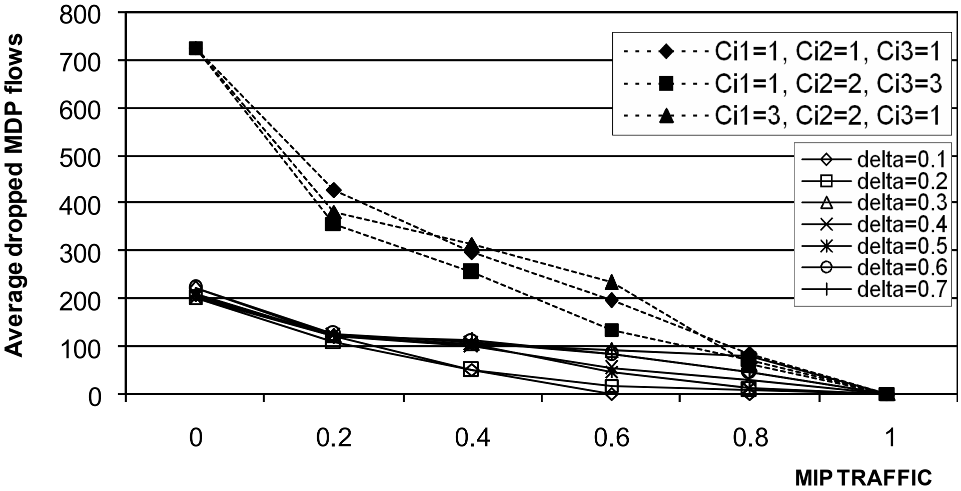


Figure 5.36. Average number of MDP dropped flows vs MIP traffic.

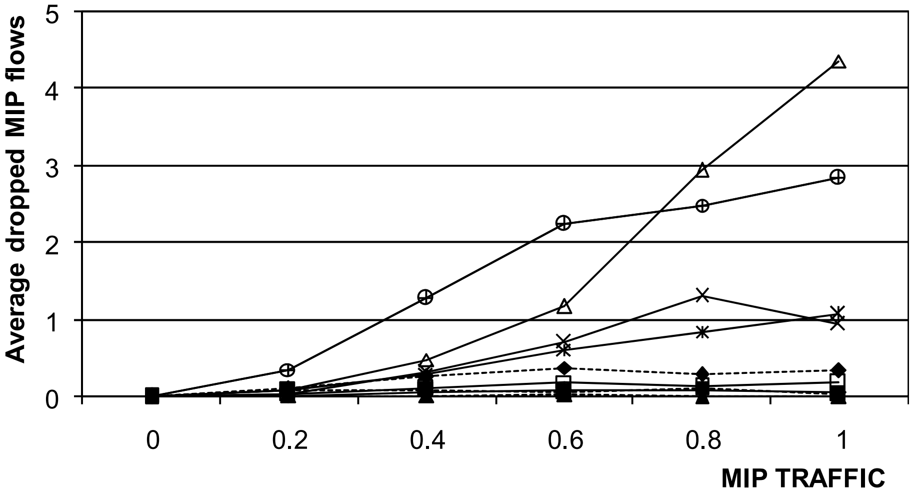


Figure 5.37. Average number of MIP dropped flows vs MIP traffic.

Figures 5.36 and 5.37 illustrate the average number of dropped flows: these curves show how the CAC module and the bandwidth reallocation algorithm guarantee to MIP flows a very low dropping probability. In the dynamic prediction case, the number of MIP dropped flows is higher than the static one, but if it is compared with the average number of admitted MIP flows, the percentage of MIP dropped flows maintains at very low and acceptable level, as well as in the static case. The minimum and maximum MIP dropping observed percentages for the static case are 0.0285% and 0.3%, while for the dynamic prediction case they are 0.31% and 0.87%; it can be concluded that both prediction schemes with the CAC and bandwidth management of chapters 3 and 5 have good performances in terms of percentage of MIP dropped flows, that maintains below 1%. For MDP users the minimum and maximum observed values for static and dynamic prediction schemes are 2.31% - 34.4% and 7.7% - 10.75% respectively: these high values are obtained because the proposed bandwidth management algorithm gives only an outage intra-cell guarantee for MDP users, so they may not find the needed bandwidth after a hand-off event; in addition if a MIP user need some bandwidth to mitigate its outage condition, the algorithm drops a MDP flow if no bandwidth is available.

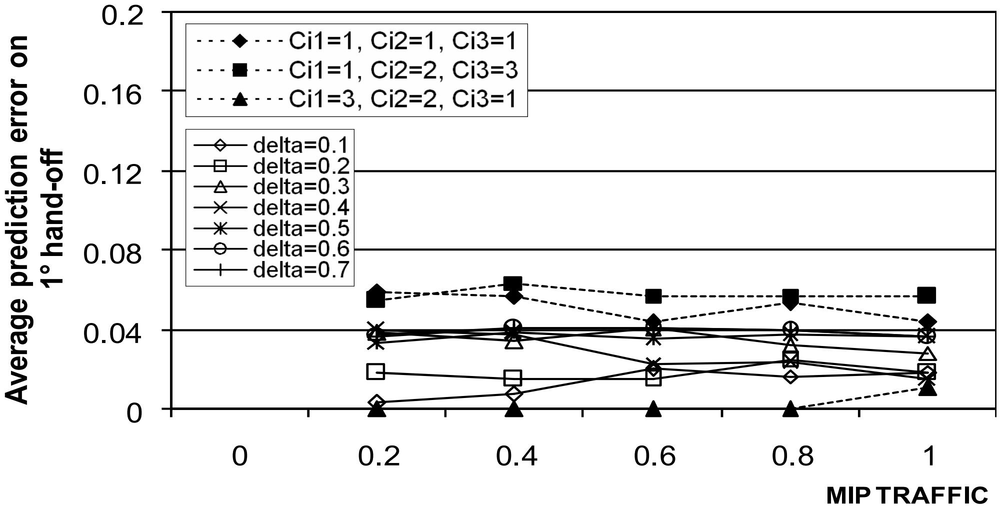


Figure 5.38. Average prediction error on 1° hand-off vs MIP traffic percentage.

In figure 5.38 it can be seen the average error committed in the prediction of possible next cells for the first hand-off event of MIP users in the static and dynamic cases; if *$N_{MIP}(ho)$* is the overall number of MIP users that have made at least *ho* hand-overs from the first cell and *$n_{MIP}(ho)$* is the overall number of MIP users that did not find a passive reservation after the *ho-th* hand-off event, then the prediction error on *ho-th* hand-off event is *$e(ho)=n_{MIP}(ho)/N_{MIP}(ho)$*. In figure 5.38 *e(1)* is shown for static and dynamic prediction schemes: first of all, for a fixed prediction scheme with a fixed set of input parameters the trend is almost constant if the MIP traffic percentage is varied;

obviously, the prediction algorithm is not affected by the number of admitted flows. The static reservation with the sequence C_{i1}=3, C_{i2}=2, C_{i3}=1 has a negligible error because there are three predicted cells for the first hand-off and the probability of an error is near to zero; however, dynamic reservations with low δ values lead to an acceptable value of *e(1)*, lower than 4%.

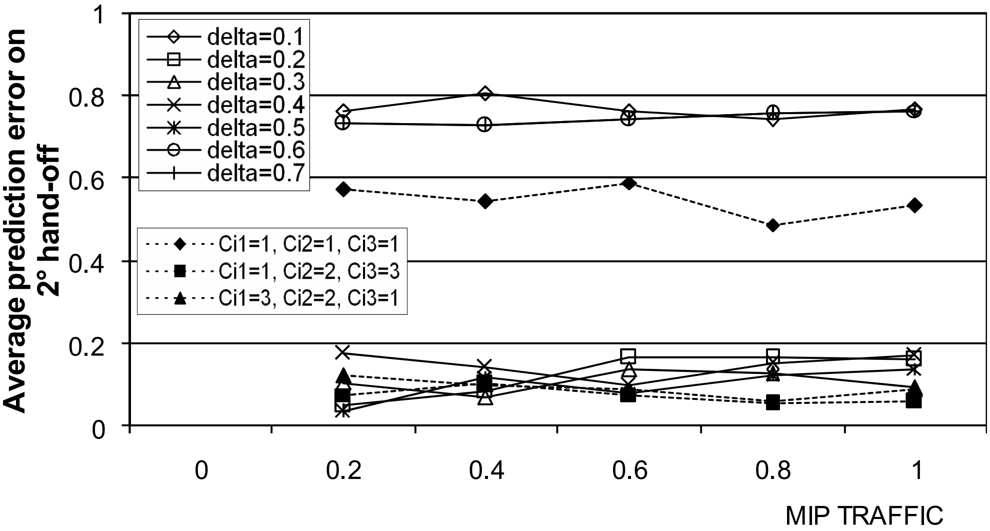


Figure 5.39. Average prediction error on 2° hand-off vs MIP traffic percentage.

Figure 5.39 depicts the trend of *e(2)*. Similar considerations of figure 5.38 can be made but, in this case, some input combinations must be excluded because they make the algorithms have a low precision in prediction making. The static policy offers best performances for the input sequence C_{i1}=1, C_{i2}=2, C_{i3}=3, while the dynamic one performs better for δ=0.3 and δ=0.5. Nevertheless, the best obtained value of *e(2)* is about 10%-12%.

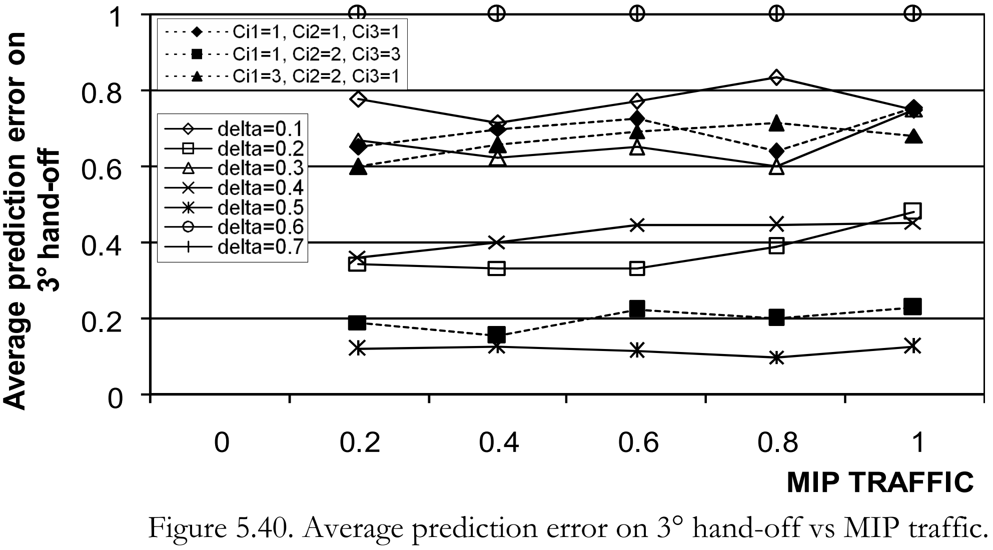


Figure 5.40. Average prediction error on 3° hand-off vs MIP traffic.

Figure 5.40 illustrates the obtained prediction error *e(3)*. As for the results of figure 5.39, some input combinations must be excluded, because they do not lead to any acceptable result; the best results have been obtained for the dynamic case with δ=0.5 and they are not comparable with those obtained with any other input values of the dynamic scheme or with the static one, that has good performances for the input sequence C_{i1}=1, C_{i2}=2, C_{i3}=3. The minimum and maximum error for δ=0.5 are 10% and 12.57%.

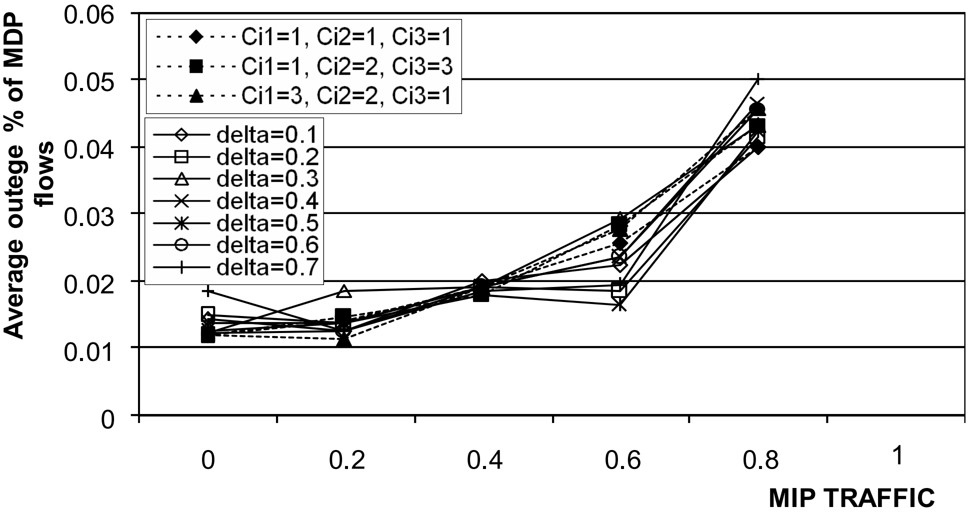


Figure 5.41. Average MDP outage percentage vs MIP traffic percentage.

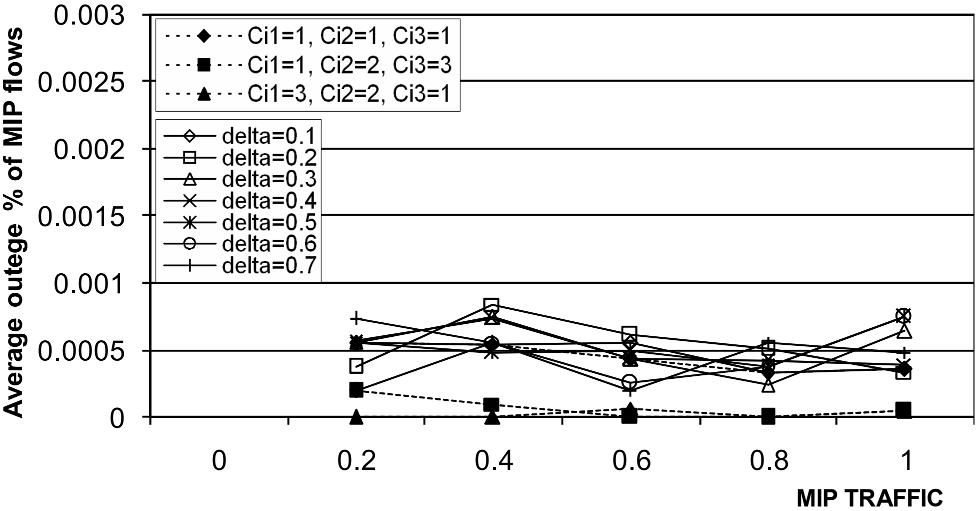


Figure 5.42. Average MIP outage percentage vs MIP traffic percentage.

Figures 5.41 and 5.42 show the average percentage of MDP and MIP users that have suffered an outage event during their active sessions: in our simulations the outage threshold has been fixed to p_{outage}=0.04. As it can be seen, the implemented CAC and bandwidth allocation algorithm ensures the respect of the fixed threshold, in fact the maximum observed values of outage percentage for MDP and MIP users are 0.0502% and 0.00084%, that are quite lower than the imposed constraint; in particular, as earlier discussed, MIP users are advantaged if compared with MDP ones, because of the guaranteed service continuity and the passive reservation policy; in addition the bandwidth allocation algorithm privileges MIP users when choosing a benefactor that has to give up a portion of its allocated resources. In terms of outage percentage of MIP and MDP flows no large differences can be observed between static and dynamic schemes. The increasing trend of the curves in figure 5.41 does not depend on the adopted prediction policy, but it is due to the higher presence of MIP flows, that have a certain bandwidth allocation priority.

From the figures above it can be concluded that both static and dynamic prediction schemes are able to guarantee a good level of QoS for different input parameters; in particular, the static scheme is able to ensure a slight higher level of bandwidth and utility (figures 5.30 and 5.31) with an acceptable system utilization (figure 5.33); the percentage of dropped MIP flows is also maintained under the value of 1% (figures 5.34 and 5.37). The best obtained input sequence for the static scheme is C_{i1}=1, C_{i2}=2, C_{i3}=3; this can be explained by considering the increasing of prediction error for higher hand-off events due to the intrinsic error that has been committed in the generic "previous step"; pre-reserving resources on a higher number of cells for the next hand-off event can balance previous prediction errors. The dynamic scheme offers slight lower performances in terms of amount of assigned bandwidth and perceived utility, but it outperforms the static one in terms of prediction error: for the first hand-off only the static sequence C_{i1}=3, C_{i2}=2, C_{i3}=1 leads to a negligible value of *e(1)*, because reserving on C_{i1}=3 cells reduces the probability of error near to zero. However the dynamic threshold-based algorithm performs better in the "long-range" prediction: the maximum value of *e(3)* is 12.57% for δ=0.5.

Concluding, the obtained results have shown that the pre-reservation phase is necessary in wireless environments if a certain level of QoS and service continuity must be ensured (the MDP dropping percentage is too much higher if compared with the one of the MIP traffic, that is lower than 1%). So, the pre-reservation phase is mandatory and two prediction schemes (static and dynamic) have been proposed and tested, in order to realize the "passive reservations" policy. The proposed CAC and bandwidth allocation schemes integrated with the FSMC model and one of the prediction algorithms ensure that the outage percentage is never higher than the fixed $p_{threshold}$. Both prediction algorithms have shown satisfactory results, but the dynamic one led to a more accurate "long-term" prediction: in particular for δ=0.5 the maximum obtained values of *e(1)*, *e(2)* and *e(3)* are 4.1%, 12% and 12.5%. The static scheme offers good performances for the input sequence C_{i1}=1, C_{i2}=2, C_{i3}=3.

5.6 The effects of statistical bandwidth multiplexing for the 2D pre-reservation scheme

The same type of simulations of previous paragraphs have been carried out while taking into account the considerations of paragraph 4.6.4 about the Statistical Bandwidth Multiplexing (SBM). For sake of simplicity, only the results for an exponent function *f(k)=ak* and δ fixed to 0.5 are shown. The results for other kind of exponent functions (*f(k)=a/k*, *f(k)=1/ak* and so on) have been observed to follow a similar trend. Only MIP traffic has been considered. Many campaigns of simulations have been carried out and for all the following figures, the two curves are obtained with the same prediction scheme, but the *NO-MUX* one does not account the CAC multiplexing effect, following the classical admission control (that is to say the call is admitted only on the free available bandwidth *S*).

Figure 5.43 shows the average received bandwidth from MIP users, in function of the number of MIP service requests per second made to the system: the course is slightly decreasing in both cases because of the higher presence of admitted flows and the maximum gap between two curves is about 4-5Kbps; the introduction of the multiplexing scheme does not introduce appreciable enhancements on the received bandwidth.

Figure 5.44 depicts the average system utilization: with the *NO-MUX* approach the obtained values in function of MIP requests per second do not exceed the bound of 10%. A so lower utilization value is unacceptable and it is due to the heavy presence of passive and unused bandwidth. In addition, the trend is constant because the system is always saturated, independently from the number of service requests. Introducing the multiplexing policy, an appreciable enhancement is visible, especially for higher number of requests.

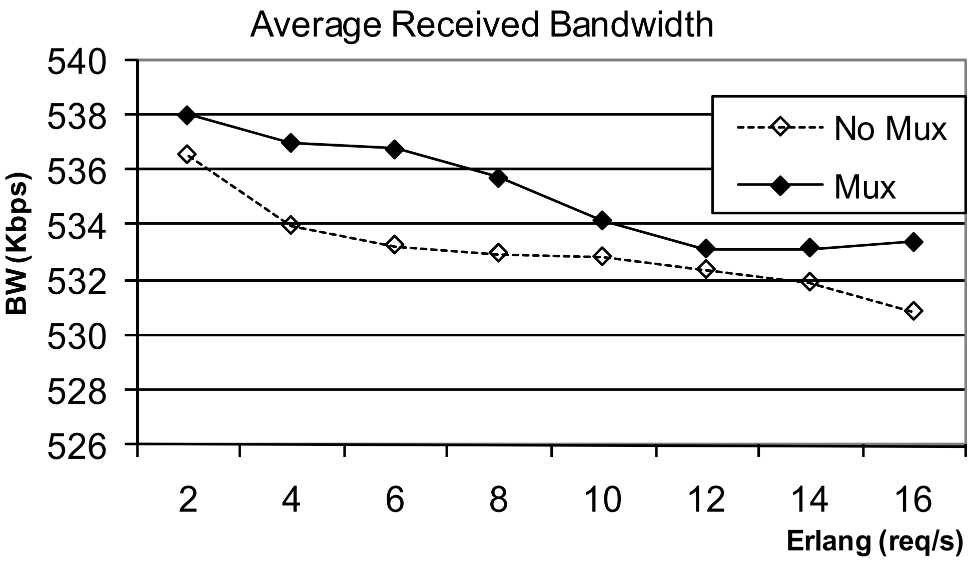


Figure 5.43. Average received bandwidth from MIP users.

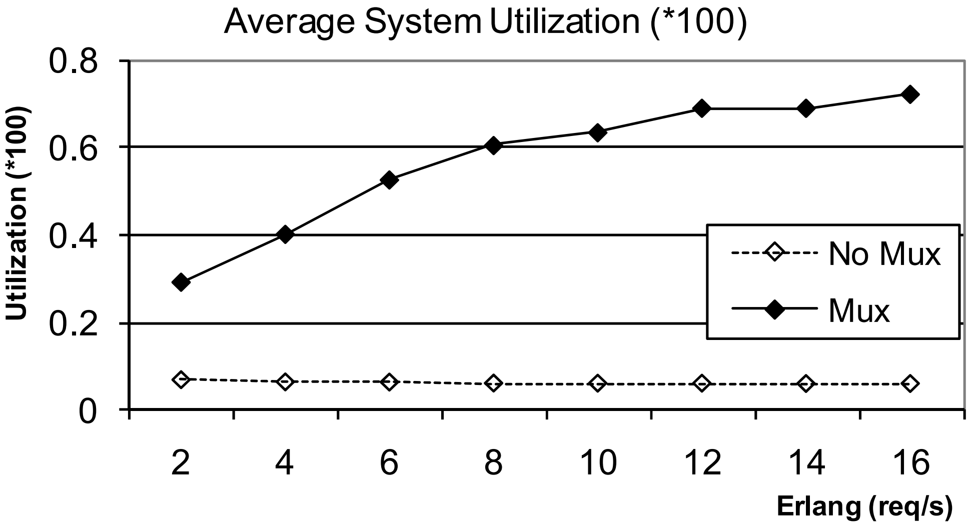


Figure 5.44. Average system utilization.

The utilization grows up from about 30% to about 70%, and these values are not comparable with those of the classical scheme. So, figure 5.44 shows the obtained enhancements in terms of system utilization.

Figure 5.45 depicts the course of the average number of admitted flows; as for the system utilization, the obtained enhancements are evident, especially for high values of requests per second. The number of admitted flows into the system goes up near to 1450, versus about 100 admitted flows of the *NO-MUX* classical scheme (also in this case the system is always saturated because of the heavy presence of passive and unused bandwidth).

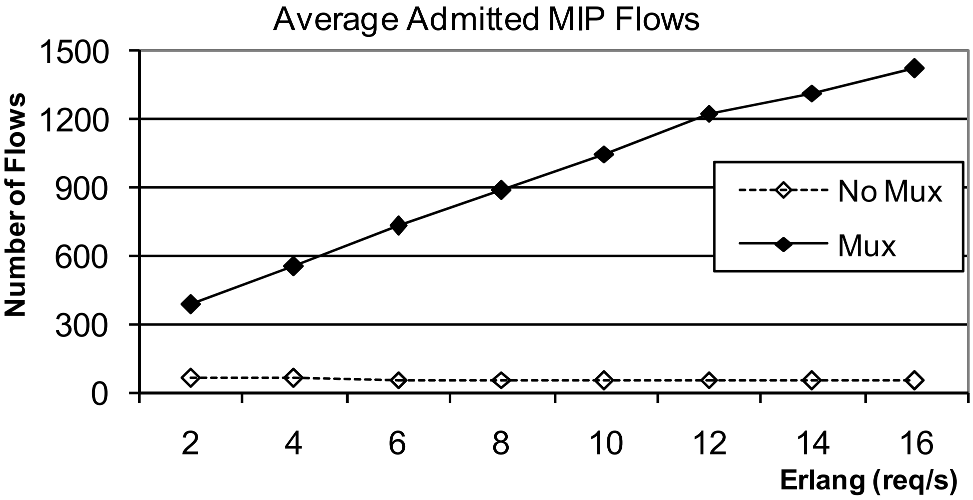


Figure 5.45. Average number of admitted flows.

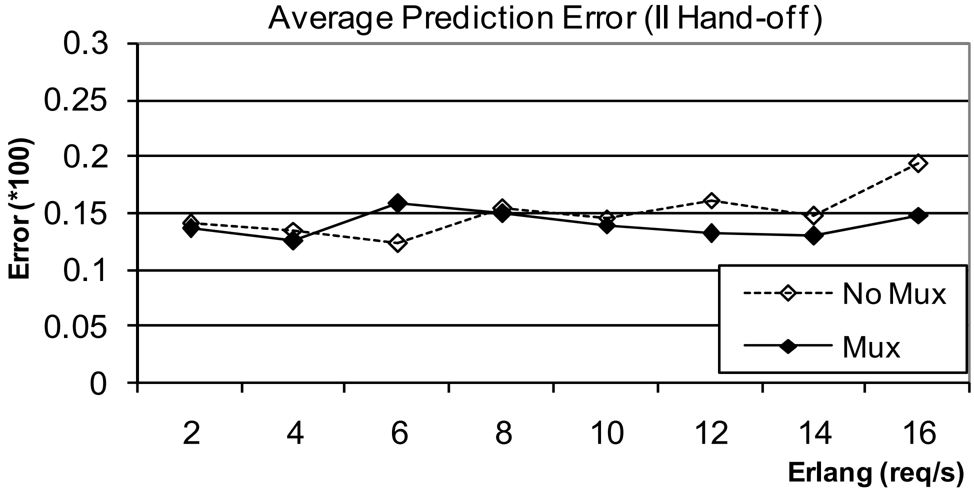


Figure 5.46. Prediction error on 2nd hand-off event.

Figure 5.46 illustrates the average error on predicting future visited cells for second hand-off event (for the first hand-off event the error is negligible as in previous cases); since the multiplexing CAC algorithm does not affect the prediction module, no big differences can be outlined between *MUX* and *NO-MUX* curves. However, the proposed algorithm gives an average error of about 15%, obviously independent from the number of service requests.

5.7 Conclusions on chapter 5

In this chapter of the PhD thesis the performance evaluation of the proposed ideas has been performed. The network simulator implemented by the author has been briefly described as a powerful tool to manage mobile hosts in a wireless environment. Initially the 1D CST-based prediction scheme has been investigated and good results have been obtained regarding the MIP QoS guarantees: system utilization can considerably increase if passive reservations are made in an adequate manner and service continuity is always guaranteed for MIP service requests. However the simulated 1D scenario is too simply if compared with real mobile environments, so it has been extended with a 2D clustered simulation scenario, where users can move according to the Random WayPoint or Smooth Random Mobility Models. The same CST-based prediction scheme of the 1D case has been employed in the 2D scenario but, although the prediction error is negligible, too many resources are wasted, because of the high number of C_r cells which are interested by passive reservations; so additional and directional information has been introduced in the prediction algorithms in order to make them more selective. Optimal results have been obtained for some combinations of input parameters and the dynamic scheme has resulted to be a good predictor for a value of δ=0.5. After a deep analysis of the QoS that is perceived by mobile MIP and MDP users, the SBM scheme proposed in chapter 4 has been evaluated, illustrating the utilization gain that can be reached with the introduction of the bandwidth multiplexing policy.

Conclusions

The work proposed in this PhD thesis has focused on some actual problems regarding wireless networks, such as link degradations, mobility effects and bandwidth management. An integrated idea, which takes into account all the considered problems, has been proposed, as a possible and efficient solution to some inherent issues of wireless networks, in particular the wirelessLANs. After a brief overview on wireless environments, the fading phenomenon has been deeply described and a channel model based on the Markov processes has been used in order to obtain a powerful way to take into account channel fluctuations during the experiments of the simulation tool. The obtained Markov chain has been "calibrated" through the use of a parameterization algorithm, already present in literature: a new approach for partitioning the received SNR range that enables tractable analysis of the packet loss and delay performance over a time-varying wireless channel has been presented. In this way, the evolution of the obtained chain reflects better real channels behaviours, giving the chance to extract precious information about packet performances.

The last studies about telecommunications and computer-science have increased the need and the availability of communication devices, such as laptops and palmtops; at the same time, the research in digital wireless communications has made possible the connection of notebooks to Internet wherever they are. For these reasons the Quality of Service is always required in wireless connections, in order to give to the final user the right level of satisfaction for the received service. The ISPNs architecture has been introduced as a way to overcome the "Best-Effort" nature of the classical Internet applications. An overview of the RSVP protocol has been given, with its natural extensions for mobile environments: MRSVP and DRSVP.

In the three years of doctoral activity, the attention has been focused on the management of QoS for mobile host, so the MRSVP has been implemented for this purpose. In particular two classes of service have been considered: MIP (for those tolerant applications which require fairly reliable delay bounds in all cells that will be visited, without the affecting of mobility effects) and MDP (for tolerant applications, which can tolerate the effects of delay violations due to mobility of hosts). The importance of passive reservations has been outlined, showing some interesting simulation results that can be obtained when the passive reservations policy is employed, in terms of received bandwidth, offered QoS and system utilization.

The pre-reservation policy is necessary if the independence from mobility must be granted, offering a service continuity also during hand-off events. The main problem of the pre-reservation of passive bandwidth is the a-priori knowledge of the number of cells that the user will visit (in a one-dimensional environment) or what cells the user will visit (in a two-dimensional environment), in order to build a correct MSPEC object in the MRSVP packets.

In this PhD thesis, a novel approach to both admit new calls and adaptively manage their bandwidth has been proposed, in respect of some important criteria, like the fairness and/or the high system utilization. The need of dynamically change the transmission rate during active sessions introduces implicit overhead. It will be also shown that introducing flexibility in the assigned rate leads to better system performances, in terms of admitted calls and system utilization. The concept of utility function has been introduced, as an indicator of the user satisfaction level.

After a description of some important concepts of utility functions, a new bandwidth allocation protocol has been introduced, with the aim of having a new scheme that can be applicable in the ISPNs systems. Since different applications can be introduced in an ISPN system, the proposed idea takes care of considering the specific utility function, as well as the wireless channel modelling. In this way some important goals can be reached: fairness among users belonging to the same class; high system utilization and QoS guarantees. The obtained results have shown that the introduction of a dynamic scheme for bandwidth management increases system performance, in terms of utilization and number of admitted flows. In addition, it has been shown that the introduction of a channel model is mandatory if channel degradations must be taken into account when dimensioning a wireless system or while serving MDP requests.

A deep analysis of users' mobility for wireless environments has been made in chapter 4; the importance of mobility prediction for wireless systems (such as WLANs) has been outlined, introducing some schemes for passive reservations enhancements.

Two mobility models have been considered, RWPMM and SRMM, but the proposed schemes are completely uncorrelated with the employed mobility model. The introduction of the proposed schemes has started with the analysis of a simplified 1D scenario, in order to give some knowledge about the CST evaluation and analysis, in terms of statistical distribution. After the quantitative analysis of CST distribution, an extension has been introduced, by considering the complete 2D space and the directional behavior of mobile hosts. So after a circular reservation, a directional one has been proposed. The obtained CST distributions have been shown, in terms of mean and standard deviation parameters. In addition, a polynomial regression has been shown to be a good way to obtain a CST evaluation by simply introducing system and mobility parameters. The static scheme and the dynamic one have been proposed and formally described by pseudo-code.

The last chapter of this work consists in a good performance analysis of the proposed schemes: initially the 1D CST-based prediction scheme has been investigated and good results have been obtained regarding the MIP QoS guarantees: system utilization can considerably increase if passive reservations are made in an adequate manner and service continuity is always guaranteed for MIP service requests. However the simulated 1D scenario is too simply if compared with real mobile environments, so it has been extended with a 2D clustered simulation scenario, where users can move according to the Random WayPoint or Smooth Random Mobility Models. The same CST-based prediction scheme of the 1D case has been employed in the 2D scenario, with additional and directional information, which has been introduced in the prediction algorithms in order to make them more selective. Optimal results have been obtained for some combinations of input parameters and the dynamic scheme has resulted to be a good predictor. After a deep analysis of the QoS that is perceived by mobile MIP and MDP users, the multiplexing scheme proposed has been evaluated, illustrating the utilization gain that can be reached with the introduction of the considered policy.

Concluding, it is evident that this PhD thesis faces different problems related to wireless communication systems: channel modelling, bandwidth management and mobility pattern prediction. The main contribution of this work is the proposal of a new idea that integrates some solutions to the above issues: having an integrated architecture, able to grant the required QoS constraints, while respecting the fairness principle with high system utilization and low dropping probability is a suitable issue, especially when channel conditions are considered. The proposed scheme has demonstrated to offer good QoS performance, in terms of bandwidth assignments, system utilization, number of admitted flows and flow dropping probability.

Bibliography

[1] W. C. Jakes, "*Microwave Mobile Communications*" (reprint), IEEE Press, 1974.

[2] J. Meurling and R. Jeans, "*The Mobile Phone Book*", Communications Week International, 1994.

[3] Q. Bi, G. L. Zysman and H. Menkes, "*Wireless mobile communications at the start of the 21st century*", IEEE Communications Magazine, vol. 39, pp. 110-116, 2001.

[4] B. O'Hara and A. Petrick, "*The IEEE 802.11 Handbook: A Designer's Companion*" (2nd edn.), IEEE Standards Publications, 2005.

[5] A. F. Molisch, "*Wireless Communications*", IEEE Press, 2005.

[6] N. Guido, "*Reti wireless: architettura, protocollo IEEE 802.11b e relative problematiche*", Computer Science Department (VE).

[7] L. Taylor, "*Hiperlan type 1 Technology Overview*". Uniform Resource Locator: www.hut.fi/~k23458/links/WLANStandards.html.

[8] M. Johnsson "*Hiperlan/2, The Broadband Radio Transmission Technology Operating in the 5 GHz Frequency Band*", H/2 Global Forum, 1999.

[9] J. Haartsen, "*Bluetooth: The Universal Radio Interface for Ad-hoc Wireless Connectivity*", Ericsson Review n. 3, 1998.

[10] D. Sweeney, M. Robert, "*Tutorial Bluetooth*", Virginia Polytechnic Institute, 2000.

[11] J.G. Proakis, "*Digital Communications*", New York, McGraw-Hill.

[12] A.S. Tanenbaum, "*Reti di Computer*", UTET editions.

[13] J. Khun-Jush and P. Schramm, Ericsson Research, Germany, G. and J. Torsner, Ericsson Research, Sweden, "*HiperLAN2: Broadband Wireless Communications at 5 GHz*", IEEE Communications Magazine, 2002.

[14] J. Hung Yeh, J. Cheng Chen and C. Chen Lee, *"WLAN standards: in particular the IEEE 802.11 family"*, IEEE POTENTIALS, 2003.

[15] IEEE, *"Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications"*, ANSI/IEEE Std 802.11, 1999 Edition (R2003).

[16] M. Gast and O'Reilly, *"802.11® Wireless Networks: The Definitive Guide"*, 2002.

[17] E. N. Gilbert, "*Capacity of a burst-noise channel*", Bell Syst. Tech. J.,vol. 39, pp. 1253-1265, 1960.

[18] E. O. Elliot, "*Estimates of error rates for codes on burst-noise channels*", Bell Syst. J., vol. 42, pp. 1977-1997, 1963.

[19] H. S. Wang, N. M. Moayeri, "*Finite-State Markov Channel – A useful model for radio communication channels*", *IEEE Trans. Vehic. Technology*, vol. 44, no. 1, 1995.

[20] R, J, McEliece, W. E. Stark, "*Channels with block interference*", IEEE Trans. Info. Theory, vol. IT-30, pp. 44-53, 1984.

[21] Sun J., "*Investigation on BER performance of CCK modulation in AWGN channels*", West Virginia University.

[22] J. Aràuz, P. Krishnamurthy, "*A study of different Partitioning schemes in first order Markovian models for Rayleigh fading channels*", IEEE, 2002.

[23] Q. Zhang, S. A. Kassam, "*Finite-state Markov model for Rayleigh fading channel*", *IEEE Trans. Communications*, vol. 47, no. 11, pp. 1688-1692, 1999.

[24] M. Hassan, M. Krunz, I. Matta, "*Markov-based channel characterization for tractable performance analysis in wireless packet networks"*, IEEE Trans. Wireless Communications, vol. 3, no. 3, 2004.

[25] D. Mitra, "*Stochastic theory of a fluid model of producers and consumers coupled by a buffer*", *Adv. Appl. Prob.*, vol. 20, pp. 646-676, 1988.

[26] J. Kim, M. Krunz, "*Bandwidth allocation in wireless networks with guaranteed packet-loss performance*", IEEE/ACM Trans. Networking, vol. 8, pp. 337-349, 2000.

[27] M. Krunz, J. Kim, "*Fluid analysis of delay and packet discard performance for QoS support in wireless networks*", IEEE Trans. Select. Areas Commun., vol. 19, pp. 384-395, 2001.

[28] M. Zorzi, R. R. Rao, L. B. Milstein, "*Error statistics in data transmission over fading channels*", IEEE Trans. Commun., vol. 46, pp. 1468-1477, 1998.

[29] A. Viterbi, "*Error bounds for convolutional codes and an asymptotically optimum decoding algorithm*", IEEE Trans. on Information Theory, vol. 13, pp. 206-269, 1967.

[30] J. K. Wolf, "*Efficient maximum-likelihood decoding of linear block codes using a trellis*", IEEE Trans. on Information Theory, vol. 24, pp.76-80, 1978.

[31] R. Steele, L. Hanzo, "*Mobile Communications"* (2nd edn.), John Wiley & Sons Ltd., 1999.

[32] Clark, D.D. Shenker, S. and Zhang, "*Supporting Real-Time Applications in an Integrated Services Packet Network: architecture and mechanism*", Proc. SIGCOMM, 1992.

[33] P. Ferguson and G. Huston, "*Quality of Service: Delivering QoS in the Internet and the Corporate Network*", Wiley Computer Books, New York, NY, 1998.

[34] I. Foster and C. Kesselman, "*The Grid: Blueprint for a New Computing Infrastructure*", Morgan Kaufmann Publishers, San Francisco, California, 1998.

[35] C. Dovrolis and P. Ramanathan, "*A case for relative differentiated services and the proportional differentiation model*", IEEE Network, vol. 13, pp. 26–34, 1999.

[36] C. Dovrolis, D. Stiliadis and P. Ramanathan, "*Proportional differentiated services: Delay differentiation and packet scheduling*", ACM Computer Communication Review, vol. 29, pp. 109–120, 1999.

[37] L. Breslau and S. Shenker, "*Best-effort versus reservations: A simple comparative analysis*", ACM Computer Communication Review, vol. 28, pp. 3–16, 1998.

[38] B. Teitelbaum, S. Hares, L. Dunn, R. Neilson, R. Vishy Narayan and F. Reichmeyer, "*Internet2 qbone: Building a testbed for differentiated services*", IEEE Network, vol. 13, pp. 8–16, 1999.

[39] D. Clark., "*The design philosophy of the DARPA internet protocols*", In SIGCOMM Symposium on Communications Architectures and Protocols, pp. 106–114, Stanford, California, 1988. ACM. Also in Computer Communication Review 18, 1988.

[40] S. Shenker and J. Wroclawski, "*General characterization parameters for Integrated Service network elements*", Request for Comments (Proposed Standard) 2215, Internet Engineering Task Force, 1997.

[41] R. Braden, D. Clark, and S. Shenker, "*Integrated services in the internet architecture: an overview*". Request for Comments (Informational) 1633, Internet Engineering Task Force, 1994.

[42] R. Braden, Ed. L. Zhang, S. Berson, S. Herzog and S. Jamin, "*Resource ReSerVation protocol (RSVP) – version 1 functional specification*". Request for Comments (Proposed Standard) 2205, Internet Engineering Task Force, 1997.

[43] S. Shenker, C. Partridge and R. Guerin, "*Specification of guaranteed quality of service*", Request for Comments (Proposed Standard) 2212, Internet Engineering Task Force, 1997.

[44] J. Wroclawski, "*Specification of the controlled-load network element service*", Request for Comments (Proposed Standard) 2211, Internet Engineering Task Force, 1997.

[45] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang and W. Weiss, "*An architecture for differentiated service*", Request for Comments (Informational) 2475, Internet Engineering Task Force, 1998.

[46] K. Nichols, S. Blake, F. Baker, and D. Black, "*Definition of the differentiated services field (DS field) in the IPv4 and IPv6 headers*", Request for Comments (Proposed Standard) 2474, Internet Engineering Task Force, 1998.

[47] M. May, J. Bolot, A. Jean-Marie and C. Diot, "*Simple performance models of differentiated services schemes for the Internet*", In Proceedings of the Conference on Computer Communications (IEEE Infocom), New York, 1999.

[48] V. Jacobson, K. Nichols and K. Poduri, "*An expedited forwarding PHB*", Request for Comments (Proposed Standard) 2598, Internet Engineering Task Force, 1999.

[49] J. Heinanen, F. Baker, W. Weiss and J. Wroclawski, "*Assured forwarding PHB group*", Request for Comments (Proposed Standard) 2597, Internet Engineering Task Force, 1999.

[50] S. Floyd and V. Jacobson, "*Random early detection gateways for congestion avoidance*", IEEE/ACM Transactions on Networking, vol. 1, pp. 397–413, 1993.

[51] D. D. Clark and Wenjia Fang, "*Explicit allocation of best-effort packet delivery service*", IEEE/ACM Transactions on Networking, vol. 6, pp. 362–373, 1998.

[52] R. Guerin and V. Peris, "*Quality-of-Service in packet networks: Basic mechanisms and directions*", Computer Networks, vol. 31, pp. 169–189, 1999.

[53] I. Stoica, S. Shenker and H. Zhang, "*Core-stateless fair queueing: Achieving approximately fair bandwidth allocations in high speed networks*", ACM Computer Communication Review, vol. 28, pp. 118–130, 1998.

[54] C. Sunghyun, K. G. Shin, "*Adaptive bandwidth reservation and admission control in QoS-sensitive cellular networks*", IEEE Transactions on Parallel and Distributed Systems, vol. 13, issue 9, pp. 882 – 897, 2002.

[55] D. Zhao, X. Shen and J. W. Mark, "*QoS performance bounds and efficient connection admission control for heterogeneous services in Wireless Cellular Networks*", Journal on Wireless Networks, vol. 8, n. 1, pp. 85-95, 2002.

[56] C. Zhu, C. Pei, J. Li, W. Kou, "*QoS-oriented hybrid admission control in IEEE 802.11 WLAN*", 19th International Conference on Advanced Information Networking and Applications, AINA 2005, vol. 1, pp. 484-487, 2005.

[57] I. R. Chen, O. Yilmaz and I. L. Yen, "*Admission Control algorithms for revenue optimization with QoS guarantees in mobile wireless networks*", Wireless Personal Communications vol. 38, pp. 357–376, 2006.

[58] J. Boyle, R. Cohen, D. Durham, S. Herzog, R. Rajan and A. Sastry, "*The COPS (Common Open Policy Service) protocol*", Request for Comments (Proposed Standard) 2748, Internet Engineering Task Force, 2000.

[59] A. Parekh, "*A generalized processor sharing approach to flow control in Integrated Services Networks*", Technical Report LIDS-TR-2089, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, 1992.

[60] M. Karlsson, C. Karamanolis, J. Chase, "*Controllable Far Queuing for meeting performance goals*", International Journal on Performance Evaluation vol. 62, pp. 278–294, 2005.

[61] R. Braden, L. Zhang, S. Berson, S. Herzog, S. Jamin, "*Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification*", Request for Comments 2205, ISI Standards Track IB, 1997.

[62] A. K. Talukdar, B. R. Badrinath, A. Acharya, "*MRSVP: A Resource Reservation Protocol for an Integrated Services Network with mobile hosts*", Wireless Networks Volume 7, Issue 1, pp. 5-19, 2001.

[63] A. K. Talukdar, B. R. Badrinath, A. Acharya, "*Integrated services packet networks with mobile hosts: Architecture and performance*", Journal of Wireless Networks, vol. 5, pp. 111–124, 1999.

[64] Charles Perkins, "*Mobile IP*", Addison-Wesley, Reading 1998.

[65] J. Veizades, E. Guttman, C. Perkins and S. Kaplan, "*Service Location Protocol (SLP)*", Request For Comment 2165, Internet Engineering Task Force, 1997.

[66] A.K. Talukdar, B.R. Badrinath and A. Acharya, "*On accommodating mobile hosts in an integrated services packet network*", Proceedings of the INFOCOM, Japan, 1997.

[67] M. Mirhakkak, N. Schult, and D. Thomson, "*Dynamic bandwidth management and adaptive applications for a variable bandwidth wireless environment*", IEEE Journal on selected areas in Communications, vol. 19, no. 10, 2001.

[68] W. Navidi, T. Camp, "*Stationary distribution of Random Way Point Mobility Model*", IEEE Transaction on Mobile Computing, vol.3, no.1, 2004.

[69] X. Fei, Y. Min-hua, Z. Peng and Z. Hui-min, "*The research on the service rate adaptation in mobile network*", Proceedings of ICCT, 2003.

[70] S. P. Abraham, A. Kumar, "*A stochastic approximation approach for Max-Min fair adaptive rate control of ABR sessions with MCRs*", INFOCOM, 1998.

[71] Y. Cao and V. O. K. Li, "*Scheduling algorithms for broadband wireless networks*", Proc. of the IEEE, Vol. 89, no. 1, pp. 76 - 87, 2001.

[72] W. H.Kuo, W. Liao, "*Utility-based optimal resource allocation in wireless networks*", Proceedings of Globecom, 2005.

[73] M. Xiao, N. B. Shroff and E. K. P. Chong "*A Utility-based power-control scheme in wireless cellular systems*", IEEE/ACM Transactions on Networking, Vol. 11, No. 2, 2003.

[74] G. Bianchi and A.T. Campbell, "*A programmable MAC framework for utility-based adaptive Quality of Service support*", IEEE Journal on Selected Areas in Comm. Vol. 18, pp. 244-255, 2000.

[75] V. A. Siris, B. Briscoe and D. Songhurst, "*Economic models for resource control in wireless networks,*" IEEE PIMRC, Lisbon, Portugal, 2002.

[76] S. Shenker, "*Fundamental design issues for the future Internet*", IEEE J-SAC, vol. 13, pp. 1176-1188, 1995.

[77] C. Bettstetter, "*Mobility modeling in wireless networks: categorization, smooth movement, and border effects*", Mobile Computing and Communications Review, vol. 5, no. 3, 2001.

[78] G. Y. Liu, G. Q. Maguire Jr., "*A predictive mobility management algorithm for wireless mobile computing and communications*", International Conference on Personal Communications, pp. 268-272, 1995.

[79] L. Lu, J. Wu, W. Chen, "*The study of handoff prediction schemes for resource reservation in mobile multimedia wireless networks*", International Journal of Communication Systems, vol. 17, pp. 535-552, 2004.

[80] B. Epstein, M. Schwartz, "*Reservation strategies for multi-media traffic in a wireless environment*", Proceedings of the 45th IEEE VTC, Chicago, U.S.A., pp. 165-169, 1995.

[81] D. Zhao, X. Shen, J. Mark, "*Efficient call admission control for heterogeneous services in wireless mobile ATM networks*", IEEE Communications Magazine, vol. 38, pp. 72-78, 2000.

[82] G. Kuo, P. Ko, M. Kuo, "*A probabilistic resource estimation and semi-reservation scheme for flow-oriented multimedia wireless networks*", IEEE Communications Magazine, vol. 39, pp. 135-141, 2001.

[83] I. Akyildiz, W. Wang, "*The predictive user mobility profile framework for wireless multimedia networks*", IEEE/ACM Transactions on Networking, vol. 12, no. 6, 2004.

[84] A. Aljadhai, T. Znati, "*Predictive mobility support for QoS provisioning in mobile wireless environments*", IEEE Journal on Selected Areas of Communications, vol. 19, pp. 1915-1931, 2001.

[85] T. Ng, I. Stoica, H. Zhang, "*Packet fair queueing algorithms for wireless networks with location-dependent errors*", Proceedings of IEEE INFOCOM, pp. 1103–1111, 1998.

[86] L. Song, U. Deshpande, C. Kozat, D. Kotz, R. Jain, "*Predictability of WLAN Mobility and its Effects on Bandwidth Provisioning*", Proceedings of IEEE INFOCOM, 25th IEEE International Conference on Computer Communications, 2006.

[87] C. Montgomery, "*Applied statistics and probability for engineers*", Third Edition, Wiley, 2003.

[88] J.Banks, J.S. Carson et al., "*Discrete-Event system simulation,*" Third Edition, Prentice Hall, 2001.

[89] M.A. Stevens, R.B. D'Agostino, "*Goodness of Fit Techniques*", Marcel Dekker, New York, 1986.

[90] C. Bettstetter, H. Hartenstein, X. Pérez-Costa, "*Stochastic properties of the Random Waypoint Mobility Model*", ACM/Kluwer Wireless Networks, vol. 10, no. 5, 2004.

[91] J. Yoon, M. Liu, B. Noble, "*Random Waypoint considered harmful*", Proc. IEEE INFOCOM Conf., pp. 1312-1321, 2003.

[92] J. Le Boudec, "*On the stationary distribution of speed and location of Random Waypoint*", IEEE Trans. Mobile Computing, vol. 4, no. 6, 2005.

[93] M. Zoonozi, P. Dassanayake, "*User mobility modelling and characterization of mobility patterns*", IEEE Journal on Select. Area on Communication, vol.15, no.7, 1997.

[94] T. Liu, P. Bahl, I. Chlamtac, "*Mobility modelling, location tracking and trajectoty prediction in wireless ATM networks*", Journal on Select. Area on Communication, vol.16, no.6, Aug.1998.

[95] J. S. Carson, J. Banks – "*Discrete-Event System Simulation*", Prentice Hall, 1984.

[96] G. Agha, "*ACTORS, a model for concurrent computation in distributed systems*", The MIT Press, Cambridge, 1986.

[97] De Rango F., Fazio P., Marano S., "*Utility-based adaptivity and partial resource reservation in wireless/mobile multimedia networks*", 1th International Symposium on Wireless Communication Systems (ISWCS'04), Mauritius, USA, 2004.

[98] De Rango F., Fazio P., Marano S., "*Resource reservation and utility based rate adaptation in wireless LAN with slow fading channels*", 7th International Symposium on Wireless Personal Multimedia Communications (WPMC'04), Abano Terme, Italy, 2004.

[99] De Rango F., Fazio P., Marano S., "*Mobility Indipendent and Dependent Predictive services management in wireless/mobile multimedia network*", IEEE Vehicular Technology Conference (VTC Fall 2004), Los Angeles, CA, USA, 2004.

[100] De Rango F., Fazio P., Marano S., "*Adaptive reservation in WLAN networks under Smooth Random Mobility Model*", 62th Vehicular Technology Conference (VTC Fall 2005), Texas, USA, 2005.

[101] De Rango F., Fazio P., Marano S., "*Cell stay time prediction for Mobility Independent Predictive Services in wireless networks*", IEEE Wireless Communication and Networking Conference (WCNC' 05), New Orleans, Los Angeles, USA, 2005.

[102] De Rango F., Fazio P., Marano S., "*An active reservation scheme with predictive estimation in WLAN networks under 2D mobility models*", 13rd International Conference on Telecommunications (ICT 2006), Madeira, Portugal, 2006.

[103] De Rango F., Fazio P., Marano S., "*Mobility prediction and resource reservation in WLAN networks under a 2D mobility models*", 63rd Vehicular Technology Conference (VTC Fall 2006), Montreal, Canada.

[104] De Rango F., Fazio P., Marano S., "*Mobility Independent Predictive services in WLAN networks with predictive reservation policy under a 2D mobility model*", Wireless Telecommunication Symposium (WTS 2006), Pomona (USA), 2006.

[105] De Rango F., Fazio P., Marano S., "Cell Stay Time analysis under Random Way Point Mobility model in WLAN networks", IEEE Communications Letters, Vol. 10, n. 11, pp. 763-765, 2006.

[106] De Rango F., Fazio P., Marano S., "*A 2D cell stay time and movement direction-based reservation scheme for WLAN clusters with active and passive advanced reservations*", 18th IEEE Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 2007), Athene, Greece, 2007.

[107] De Rango F., Fazio P., Marano S., "*A new 2D direction-based predictive reservation scheme for WLAN environment with passive advanced reservations*", IEEE Wireless Communication and Networking Conference (WCNC 2007), Hong Kong, China, 2007.

[108] De Rango F., Fazio P., Marano S., "*A new threshold-based predictive reservation scheme*", 65th Vehicular Technology Conference (VTC Spring 2007), Dublin, Ireland, 2007.

ACKNOWLEDGEMENTS

Arcavacata di Rende, November 20th,

I would like to thank some people who have encouraged me in this long path.

Thanks to my father Tommaso, who has prematurely left me in 2005. He has always given me some precious words, in order to face the ordinary problems of life. I am in debt to him, but I cannot reply what he has made for me in his life. I only hope he will be happy for my work and my behaviour, that is everyday inspired to his teachings. Dad, please stay near me everyday of my life, I miss you so much.
Thanks to my mother Rita for her worries about me and for everything she has made for me as an affectionate and sweet mother.
Thanks to my unlucky brother Francesco, far having encouraged me in every bad moment.
Thanks to my grandparents Maria, Erminia, Francesco and Peppino.
Thanks to Sabrina, for having beard me since two years. I strongly hope it can continue.

I cannot forget to thank my co-workers: thanks to Mauro for his great regard for me and for his precious company during work-hours; thanks to Floriano for his precious words, thanks to Fiore (called "My Disciple"), Franco (called "Mr Mitsuyoshi Anzai"), Andrea and Antonio (called, "My Friend of coffee").

Thanks to my tutor Prof. Salvatore Marano for the great coordination work that he has made for me during these long years.

Thanks to Antonio Carpino and his wife for their contribution to this work.

Thanks to all my friends and all the people who helped me to reach this finishing line. Thanks so much.

SHORT BIOGRAPHY

Peppino Fazio was born in Catanzaro, Italy, in 1977. He received the Degree in Computer Science Engineering with Electronic and Telecommunication specialization in 2004 from the University of Calabria, Faculty of Computer Science Engineering, D.E.I.S. Dept., Arcavacata, Rende, Cosenza, Italy.

In November 2005 he joined the D.E.I.S. Dept. as PhD student. His main research interests are WirelessLAN, Quality of Service architectures, Mobility models and mobility prediction algorithms.

LIST OF PUBLICATIONS

[1] De Rango F. , Fazio P. , Veltri F. , Marano S. , " Interference Aware Routing Protocols over Ad Hoc UWB Networks". "4th International Symposium on Wireless Communication Systems (ISWCS 2007)", Trondheim, Norway, Oct. 17-19, 2007.

[2] De Rango F. , Tropea M. , Santamaria A. , Veltri F. , Fazio P. , Marano S. , " Multi-Mode DVB-RCS Satellite Terminal with Software Defined Radio". "IEEE Wireless Communication and Networking Conference (WCNC 2007)", Hong Kong, China, March 11-15, 2007.

[3] De Rango F. , Fazio P. , Marano S. , " A New Threshold-Based Predictive Reservation Scheme". "65th Vehicular Technology Conference (VTC Spring 2007)", Dublin, Ireland, Apr. 22-25, 2007.

[4] De Rango F. , Veltri F. , Fazio P. , Marano S. , " BER Regression Analysis of DS-UWB based WPAN". "65th IEEE Vehicular Technology Conference (VTC Spring 2007)", Dublin, Ireland, Apr. 22-25, 2007.

[5] De Rango F. , Fazio P. , Marano S. , " A new 2D Direction-Based Predictive Reservation Scheme for WLAN Environment with Passive Advanced Reservations". "IEEE Wireless Communication and Networking Conference (WCNC 2007)", Hong Kong, China, March 11-15, 2007.

[6] De Rango F. , Veltri F. , Fazio P. , Marano S. , " Markov Chain Channel Modeling based on Degradation Level Concept for Ultra Wideband WPAN". "10th International Symposium on Performance Evauation Of Computer and Telecommunication Systems (SPECTS 2007)", San Diego, CA, USA, July 16-18, 2007.

[7] De Rango F. , Santamaria A. , Veltri F. , Tropea M. , Fazio P. , Marano S. , " Multi-Satellite DVB-RCS System with RCST based on Software Defined Radio". "65th IEEE Vehicular Technology Conference (VTC Fall 2007)", Baltimore, USA, Oct. 22-25, 2007.

[8] De Rango F. , Fazio P. , Marano S. , " A 2D Cell Stay Time and Movement Direction-Based Reservation Scheme for WLAN Clusters with Active and Passive Advanced Reservations". "18th IEEE Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 2007)", Athene, Greece, Sept. 3-7, 2007.

[9] De Rango F. , Veltri F. , Fazio P. , Santamaria A. , Tropea M. , Marano S. , " FER Regression Analysis of DS-UWB-based WPAN". "Wireless Telecommunication Symposium (WTS 2007)", Pomona, CA, USA, Apr. 27-29, 2007.

[10] De Rango F. , Fazio P. , Marano S. , " Mobility Independent Predictive Services in WLAN Networks with Predictive Reservation Policy under a 2D Mobility Model". "Wireless Telecommunication Symposium (WTS 2006)", Pomona (USA), 2006.

[11] De Rango F. , Fazio P. , Veltri F. , Marano S. , " Distance-Dependent BER Evaluation of DS-SS IEEE 802.15.3a Physical Layer under Multiple User Data-Rates and Multi-User Interference". "13rd International Conference on Telecommunications (ICT 2006)", Madeira (Portugal), 2006.

[12] De Rango F. , Fazio P. , Veltri F. , Marano S. , " PER Analysis and Performance Evaluation of DS-SS UWB Networks". "17th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 2006)", Helsinki, Finland, Sept. 11-12, 2006.

[13] De Rango F. , Veltri F. , Santamaria A. , Tropea M. , Fazio P. , Marano S. , " Software Defined Radio based Multi-Mode Satellite Terminal over DVB-RCS Platform". "12th Ka and Broadband Communications Conference", Naples, Italy, Sept. 27-29, 2006.

[14] Fazio P. , De Rango F. , Veltri F. , Marano S. , " Performance Evaluation of the packet Error Rate of DS-SS Physical layer in UWB Networks". "Canadian Conference on Electrical and Computer Engineering (CCECE 2006)", Ottawa (Canada), May 7-10, 2006.

[15] De Rango F. , Fazio P. , Veltri F. , Marano S. , " Time and Distance Dependent UWB Channel Modelling: BER and PER Evaluation for DS-SS Modulation". "63rd Vehicular Technology Conference (VTC Fall 2006)", Montreal, Canada, Sept. 25-28, 2006.

[16] De Rango F. , Fazio P. , Marano S. , " Mobility Prediction and Resource Reservation in WLAN Networks under a 2D Mobility Models". "63rd Vehicular Technology Conference (VTC Fall 2006)", Montreal, Canada, 2006.

[17] De Rango F. , Fazio P. , Marano S. , " An Active Reservation Scheme with Predictive Estimation in WLAN Networks under 2D Mobility Models". "13rd International Conference on Telecommunications (ICT 2006)", Madeira (Portugal), 2006.

[18] De Rango F. , Fazio P. , Veltri F. , Marano S. , " DS-SS UWB Wireless Personal Area Network: BER and PER Evaluation under a MMSE Receiver". "International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS 2006)", Calgary, Canada, July 31 - August 2, 2006.

[19] De Rango F. , Fazio P. , Marano S. , " Cell Stay Time Prediction for Mobility Independent Predictive Services in Wireless Networks". "IEEE Wireless Communication and Networking Conference (WCNC'05)", New Orleans, Los Angeles, USA, 13-17 March, 2005.

[20] De Rango F. , Tropea M. , Fazio P. , Marano S. , " Call Admission Control with Statistical Multiplexing for Agrregate MPEG Traffic in a DVB-RCS Satellite Network". "IEEE Global Telecommunication Conference (Globecom'05)", St.Louis, MO, USA, 28 Nov. - 2 Dec., 2005.

[21] De Rango F. , Fazio P. , Marano S. , " Adaptive Reservation in WLAN Networks under Smooth Random Mobility Model". "62th Vehicular Technology Conference (VTC Fall 2005)", Texas, USA, 25-28 Sept., 2005.

[22] De Rango F. , Fazio P. , Marano S. , " Mobility Indipendent and Dependent Predictive Services Management in Wireless/Mobile Multimedia Network". "IEEE Vehicular Technology Conference 2004 Fall", Los Angeles, CA, USA, September, 2004.

[23] De Rango F. , Fazio P. , Marano S. , " Resource Reservation and Utility based Rate Adaptation in Wireless LAN with Slow Fading Channels". "7th International Symposium on Wireless Personal Multimedia Communications (WPMC'04)", Abano Terme, Italy, 12-15 Sept., 2004.

[24] De Rango F. , Fazio P. , Marano S. , " Utility-based Adaptivity and Partial Resource Reservation in Wireless/Mobile Multimedia Networks". "1th International Symposium on Wireless Communication Systems (ISWCS'04)", Mauritius, USA, 20-22 Sept., 2004.

[25] De Rango F. , Fazio P. , Marano S. , " Cell Stay Time Analysis under Random Way Point Mobility Model in WLAN Networks". IEEE Communications Letters, 2006, Vol. 10, n. 11, pp. 763-765.

[26] De Rango F. , Tropea M. , Fazio P. , Marano S. , " Overview on VoIP: Subjective and Objective Measurement Methods". International journal of computer science and network security, 2006, Vol. 6, n. 1B, pp. 140-153.

[27] De Rango F. , Tropea M. , Fazio P. , Marano S. , " A Scalable Approach for QoS Management in Next Generation Multimedia GEO-Satellite Networks". ASSI Satellite Communication Letter, 2006, Vol. 6, n. 1, pp. 10-21.

Submitted:

De Rango F., Fazio P., Marano S., "Resource Reservation and Call Admission Control for Adaptive Wireless Network Under 2D Mobility Model", Wireless Network Journal, 2007.

De Rango F., Fazio P., Marano S., "Utility-Based Predictive Services for Adaptive Wireless Networks with Mobile Hosts", IEEE Transaction on Vehicular Technologies, 2007.