

**EVEREST**Information Society
Technologies**EVEREST IST-2002-001858****D20*****Final report on the evaluation of RRM/CRRM algorithms*****Contractual Date of Delivery to the CEC: 31.10.2005****Actual Date of Delivery to the CEC: 18.11.2005****Editor: Oriol Sallent (UPC)****Author(s): see list****Participant(s): UPC, KCL, PTIN, TID, TEL, TI****Workpackage: WP3****Est. person months: 34****Security: PU****Nature: Report****Version: 001****Total number of pages: 317****Abstract:**

This deliverable provides a definition and a complete evaluation of the RRM/CRRM algorithms selected in D11 and D15, and evolved and refined on an iterative process. The evaluation will be carried out by means of simulations using the simulators provided at D07, and D14

Keyword list: UTRAN, GERAN, WLAN, heterogeneous networks, RRM, CRRM

DISCLAIMER

The work associated with this report has been carried out in accordance with the highest technical standards and the EVEREST partners have endeavoured to achieve the degree of accuracy and reliability appropriate to the work in question. However since the partners have no control over the use to which the information contained within the report is to be put by any other party, any other such party shall be deemed to satisfied itself as to the suitability and reliability of the information in relation to any particular use, purpose or application.

Under no circumstances will any of the partners, their servants, employees or agents accept any liability whatsoever arising out of any error or inaccuracy contained in this report (or any further consolidation, summary, publication or dissemination of the information contained within this report) and/or the connected work and disclaim all liability for any loss, damage, expenses, claims or infringement of third party rights.

DOCUMENT HISTORY

Date	Version	Status	Comments
20-04-05	001	Int	Preliminary version for discussion
01-06-05	002	Int	First run of partner's contribution
01-09-05	003	Int	Second run of partner's contribution
19-10-05	004	Int	Final Partner's Contribution
2-11-05	005	Int	Section conclusion: Partner contributions
15-11-05	006	Int	Final version for PCC approval
18-11-05	001	Apr	Aproved fina version

Authors List

Andres ALAYON-GLASUNOV (TEL)
Teresa ALMEIDA (PTIN)
Andrea BARBARESI (TI)
Sergio BARBERIS (TI)
Paola BERTOTTO (TI)
Filipe CABRAL-PINTO (PTIN)
Ferran CASADEVALL (UPC)
Massimo COLONNA (TI)
Anders DAHLÉN (TEL)
Per EMANUELSSON (TEL)
Robert FAROTTO (TI)
Vassilis FRIDERICKOS (KCL)
Ali GHORASHI (KCL)
Xavier GELABERT (UPC)
Álvaro GOMES (PTIN)
Héctor GONZÁLEZ SANCHÍS (TID)
Paolo GORIA (TI)
Mikio IWAMURA (KCL)
Peter KARLSSON (TEL)
Rickard LJUNG (TEL)
Jakub MAJKOWSKI (UPC)
Valdemar MONTEIRO (PTIN)
Nima NAFISI (KCL)
Jordi PEREZ-ROMERO (UPC)
João REBELO (PTIN)
Oriol SALLEN (Editor, UPC)
Juan SÁNCHEZ GONZÁLEZ (UPC)
Alessandro TROGOLO (TILAB)
Anna UMBERT (UPC)
Avelina VEGA NOVELLA (TID)
Lin WANG (KCL)
Giovanna ZARBA (TI)

EXECUTIVE SUMMARY

The scope of this document is to present a wide range of studies in the field of RRM/CRRM strategies that have been developed in the framework of the IST-project EVEREST. RRM/CRRM solutions envisage an optimised utilisation of scarcely available radio resources for the support of mixed services within heterogeneous networks.

Initially, RRM aspects for different RATs are covered, since these are the pillars for the efficient development of common strategies. In particular, UTRAN, GERAN and WLAN have been covered, UTRAN being more extensively treated because of the larger number of dimensions involved (i.e. frequency, time, code and also power) in the radio resource management problem. Finally, CRRM mechanisms have been targeted and strengthened in order to provide consolidated views and solutions.

As a final report of EVEREST's activities on RRM and CRRM for the whole duration of WP3, this document intends to capture the different developments achieved from March 2004 to October 2005. For those aspects already included in previous deliverables (D11 "First report on the evaluation of RRM/CRRM algorithms", issued October 2004, and D15 "Report on the evaluation of RRM/CRRM algorithms", issued March 2005) only some relevant concepts and a few representative results are captured here, so as to maintain the self-contained nature in the present final report. Strategies and algorithms developed since March 2005 are described with a higher level of detail, so that these activities are suitably reported in an official deliverable. Thus, the extension devoted to every single topic covered in the following subsections is not necessarily indicative of its relevance for the whole EVEREST project.

The document is structured as follows. After an introduction in Section 1, Section 2 is devoted to RRM mechanisms and considerations for UMTS. Section 3 covers RRM for GERAN, while Section 4 is devoted to WLAN. Then, Section 5 is devoted to CRRM presentation and evaluation of the different proposed alternatives. Finally, Section 6 summarises the conclusions reached.

Table of Contents

EXECUTIVE SUMMARY.....	V
1 RRM IN A BEYOND 3G FRAMEWORK.....	1
2 RRM ISSUES FOR UMTS	2
2.1 INTRODUCTION.....	2
2.2 A NEW FRAMEWORK FOR CAPTURING COUPLING AMONG CELLS.....	2
2.2.1 Introduction.....	2
2.2.2 Applicability example: Congestion control.....	3
2.3 INTEGRATED VOICE/DATA IN CDMA SYSTEMS FRAMEWORK	5
2.4 INDOOR TRAFFIC.....	8
2.4.1 Introduction.....	8
2.4.2 Results.....	8
2.5 TRAFFIC HOT-SPOTS	9
2.5.1 Introduction.....	9
2.5.2 Pilot Adjustment Algorithm: Downlink case.....	9
2.6 STATIC TRAFFIC	9
2.6.1 Results.....	10
2.6.2 Impact of mobility	11
2.7 REPEATERS	12
2.7.1 Repeaters usage in WCDMA systems.....	12
2.7.2 Analysed layout	12
2.7.3 Simulation results.....	12
2.8 MULTIPLE RF CARRIERS	14
2.8.1 Admission control, congestion control and coverage control.....	15
2.8.2 QoS measures.....	16
2.8.3 Assumptions and reflections.....	16
2.8.4 Methods.....	16
2.8.5 Simulation study.....	17
2.8.6 Conclusions.....	21
2.9 HIERARCHICAL CELL STRUCTURES	22
2.9.1 Cell selection/reselection criteria in HCS.....	22
2.9.1.1 Generalities.....	22
2.9.1.2 The mobility class evaluation	23
2.9.1.3 The cell-reselection algorithm	24
2.9.1.4 Some guide-lines for setting the cell-reselection parameters.....	26
2.9.1.5 Simulated layout.....	27
2.9.2 Capacity enhancement with HCS.....	32
2.9.2.1 Analysed layout.....	32
2.9.2.2 Simulation results	33
2.9.3 RRM and mobility issues in HCS idle mode.....	34
2.9.3.1 Introduction	34
2.9.3.2 Cell reselection criteria with HCS	34
2.9.3.3 Capacity analysis.....	39
2.9.3.4 Conclusions	42
2.9.4 Application of the derivatives framework to HCS.....	42
2.9.4.1 Frequency allocation schemes	43
2.9.5 Non-Real Time Packet Transmission for a Microcell (Hotspot) Embedded in CDMA Macrocell Systems 44	
2.9.5.1 Results and dicussion	45
2.10 TRANSPORT CHANNEL TYPE SWITCHING.....	47
2.10.1 Introduction.....	47
2.10.2 Bidirectional traffic model for the WWW service.....	47
2.10.3 Traffic measurements and TCTS algorithm	50
2.10.4 Main simulation inputs.....	53
2.10.5 Simulation results.....	55
2.10.5.1 DCH-only versus Transport Channel Type Switching comparison.....	55

2.10.5.2	TCTS active	62
2.10.6	Conclusions	68
2.11	ADMISSION CONTROL	69
2.11.1	Scenario description	69
2.11.2	Services, Users and Services&Users prioritization	70
2.11.3	Services prioritization study in an indoor scenario	70
2.11.3.1	Scenario description	71
2.11.3.2	Admission control strategies	73
2.11.3.3	Conclusions	77
2.12	DIFFSERV AWARE SCHEDULING	77
2.12.1	DiffServ Marking	77
2.12.2	Motivation of Color Aware RRM	78
2.12.3	Applicability example: Color aware link adaptation protocol	81
2.12.3.1	Proposed scheme	81
2.12.4	DiffServ aware RRM for CDMA	83
2.12.5	Applicability example: Color aware coverage control	86
2.13	HIGH SPEED DOWNLINK PACKET ACCESS	87
2.13.1	Scheduling methods for TCP traffic over HSDPA	88
2.13.2	Reference scheduling evaluations	89
2.13.2.1	Hybrid Automated Request	89
2.13.2.2	Scheduling	89
2.13.2.3	Results	90
2.13.3	Advanced scheduling evaluations	92
2.13.3.1	Introduction	92
2.13.3.2	Simulation assumptions	92
2.13.3.3	Performance metrics	97
2.13.3.4	Simulation results	98
2.13.3.5	Summary	102
2.13.4	Performance Enhancement for HSDPA	103
2.13.4.1	Biased Adaptive Modulation/Coding to Provide VoIP QoS over HSDPA	103
2.13.4.2	DiffServ-aware priority queuing improves IP QoS support on HSDPA	113
2.13.4.3	Supporting Heterogeneous Traffic in HSDPA	117
2.13.4.4	Code Multiplexing of Multiple Access Users in HSDPA	125
2.14	RAN SHARING	127
2.14.1	Simulation study	128
2.14.2	Multicell RAN Sharing	129
2.14.2.1	Introduction	129
2.14.2.2	Proposed sharing algorithms	130
2.14.2.3	Scenario Model	132
2.14.2.4	Results	133
2.15	LOCATION AWARE RESOURCE RESERVATION	134
2.15.1	Introduction	134
2.15.2	Resource Reservation Algorithm	136
2.15.3	Simulation model	138
2.15.4	Results	138
2.15.5	Conclusions	142
3	RRM ISSUES FOR GERAN	142
3.1	INTRODUCTION	142
3.2	ADMISSION CONTROL	143
3.3	VIDEO STREAMING OVER GPRS	145
4	RRM ISSUES FOR WLAN	148
4.1	INTRODUCTION	148
4.2	ADMISSION CONTROL FOR IEEE 802.11A/B/G	149
4.2.1	Analytical model to estimate the performance of MAC 802.11 DCF for real-time services (step 1) 149	
4.2.2	Capacity region for the WLAN hot-spot (step 2)	150
4.2.3	Capacity region based Admission Control (step 3)	151
4.2.4	Validation of the analytical model for the performance evaluation of IEEE 802.11a/b/g WLAN	152
4.2.4.1	Introduction	152
4.2.4.2	Previous work	152
4.2.4.3	Overview of the Simulation work	153

4.2.4.4	Simulated scenarios	155
4.2.4.5	Results	156
4.2.4.6	Conclusions	164
4.3	ADMISSION CONTROL FOR IEEE 802.11E CONTENTION ACCESS.....	164
4.3.1	Enhanced Distributed Admission Control Algorithm.....	166
4.3.2	EDAC performance evaluation	170
4.3.3	Conclusions.....	174
4.4	SERVICE PRIORITY – QoS ENHANCEMENTS IN 802.11B.....	174
4.4.1	Hierarchical Token Bucket.....	174
4.4.2	Comparison of the legacy DCF and DFS, DRR service differentiation schemes.....	175
4.5	SERVICE PRIORITY - QoS ENHANCEMENTS IN 802.11E.....	177
4.5.1	Enhanced distributed channel access (EDCA).....	178
4.5.2	Prioritisation in EDCA.....	180
4.5.3	Simulation Results.....	180
4.5.4	On the use of the EDCA Transmission Opportunity (TXOP) mechanism for improving the WLAN system performances.....	182
4.5.4.1	Influence of stations working at lower transmission rates on system performance	182
4.5.4.2	Implementation of TXOP mechanism for system performance improvement.....	185
4.5.4.3	Optimum TXOP limit for each type of traffic	187
4.5.4.4	Dynamic TXOP limit configuration	190
4.5.4.5	Packets fragmentation to enhanced QoS guarantees for high priority traffics.....	193
4.5.4.6	Conclusions	194
5	COMMON RRM.....	195
5.1	INTRODUCTION: THE CRRM FRAMEWORK.....	195
5.1.1	CRRM functional model.....	196
5.1.2	CRRM functionalities.....	197
5.1.3	CRRM implementation.....	198
5.1.4	Scope of this chapter.....	199
5.2	RRM POLICIES IN HETEROGENEOUS NETWORKS.....	199
5.3	SERVICE-BASED RAT SELECTION POLICIES	201
5.3.1	Introduction.....	201
5.3.2	Performance of basic policies	201
5.3.2.1	Simulation Environment.....	202
5.3.3	Radio network considerations.....	206
5.3.3.1	Combination of basic policies: n-complex policies	207
5.3.4	Vertical Handover.....	210
5.3.4.1	Loose and Tight Interworking between Vertical and Horizontal Handover	211
5.3.5	Conclusions.....	223
5.4	LOAD BALANCING - BASED RAT SELECTION	224
5.4.1	Introduction.....	224
5.4.2	Initial RAT selection.....	225
5.4.2.1	Performance evaluation	227
5.4.3	Scenario including vertical handover	230
5.4.3.1	Load and service-class distribution.....	230
5.4.3.2	Vertical Handover Rates.....	232
5.4.3.3	Performance evaluation of voice users	233
5.4.3.4	Performance evaluation of interactive users	235
5.4.3.5	Throughput performance	235
5.4.3.6	Admission probability	236
5.4.4	Conclusions.....	236
5.5	PATH LOSS - BASED RAT SELECTION.....	237
5.5.1	Introduction.....	237
5.5.2	Preliminary theoretical evaluation	239
5.5.3	Evaluation in a dynamic scenario.....	244
5.5.3.1	Initial RAT selection and vertical handover algorithms	244
5.5.3.2	Evaluation in a single service scenario	245
5.5.3.3	Multi-service scenario	254
5.5.4	Conclusions.....	263
5.6	RAT PRIORITY LIST-BASED RAT SELECTION.....	264
5.6.1	Introduction.....	264
5.6.2	Initial RAT Selection Algorithm.....	264
5.6.2.1	Considered initial RAT Selection Strategies	267

5.6.2.2	Conclusions	270
5.6.3	<i>Study on sharing load and service prioritization</i>	270
5.6.3.1	Load sharing strategies: GERAN only can provide voice.	270
5.6.3.2	Admission Control based on Service Prioritization: GERAN only can provide voice.	271
5.6.3.3	Admission Control Based on Service Prioritization: GERAN can provide both, voice and data services	271
5.7	PERCEIVED TCP THROUGHPUT IN CRRM FRAMEWORK	275
5.7.1	<i>Simulation assumptions</i>	275
5.7.2	<i>Performance results</i>	278
5.8	IMPACT OF MULTI-MODE TERMINALS ON CRRM PERFORMANCE.....	282
5.8.1	<i>Introduction</i>	282
5.8.2	<i>Simulation results</i>	284
5.8.2.1	Throughput Performance	284
5.8.2.2	Delay Performance.....	285
5.8.2.3	Approach using EGPRS dedicated slots	286
5.8.3	<i>Conclusions</i>	288
6	CONCLUSIONS	288
	APPENDIX A	299
	A.1 SIMULATION SETUP	299
	A.2 SIR CALCULATIONS	300
	A.3 CQI ESTIMATION	301
	APPENDIX B	302
7	REFERENCES	304
8	ABBREVIATIONS	313

1 RRM IN A BEYOND 3G FRAMEWORK

The provision of heterogeneous network topologies in beyond 3G systems is conceptually a very attractive notion; however, it is certainly a challenge to the network designer. Here, coupling between the networks of possibly different characteristics can be provided, leading to open, loose, tight and very tight coupling. The stronger the coupling the better resources are being utilized leading to an optimum of performance. However, this comes along with an increased effort in the definition and implementation of required interfaces. A suitable trade-off for specific systems thus ought to be determined.

In either case, available radio resources of coupled networks will have to be managed jointly, up to the degree allowed by the coupling mechanism. Targeted is an optimum solution in terms of throughput, cost per packet, development and deployment cost, etc. Radio resource management (RRM) strategies are responsible for an utmost efficient utilisation of the air interface resources in the Radio Access Network (RAN). Any stand-alone wireless systems or heterogeneous hybrids thereof, rely on RRM strategies to guarantee a certain prior agreed QoS, to maintain the planned coverage area, to offer high capacity, etc. Without them, the most efficient physical transmission system coupled into the most sophisticated IP core network would fail. For a realistic network deployment it is utmost important to devise optimum but tangible strategies for managing available resources of the RANs attached to a given CN.

The QoS concept can be understood in many different forms and levels. Although QoS is inherently subjective, the realization of QoS in a communication network needs to be associated with some QoS quantitative parameters that characterize a satisfactory user's perceived service.

QoS provisioning in mobile systems is particularly complex given the large number of effects that tend to degrade the communication links. Fading in propagation conditions, interference, channel distortion, etc. are only some examples. Although the ultimate objective of an end-to-end QoS will be retained, the focus will be placed here on QoS in the radio segment.

3G and Beyond systems will offer an optimization of the capacity over the air interface by means of designing efficient algorithms for Radio Resource and QoS Management. These functionalities are very important in the framework of 3G systems because the system relies on them to guarantee a certain target QoS, to maintain the planned coverage area and to offer a high capacity, which somehow are contradictory to each other (e.g. capacity may be increased at the expense of a coverage reduction; capacity may be increased at the expense of a QoS reduction, etc.).

The management of radio resources can be seen as a problem with multiple dimensions. Every RAT is based on specific multiple access mechanism exploiting in turn different orthogonal dimensions, such as frequency, time and code. Then, RRM mechanisms are needed for every considered RAT: GERAN, UTRAN and WLAN in the case of EVEREST. CRRM is based on the picture of a pool of radio resources, belonging to different RATs but commonly managed. Then, the additional dimensions introduced by the multiplicity of RATs available provide further flexibility in the way radio resources can be managed and, consequently, overall improvements may follow.

For a general description of the Radio Resource Management (RRM) and Common RRM (CRRM) problems in the framework of EVEREST the reader is referred to [1].

2 RRM ISSUES FOR UMTS

2.1 INTRODUCTION

WCDMA access networks, such as the considered in UTRA-FDD proposal [2], provide an inherent flexibility to handle the provision of future 3G mobile multimedia services. UMTS will offer an optimization of capacity in the air interface by means of efficient Radio Resource Management (RRM) algorithms. 3GPP has provided a high degree of flexibility to carry out the RRM functions, which at the same time need to be consistent for both uplink and downlink.

Flexibility is a good and necessary property from the system perspective indeed, to cope for example with future services uncertainties. At the same time, it will be difficult to cope with flexibility from the radio network perspective, since flexibility requires proper management mechanisms. Clearly, the benefits deriving from flexibility justify the research and engineering efforts for smart radio resource management algorithms development. By doing so, the potentials of technological advances (e.g. UMTS) can be fully exploited.

RRM is a complex problem with many factors influencing in the achieved performance and with many mixing effects. Furthermore, the RRM problem has multiple dimensions and multiple functionalities that, either in a more direct or indirect way, impact on the air interface. Then, a crucial aspect is to identify relevant issues and fundamental elements influencing on the overall RRM process, then achieving a wide-scope and open-minded perspective.

In this context, the subsections detailed below describe a variety of studies related to RRM for UMTS. Studies include analytical frameworks (Section 2.2 and 2.3), traffic and deployment characteristics influencing the RRM context (Section 2.4 to Section 2.9) as well as some more specific aspects dealing with particular RRM strategies (Section 2.10 to Section 2.13). Finally, the influence of RRM in RAN sharing is analysed in Section 2.14 and the applicability of location-aware RRM mechanisms has received some attention in Section 2.15.

2.2 A NEW FRAMEWORK FOR CAPTURING COUPLING AMONG CELLS

2.2.1 Introduction

An innovative mathematical framework capturing the air interface coupling among the different cells in the scenario has been developed based on the derivatives of the cell uplink load factor and the downlink transmitted power [1][17]. This framework is presented in a compact formulation for both uplink and downlink, which is claimed to be novel for both directions to the author's best knowledge. The multiple issues impacting the radio interface behavior and the much higher degree of coupling among them deriving from the WCDMA nature, where users transmit at the same time and on the same carrier, will be shown on a more visible form than more classical formulations.

On the other hand, one of the main traffic characteristics in cellular networks is the non-homogenous spatial distribution. Although, network planning can consider this fact, the high dynamics associated to traffic clearly need additional mechanisms to cope with the potential problems on the network performance for traffic profile distributions significantly different from those expected in the network planning phase.

In this context, mechanisms supporting smart load control actions could be of great interest and could be applied at different levels. The first issue to include in a smart load control would be the ability to detect those cells mostly affecting the reference cell. Thus, the second main contribution of the developed formulation is the exploitation of the developed analytical

model by obtaining the derivative of the reference cell uplink load (alternatively the transmitted power in downlink) with respect to any neighboring cell. These derivatives allow to identify the most critical cells and users influencing a reference cell, then the definition of smart RRM algorithms (e.g. admission control, congestion control, packet scheduling) may follow.

Details on the analytical development of the derivatives framework are provided in [1] (Section 2.2). Further refinements on the analytical formulation were extended in [17] (Section 2.1).

2.2.2 Applicability example: Congestion control

Load control mechanisms should be devised to face situations in which the system has reached a congestion status and therefore the QoS guarantees are at risk due to the evolution of system dynamics (mobility aspects, increase in interference, etc.). Congestion occurs when the admitted users can not be satisfied with the normal services agreed for a given percentage of time because of an overload.

Usually the network is planned to operate below a certain maximum load factor (alternatively in the downlink, a certain fraction of the maximum power available at the cell site). Then, the congestion control procedure will be triggered when the load factor increases over a certain threshold during a certain amount of time. Afterwards, actions that must be taken in order to maintain the network stability by reducing the cell load.

The algorithm in the uplink direction would operate in the following steps, and the objective is to reach a load factor lower or equal than η_T in cell 0.

Step 1.- Select the cell k with maximum: $\left(\eta_{k,NRT} \cdot \frac{\partial \eta_0}{\partial \eta_k} \right)$

where $\eta_{k,NRT}$ is the amount of load factor devoted to NRT traffic in cell k.

Step 2.- The reduction to be achieved in the selected cell k is given by

$$\Delta \eta_k = \frac{\Delta \eta_0}{\left(\frac{\partial \eta_0}{\partial \eta_k} \right)} \quad \text{where } \Delta \eta_0 = \eta_0 - \eta_T \text{ is the desired reduction in cell 0.}$$

Step 3.- Order the NRT users in cell k in a table in decreasing order of the factor $I_{i_k,0}^{UL}$, where:

$$S_{j,0}^{UL} = \sum_{i_j=1}^{n_j} \frac{L_{i_j,j}}{L_{i_j,0}} \frac{1}{\left(\frac{E_b}{N_o} \right)_{i_j} R_{i_j} + 1} = \sum_{i_j=1}^{n_j} I_{i_j,0}^{UL} \cdot \quad (1)$$

Step 4.- Inhibit the transmissions of the users in the table until reaching the desired reduction of $\Delta \eta_k$ or until having inhibited all the users.

Step 5.- Measure η_0 and if it is still higher than η_T return to step 1.

Similarly, an algorithm for the downlink direction is presented in [1] (Section 2.2.5.2).

In order to show the suitability of the algorithms envisaged, some reference algorithms are considered for comparison purposes. In particular, two other algorithms that operate in the same steps are denoted as:

- Aleat: In step 1 the base station to be reduced is chosen randomly between the six neighbouring cells. In step 3 the NRT users of the selected cell are not ordered.
- Interact: In step 1 the base station which power is to be reduced is chosen as the base station with the highest NRT load/power. In step 3 the NRT users of the selected cell are not ordered.

In this case, the cell radius is $R=1000m$. Traffic is uniformly distributed in all cells and also within each of these cells (i.e. scenario is homogeneous both at intercell and intracell level). 10 RT users are allocated in the central cell. A variable number of NRT users (50 to 300) are distributed in the rest of the cells. The RAB considered is 64/384 Kbps for both RT and NRT.

The objective in the load control algorithms is to keep 0.8 in uplink cell load factor and 0.8 for the DL fraction of power in the reference cell.

It can be observed in Figure 1 and Figure 2 that the algorithm based on derivatives achieves the load reduction with the lowest NRT throughput reduction. The difference with respect to the “aleat” algorithm is very high (notice that “aleat” inhibits all the NRT users in the neighbouring cells and its reduction is around 80%). On the contrary, both algorithms based on derivatives and on interactive load perform much better.

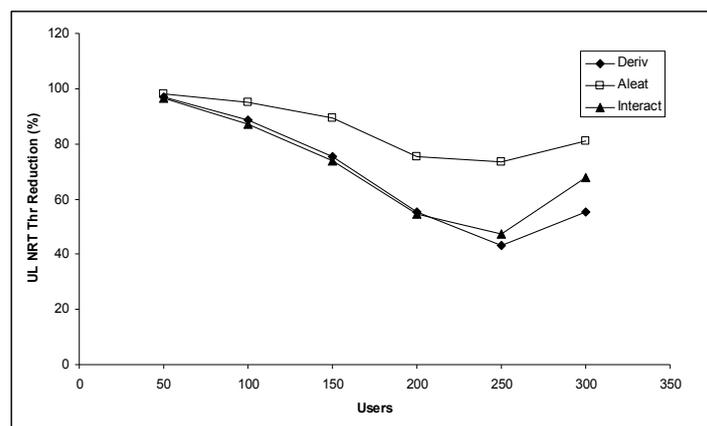


Figure 1 UL NRT Throughput reduction in the neighbouring cells

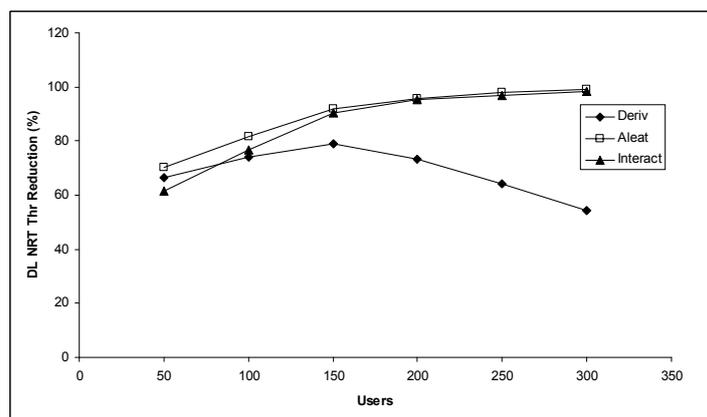


Figure 2 DL NRT Throughput reduction in the neighbouring cells

2.3 INTEGRATED VOICE/DATA IN CDMA SYSTEMS FRAMEWORK

In this section we present a new model to analyze the performance of an integrated voice/data CDMA system based on the interference-analysis method proposed in [3][9]. Since most packet services use the DFT (Defer First Transmission) transmission mode associated with ARQ (Auto-Repeat-reQuest) schemes, in congestion control state, congestion control function usually delays the PS data service to guarantee CS service quality. Thus this research focuses on the impacts of introduction of NRT data service with congestion control. In CDMA cellular systems, due to their interference-limited characteristic, the congestion control in each cell is usually designed to maintain the interference level below the maximum allowed interference level (MAIL) at BS (Base Station) for the uplink by control number of data users' transmission. To achieve this with the universal frequency reused CDMA cellular system, the congestion control is suppose to be a function of both inter-cell and intra-cell interference. With this function, a new user's transmission in one cell contributes to the inter-cell interference in other cells, and so the congestion control in the other cells will act to control its own data users' transmission; this will then affect the interference in the original cell, and so on. Thus the inter-cell interference and intra-cell interference interact with each other through the congestion function.

In [1] section 2.3.3, three types of services LCD, voice and data are considered in this integrated CDMA cellular system. So a three-dimensional Markov process can be used to describe the system with three state variables $k_{LCD,t}$, $k_{vo,t}$, and $k_{d,t}$. They are respectively the number of active LCD users, the number of active voice users and the number of queuing data users. With the procedure presented in [1] section 2.3.3, the three-dimensional Markov process can be solved by three one-dimensional queues: a LCD queue with its steady state probability denoted by $\pi_{k_{LCD}}^{LCD}$, and a voice queue with steady state probability $\pi_{k_{vo}}^v$, which are easily to be derived based on section 2.3 description, and a conditional PS data queue with its steady state probability $\pi_{k_d}^d(k_{LCD}, k_{vo}, I_{inter})$.

Then the expressions for calculating performance measures such as outage probability, average data delay and data blocking probability can be derived. Let $G(w)$ be the density function of inter-cell interference, which is still assumed to be a Gaussian distribution as in [10]. Then the outage probability is given as

$$P_o = \Pr(I_{inter} + \rho + \eta > I_{max})$$

$$= \int_{I_{max}-\eta}^{\infty} G(w)dw + \int_0^{I_{max}-\eta} \int_{I_{max}-\eta}^{\infty} f_{\rho}(x)G(w)dw dx \quad (2)$$

where ρ is the random variable representing the intra-cell interference in terms of total received power at the BS, and is given by $\rho \equiv k_{LCD}S_{LCD} + k_{vo}S_{vo} + K_2S_d$. Moreover $f_{\rho}(x)$ is the density function of ρ , which is basically the combined distribution of the signal powers of each class of user and traffic load in terms of the number of transmitting users given by the above Markov chain model.

Thus the *throughput* is given as

$$S \equiv \int_0^{\infty} \sum_{k_{LCD}=0}^{N_{LCD}} \sum_{k_{vo}=0}^{N_{vo}} \sum_{k_d=0}^{BU} \pi_{k_{vo}}^{vo} \pi_{k_{LCD}}^{LCD} \pi_{k_d}^d(k_{LCD}, k_{vo}, w) K_2(k_{LCD}, k_{vo}, k_d, w)(1 - PER)G(w)dw \quad (3)$$

Then the average message delay, based on its definition, is given as

$$D = \frac{\int_0^{\infty} \sum_{k_{LCD}=0}^{N_{LCD}} \sum_{k_{vo}=0}^{N_{vo}} \sum_{k_d=0}^{BU} \pi_{k_{vo}}^{vo} \pi_{k_{LCD}}^{LCD} \pi_{k_d}^d (k_{vo}, k_{LCD}, w) \cdot k_d \cdot L \cdot G(w) dw}{S} \quad (4)$$

where the integral above the dominator represents the average traffic load in terms of the total number of packets in the buffer.

And the blocking probability is given as

$$P_b = \sum_{d=1}^{\infty} \int_0^{\infty} \sum_{k_{LCD}=0}^{N_{LCD}} \sum_{k_{vo}=0}^{N_{vo}} \sum_{k_d=0}^{BU} A(d) \pi_{k_{LCD}}^{LCD} \pi_{k_{vo}}^{vo} \pi_{k_d}^d (k_{vo}, k_{LCD}, w) \left(\frac{\max(0, d + k_d - BU)}{d} \right) G(w) dw \quad (5)$$

From (3) to (5), it is found that, if inter-cell interference is determined, the steady state probability of the systems will be obtained, and then the performance measures can be computed from (3) to (5).

In [1] section 2.3.3, we develop two analytical models which is able to derive the inter-cell interference through an recursive process for perfect power control and imperfect power control. With these models, we exams the interactions between the PS data and CS traffic under the congetion control function. Study of the congestion control function suggests that the control threshold T , which determines the trade-off between CS quality and PS performance, needs to be carefully dealt with in the multi-cell environment. As shown in Figure 3, in genenenral, with a larger T , data traffic is able to take more resource and thus a higher throughput is achieved for data users and fewer voice users can be supported in the system, and vice versa. The investigation is also extended to the system performance with power control errors. Figure 4 shows effects of PS data on the voice capacity with various power control errors. In this figure, with a fixed power control error, the curve follows very similar trend as those without power control error (as shown in Figure 3), i.e. the capacity loss is getting less as data traffic load increasing. As power control error increases, the voice capacity loss due to the introduction of data increases and this increase is getting larger and larger. So with larger power control error, a smaller T values would favour voice capacity. The numerical results in Table 1-2 on system capacity performance show the basic trend of capacity reduction caused by power control error. In this study, it is found that, compared with the single cell case, capacity loss in multi-cell situation decreases as the power control error increases, because as the power control error increases, the outage probability caused by PCE become more and more dominant. Performance of PS in the integrated PS/CS environment is also studied with power control errors. The results in Figure 5 show that, when the power control error becomes larger, the degradation in PS performance becomes worse. Therefore a close and fast power control scheme would be very useful for packet transmission as well.

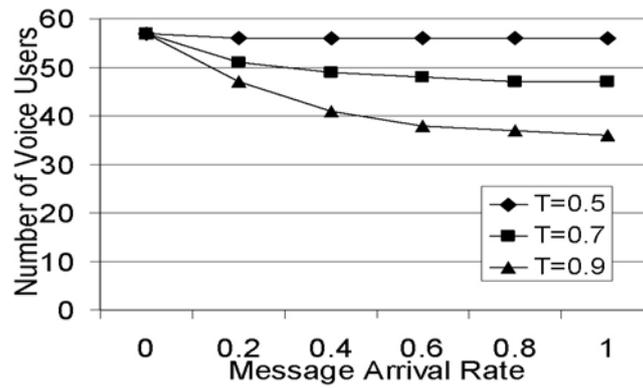


Figure 3 Effects of PS data on voice capacity with various T values

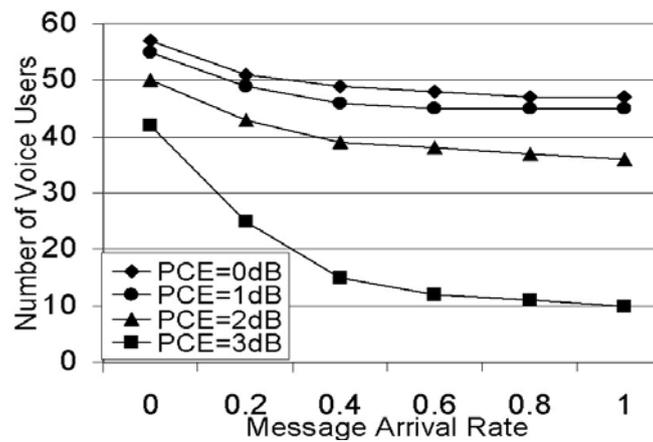


Figure 4 Effects of PS data on Voice Capacity with Various Power Control Errors

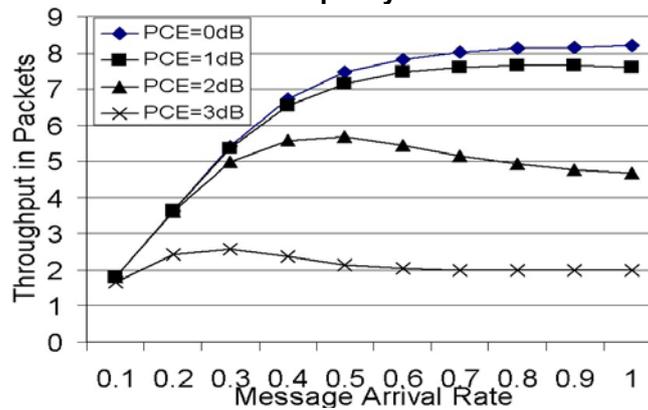


Figure 5 Effect of Power Control Error on PS Throughput (1LCD, 20Voice)

Table 1 Single-cell Voice Capacity (I_{max} is normalized to background noise η)

$I_{max}(dB)$ \ PCE (dB)	6	10	15
0	92	113	122
1	88	109	118
2	77	96	104
3	60	75	83

Table 2 Multi-Cell Voice Capacity

$I_{max}(dB)$	6	10	15

PCE(dB)			
0	57(38%)	76(32.7%)	87(28.7%)
1	55(37.5%)	74(32.1%)	85(28%)
2	50(35%)	67(30%)	76(26.9%)
3	42(30%)	56(25%)	64(22.9%)

2.4 INDOOR TRAFFIC

2.4.1 Introduction

Indoor traffic is very important in 2G networks, as remarkable traffic load is originated and terminated inside buildings. Nevertheless, the implications of indoor traffic in 3G W-CDMA based systems may significantly differ from 2G TDMA-based solutions because transmitted power levels are the key radio resources in W-CDMA. The higher power levels needed for indoor service will lower cell capacity both for uplink and downlink traffic. Consequently, it can be important for a network operator to quantify the impact that indoor traffic may have on the overall system efficiency in order to devise suitable deployment guidelines (i.e. how fast the transition from outdoor macrocell sites to indoor micro and picocells distributions should be carried out). Besides, taking into account different radio bearer services (e.g. 64 kb/s uplink with 64 kb/s downlink; 64 kb/s uplink with 384 kb/s downlink, etc.) different levels of capacity degradation as well as different link direction constraints may appear.

In this framework, this section is intended to devise the impact on system capacity deriving from different percentages of indoor traffic in the scenario. Firstly, by means of a simple analytical model it will be shown that uplink direction reveals to be much more sensitive to indoor traffic than downlink, so that higher degradations are expected in the uplink. Secondly, different radio bearers with different asymmetry levels are studied by means of system level simulations in both uplink and downlink directions. The complete set of obtained results may help to provide the indications on how and when new infrastructure and/or new cell hierarchies need to be deployed in a given scenario as user density and demanded services evolve.

A formula comparing the power increase in uplink and downlink is derived in [1] (Section 2.4). From this, it can be observed that in the downlink the power increase depends mainly on noise power while in the uplink it depends mainly on noise plus system interference. As a result, the power increase will be higher in the uplink direction and the higher the load in the system the higher the difference with respect to the downlink will be. Consequently, a lower degradation caused by indoor traffic is expected in the downlink when compared to the uplink. The rationale behind this effect is the higher protection against interference for downlink indoor users provided by the in-building penetration loss.

2.4.2 Results

In order to attain the capacity degradation caused by indoor traffic in both uplink and downlink a set of system level simulations have been made.

Besides, Table 3 presents the capacity loss for both uplink and downlink as the fraction of indoor traffic increases. Capacity is defined as the maximum number of users in the scenario that guarantees a BLER $\leq 2\%$. It can be observed that the degradation due to indoor users is much more significant in the uplink than in the downlink direction. Particularly, when half of the users are indoor ($p=0.5$), the reduction in the uplink is 88% while in the downlink it is only 15%.

When an asymmetric service like the RAB 64/384 kb/s is considered, results in terms of capacity degradation with respect to the $p=0$ case are also shown in Table 3. Although the capacity reduction for the DL is higher with the 384 kb/s service than with the 64 kb/s service, it is still much lower than the reduction in the uplink.

Table 3 Capacity loss (%) relative to the case with no indoor traffic ($p=0$) for 64/64 kb/s and 64/384 kb/s radio bearer

	UL 64 Kb/s	DL 64 Kb/s	DL 384 Kb/s
P=0.1	19.2%	9.3%	12.5%
P=0.2	37.7%	11.6%	18.8%
P=0.5	88.4%	15.3%	25.0%

2.5 TRAFFIC HOT-SPOTS

2.5.1 Introduction

In a real mobile network, there are certain geographical areas with high density of users. In these areas, a high demand of radio resources may appear. In order to assure the user QoS (Quality of Service) requirements in a hotspot, it is necessary a proper radio network planning (e.g. by a proper definition of the different base station locations, transmission powers, pilot channel power, etc.). However, hotspots characteristics (such as geographical location, etc.) are not always known a priori, so an unexpected increase in the demand of resources by a hotspot can have a relevant impact on network performance. Therefore, it is prime important a proper evaluation of the effect of these hotspot peculiarities on the network behaviour. Moreover, the existence of dynamic hotspots (i.e. a group of mobile users that move following a certain mobility pattern, such as the way out of a railway station, the exit of a football match, etc) and its impact on system performance is as well an important issue, especially on the forward link, where the user location distribution affects directly on the base station power allocation [10] (i.e. many users far from the Node-B may demand high levels of power causing that the base station has not enough power to satisfy all users demands).

These dynamic changes in the network may cause overload situations where the user QoS requirements can not be guaranteed. These overload situations can be prevented by RRM (Radio Resource Management) mechanisms (e.g. admission control or congestion control algorithms) which determine how the radio interface is used and shared among the users. Another technique that is commonly used is to adjust the transmission pilot power of the hotspot cells and its adjacent cells [13][15].

2.5.2 Pilot Adjustment Algorithm: Downlink case.

In order to manage dynamic traffic hotspots not only network planning but also RRM algorithms must be considered. Non-uniform user distributions make that certain cells may be overloaded while the load of other cells is quite low. The objective of the proposed pilot adjustment algorithm is to reduce these differences in the load of the different cells by shedding traffic from overloaded cells to low loaded cells. By doing this, a more uniform traffic distribution will be obtained and then, these overload situations will be reduced.

The load balancing technique reduces the power limitation probability of the different base stations, and this in turn reduces the dropping probability as it was shown in [17] (Section 2.2.1.4).

2.6 STATIC TRAFFIC

The flexibility in the provision of multiple bit rate services in 3G communication systems will

allow users to benefit from services better adjusted to their specific requirements. In addition to higher bit rates, the support of high QoS (Quality of Service) will also be crucial for 3G success from the user point of view. It should be provided by means of a proper utilization of the air interface resources, which at the same time should assure the planned coverage area and offer a high system capacity to maximize operators' revenue.

In this context, it becomes prime important to identify the key elements characterising the different services and anticipate the required mechanisms to support these services through the air interface in a suitable and optimised manner. In particular, users that receive data traffic services, typically with laptops in scenarios like offices, airports, etc., use to be static or, at least, with a very limited mobility. This fact gives room to propose more sophisticated RRM strategies, which may provide significant performance improvements.

The exploitation of the usual static nature of data traffic in W-CDMA for admission control purposes has not been addressed so far in the open literature. In this case, the adoption of an advanced admission control policy that takes path loss reports into account results in a significant improvement of the system performance.

Under this framework, the proposed algorithm, denoted as PLEBAC (Path Loss Estimation Based Admission Control) benefits from the easier predictability in terms of power consumption of static data users. It makes use of the measurement reports provided by the terminal during the call set-up process in order to have a more accurate estimation of the required power along the connection time. The proposed algorithm is evaluated under different conditions of service bit rate and cell radii and compared against a reference algorithm and compared against an algorithm that simply considers an average power consumption, denoted as PABAC (Power Averaged Based Admission Control).

2.6.1 Results

In order to assess the potential of the proposed PLEBAC algorithm, a set of system level simulations have been carried out. Different scenarios have been considered with different cell radii, ranging from 500 m up to 2 Km.

When extending the analysis to other scenarios, Figure 6 illustrates the throughput gain of the PLEBAC admission control with respect to PABAC for different cell radii, when the offered bit rate is 256 kb/s and 384 kb/s. Notice that the gain is bigger for larger cell radii and higher service bit rates. In these cases, the higher power demand of users with larger path loss leads to performance degradation not only for these users but also for other users that are located closer to the base station. As a result, an algorithm like PLEBAC, which takes into account user's path loss, improves the overall performance while at the same time it guarantees the quality of the accepted users. Furthermore, the higher the service bit rate, the higher the contribution of a user to the total system throughput, and consequently, a bad admission or a bad rejection turns into larger throughput reductions. As a result, it can be concluded that the PLEBAC strategy is better adapted to user's distribution in the network.

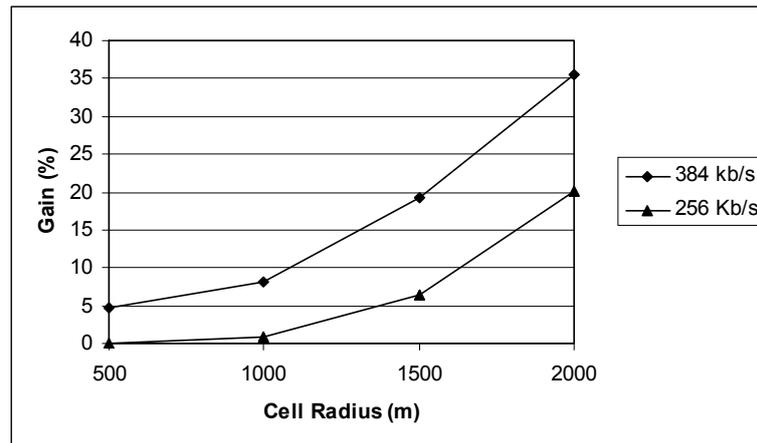


Figure 6 Throughput gain as a function of the cell radius

The proposed algorithm has been evaluated in different scenarios with different cell radius, service bit rate and offered load. Results show that PLEBAC outperforms PABAC in all the cases thanks to a lower error between the power estimation in the admission control phase and the actual transmitted power along the connection lifetime. The throughput gain is very significant especially in those scenarios with large cell radius and high bit rates, where higher power consumption is required and therefore bad admissions and bad rejections in the admission control phase may lead to important degradations for all the accepted users.

2.6.2 Impact of mobility

According to results shown in previous section, an extension for hybrid static/dynamics scenarios is developed here. In particular, it will be studied a Combined PLEBAC/PABAC Based Admission Control (CPBAC) which uses either PLEBAC or PABAC estimation depending on the requesting user speed. The obtained results will show that this combined admission control provides better system performance than PLEBAC or PABAC alone.

The proposed CPBAC algorithm consists on a combined admission control algorithm that takes advantage of user speed information in order to obtain a more accurate estimation of the power increase estimation ΔP_T . Then, depending on the requesting user speed the CPBAC algorithm will use either PLEBAC or PABAC estimation.

The obtained results show that for low speed users, PLEBAC algorithm is more adequate while for high speed users PABAC provides better performance. In Table 4, the obtained throughput for PLEBAC, PABAC and this new proposed CPBAC algorithm is shown. It has been considered 50% of indoor users at 3km/h and 50% of outdoor users at 50km/h, in both cases the bit rate is 384kbps. If the call duration is 30seconds, PLEBAC performs better than PABAC. If the call duration is 3minutes, PABAC performs better than PLEBAC. However, in both cases, CPBAC provides higher throughput, as shown in Table 4. In this scenario, CPBAC algorithm, leads to an improvement in the base station throughput of around 3% higher than PLEBAC or PABAC algorithms alone.

Table 4 Obtained throughput: Indoor 3km/h, outdoor 50km/h (rb=384kbps).

Call duration	PLEBAC	PABAC	CPBAC
30 seconds	757.789 kbps	734.975 kbps	773.073 kbps
3 minutes	749.181 kbps	755.229 kbps	774.421 kbps

2.7 REPEATERS

This section summarizes the most relevant results obtained during the analysis of the introduction of Repeaters in a UMTS network, reported in [1], section 2.7.

A general reference covering this topic is [33].

2.7.1 Repeaters usage in WCDMA systems

The possible usage situations for repeaters in a WCDMA system are the following:

- coverage extension: in order to cover the so called “dead spot” (areas not covered during the first deployment of the network);
- capacity extension: in order to increase the capacity of a base station with an increased traffic load;
- soft-handover region reduction: thanks to the repeaters it is possible to reduce the soft-handover areas for already-connected users.

When a repeater is introduced in a WCDMA system, the first effect that it is possible to observe is the increase of the noise figure of the base station. Considering the coverage area, with the increase of the noise figure of the base station we have the receiver sensitivity making worse, with the consequence that the base station coverage decreases, although the total cell radius is increased due to the introduction of the repeater.

In terms of capacity, it is a challenge to analyse the effects due to the repeaters in a WCDMA system, like UMTS. In fact, the introduction of the repeaters leads to a modification of the interference characteristics of the scenario.

2.7.2 Analysed layout

The considered layout is showed in [1], section 2.7; the simulation area consists of:

- 16 NodeBs with omni-directional antennas
- 4 repeaters related to 4 hot-spots

In the simulations, the number of users has been varied from 10 to 90 per cell and from 0 to 40 per hot-spot. The considered service is voice for all users. In the performed simulations, the code limitation in downlink was considered, but not the Admission Control on the uplink.

2.7.3 Simulation results

In Figure 7 it is showed the percentage of served users when varying the repeater gain. We have considered two load situations: 50 users per cell and 20 users per hot-spot; 60 users per cell and 25 users per hot-spot.

In each traffic situation, three different conditions have been analysed:

- Neglecting the effects of the repeaters on the noise figure and the blocking due to codes in downlink (case III)
- Considering the effects of the repeaters on the noise figure and neglecting the blocking due to codes in downlink (case II)
- Considering both the effects of the repeaters on the noise figure and the blocking due to codes in downlink (case I).

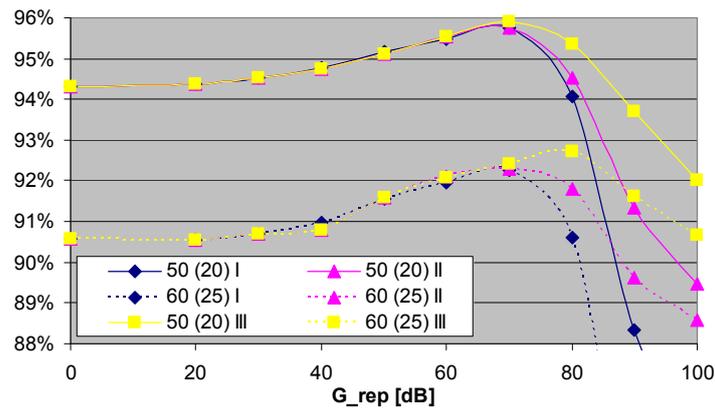


Figure 7 – Percentage of served users (I=noise figure modified, blocking on codes in downlink; II=noise figure modified, no blocking on codes; III=no noise figure modified, no blocking on codes)

The percentage of served users in the system increases with repeater gain values lower than 70-80 dB, when the maximum capacity value is obtained. Increasing the repeater gain, the influence of the repeater on the surrounding area increases, then the *best server area* of its donor base station increases also. Thus, in the hot-spot area the out-of-service users decrease and, at the same time, the number of users transmitting to the donor cell through the repeater increases.

Let us consider the case III of Figure 7. With repeater gain values greater than 70 dB, the traffic collected through the repeater causes a decrease of the system capacity. This decrease is due to outage in uplink of the terminals connecting both to the donor cell and to the adjacent cells.

In the case II of Figure 7 (in this case the noise figure is not constant), the number of terminals connected directly to the donor cell decreases more. Thus, the performances of the case II are worse than the case I with equal G_{rep} .

In the case I, we can notice that the performances are worse than in case II, since this time there is a number of users that have been blocked due to code blocking.

The case I of Figure 7 is considered in Figure 8. In this case, the code blocking leads to a decrease of served users also in the hot-spot with high gain values, while the out-of-service terminals increase in all the system.

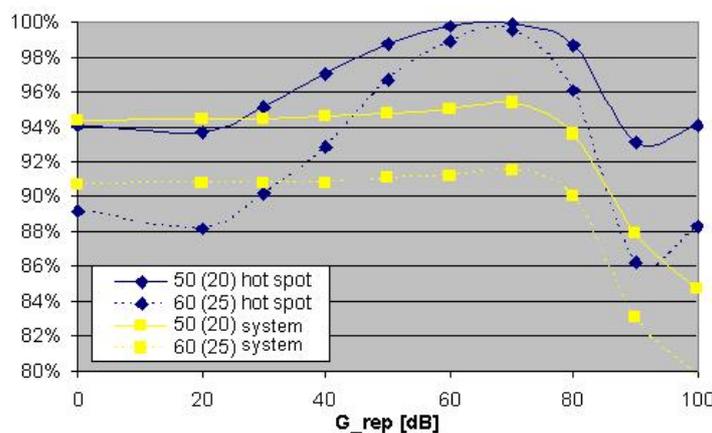


Figure 8 - Served users percentage in hot-spots and in the remaining parts of the system (case I of Figure 7)

In all the simulations performed, the downlink has never been blocking. In fact, the donor cell transmission power and the adjacent cells transmission power are different only with high value of repeater gain.

In the case 50 (20), the power decrease in adjacent cells is due to the decrease of load taken away from the cells with repeater, while in the case of the donor cells, the power decrease is due to the decrease of the users served directly by the station.

In the case 60 (25), the behavior of both the curves is quite similar, but there is a little power increase with 70-80 dB of repeater gain. In this case, the required power at the base station (due to the increasing number of connections) is not enough balanced by the transmitted power from the repeater.

With the obtained results, we can conclude that the introduction of repeaters in a WCDMA network allows increasing the capacity of the system. Moreover, it is possible to characterize a value for the repeater gain that maximizes the capacity.

The capacity gain depends on the considered scenario, however in the simulations presented it does not seem to be so high.

Further simulations have been performed varying the ratio between cell load and hotspot load, in order to evaluate the system capacity gain. For each simulation performed the maximum capacity increase varying the repeater gain has been evaluated and it is reported in Figure 9. From the next figure it is clear that the capacity gain obtainable using a repeater is generally low, unless scenarios with a high ratio between hotspot users and cell users are considered.

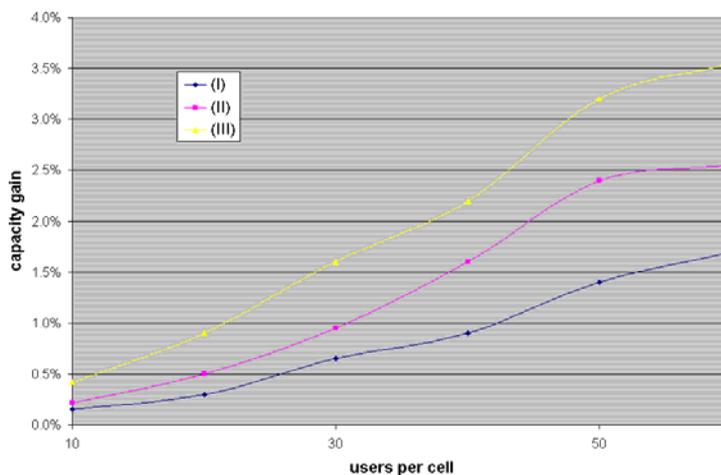


Figure 9 – Maximum capacity gain reachable with the repeater in the different scenarios considered

2.8 MULTIPLE RF CARRIERS

Here we assume that an UMTS operator uses two frequencies in a hot spot area. The scenario in [16] with an indoor traffic Hot Spot within an urban area is studied. A three sector site for frequency f_2 is used together with a macro layer of cells of frequency f_1 , see Figure 10.

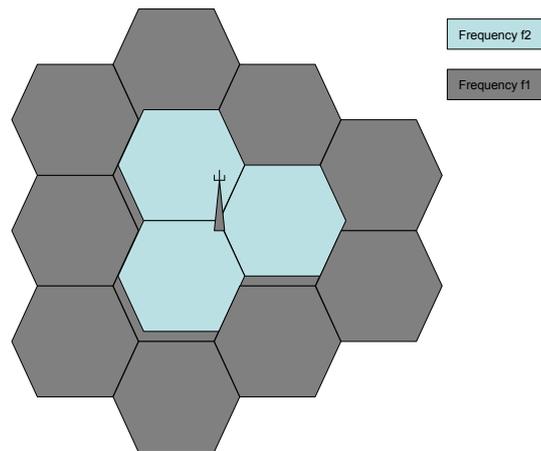


Figure 10. The studied scenario is a hot spot within urban area where a three sector site on frequency f2 is co-sited with a three sector site on frequency f1, which is a macro layer.

In this study we consider a mix of CS (speech) and PS (HTTP) traffic. Of the offered load in Bytes 25% is speech and 75% is HTTP as described in [16].

In principal, the same simulator as in the shared network study in [17] is used in this study. Inter-frequency cells have been added and some enhancements have been made to the simulator. The simulator does not include inter-frequency handover, i.e., handover between frequency f1 and f2, or vice versa. Admission and congestion control is the same as for the reference method in the shared network study in [17].

The goal of this study is to try to identify methods that increase bit rate and/or reduce blocking and dropping rates. We will study methods controlling the load on the two frequencies and the impact of controlling services to different frequencies.

As a reference method we let 50% of the load be allocated to frequency f1 and 50% of the load to frequency f2. Then we will analyse the impact of

- allocating more load to f2 than on f1,
- redirecting blocked service requests on frequency f1 to frequency f2 (or vice versa),
- allocating all speech to one frequency and
- turning off soft handover on frequency f2.

Before describing these methods in more detail we will briefly describe the admission control, congestion control and coverage control mechanisms used in this study. We will also describe the QoS measures used.

2.8.1 Admission control, congestion control and coverage control

Admission control is applied to guarantee a good QoS, i.e., to avoid overloading the system. Therefore, service requests are blocked if the OVSF code usage or power usage of a cell excides a limit. If total OVSF code usage or total power usage of a cell excides a limit at a service request, the system may reduce the bit rate of already allocated radio links. A service request is denied when OVSF code usage or power usage of a cell excides the limit even if bit rate of already allocated radio links would be reduced. At block of service requests we never reduce bit rate of allocated radio links below the bit rate of the service request. At power block of speech requests we allow reducing bit rate of dedicated PS radio links to 64 kbps. This method is used at blockage of most service requests.

The simulator is provided with congestion and coverage control methods that removes allocated radio links. The congestion control removes radio links if the total power excides a

certain limit, and the coverage control removes a radio link if it requires more power than the maximum allowed power for that radio link type. The congestion control prioritises to remove PS radio links over CS radio links.

2.8.2 QoS measures

The QoS quantities studied here are: CS and PS blocking, CS dropping and PS average bit rate per file. A CS block is registered if a CS service request is not admitted by the admission control (no queuing line is applied). A CS drop is registered when congestion or coverage control removes a CS radio link. For each file transfer the bit rate is derived and the average bit rate is registered. The PS service is assumed to be elastic, and if admission control denies a PS service request that service request is queued. A PS blocking is registered for each 10 seconds that a PS service request is queued. Moreover, if the congestion or coverage control removes a PS radio link the UE can make a new request of service, which might be queued. Started file transfers are then retained from where it was interrupted. No PS drop is registered. In practise the TCP session may timeout if the queuing delay is to long. Here it will give a lower registered bit rate and a registered block.

2.8.3 Assumptions and reflections

The studied scenario is a hotspot cell where the other-to-own cell interference ratio (I-factor) distribution is such that there is a high probability for a terminal to get a low I-factor compared to a normal urban cell scenario. This means that the studied cells can carry higher load than normal urban cells. Since frequency f1 is a macro cell layer with more neighbour cells than frequency f2, the average other-to-own cell interference ratio is higher for f1 than for f2. Hence, a cell on frequency f2 should be able to carry more load than a cell on frequency f1. When using the reference method, it is mostly common that admission control block a service request due to OVSF code usage has reached a maximum limit. This is common for both frequencies, but in particularly common for frequency f2 where almost all admission control denies are due to OVSF code blocks.

A UE that moves out of the coverage of frequency f2's three cells has to perform inter-frequency handover to frequency f1. The inter-frequency handover has a very high latency (around 10 seconds) due to compressed mode measurements on the other frequency. Call drops might therefore be common at inter-frequency handovers. The simulator does not allow simulating UEs moving around. Hence, call drops due to inter-frequency handovers are not shown in the simulation results.

The traffic is speech and HTTP only. Of the offered load in Bytes 25% is speech and 75% is HTTP as described in [16]. Furthermore, we assume zero delay for switching down bit rate of allocated radio links, and zero delay for setting up new radio links. Since TCP traffic is bursty, an allocated PS radio link is removed first when it has been inactive for 1 second.

Note that at other hot spot scenarios the results may be different.

2.8.4 Methods

As a reference method 50% of the speech requests are allocated to frequency f1 and the rest of the speech requests are allocated to frequency f2. The same allocation probability is applied for HTTP. Thus the traffic on frequency f1 is a mix of approximately 25% speech and 75% HTTP. This is also the case for f2.

As indicated above an f2 cell should be able to carry more load than an f1 cell. To illustrate this and the importance of good camping of UEs we will allocate more and more load to f2. From this we may also see how sensitive the QoS measures are to different cell camping distributions.

A way of controlling the load of the two frequencies is to redirect a service request to the serving cell to the other frequency. We will study the potential benefit of redirecting speech and HTTP requests separately. Hence, in one study we redirect a UE to f2 if its speech request is blocked on frequency f1, and vice versa. The speech request is redirected only once to the opposite frequency and a speech block is registered if the request is blocked on both f1 and f2. In another study we redirect HTTP requests that are blocked on frequency f1 to f2, and vice versa. Here the redirection continuous after a block on f1 and f2, since the service request is queued. A HTTP block is registered after each 10 seconds of queuing. The delay of speech redirection is zero seconds and the HTTP redirection delay is 100 ms. These delays are unrealistic and the HTTP bit rate will in practice be worse than what we will obtain here.

Due to high probability of dropped speech calls when a UE moves out of coverage of frequency f2 it can be attractive to allocate all speech to frequency f1. Therefore, we will study allocation of all speech to f1 and some of the HTTP traffic. If 33% of the HTTP load is allocated to f1 then 50% of the offered load (including the speech load) in bytes will be allocated to frequency f1. With this as a starting point, we will study the impact of allocating more and more HTTP traffic on f2.

As a totally opposite strategy we will study the impact on capacity if allocating all speech on f2. Since a speech service probably will use more power per OVFS code usage it is likely that this strategy will increase capacity because f2 capacity is mostly OVFS code usage limited.

Another way to increase capacity in a OVFS code limited scenario may be to turn off soft handover. Soft handover implies that OVFS codes are allocated for a radio link in more than one cell. By turning off soft handover some OVFS codes are set free and power usage are increased. We analyse the potential gain of this strategy by turning off soft handover on frequency f2.

2.8.5 Simulation study

Next we will show the simulation results. Note that the results are an average of a rather few simulations, and hence the results is mainly for guidance rather than quantification.

Figure 11 below shows the impact of allocating more and more speech and HTTP load to frequency f2. We see that it is the PS blocking that is affected mostly as the load increases. The PS blocking is as smallest when 52% of the load is allocated to frequency f2, but it starts to increase if f2 is loaded with more load. The speech blocking and dropping is fairly insensitive to shifting more and more load from frequency f1 to f2 (at least in the load range tested here). Hence, the admission control method protects the speech users from deteriorating QoS when load increases, as it should. The PS bit rate does not change much either as f2 carries more load. Even though it is difficult to see in Figure 11, speech blocking is lowest when f2 carries 52% of the load. The obtained speech blocking is 7% lower with 52% of the load on f2 than at 50%. The PS bit rate, which is 2% higher at 52% load on f2 than at 50% load on f2, is also highest at 52% load on f2. The PS blocking is 13% lower when 52% of the load is allocated to f2 than when f2 gets 50% of the load.

In conclusion, the system capacity is highest when 52% of the load is allocated to frequency f2. Note that this is not a general result when using a second frequency three sector site. At other hot spot scenarios the I-factor may be different for both frequency f1 and f2, which may lead to the highest capacity at another load distribution.

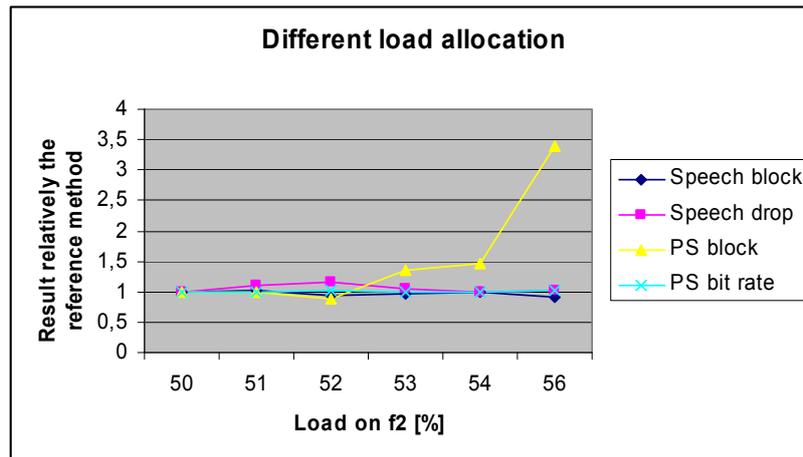


Figure 11. 50 to 56 percent of the load, speech and HTTP, is allocated to frequency f2.

The result obtain when redirecting service requests blocked on frequency f1 to f2, and vice versa, is shown in Figure 12. The first result in Figure 12, called “CS”, is the result of redirecting speech service requests. The second result, called “PS”, is the result of redirecting HTTP service requests when ever they are not admitted on the serving frequency. The third result, “PS after 10s”, is when HTTP service requests are redirected first after been queued 10 seconds on the serving frequency, i.e., after the HTTP service request has registered a PS block on the current frequency.

We see in the result that both speech and HTTP gains when redirecting speech services. The speech blocking is however only reduced about 6%, which is fairly little. On the other hand, when redirecting PS services, the PS blocking is reduced 84%. The speech service does not improve its QoS when HTTP requests are redirected. When redirecting a HTTP service first after 10 seconds, the PS blocking is also improved a lot. For this case the PS blocking is reduced 68%.

Note that the PS bit rate do not increase when redirecting PS services. In practise the PS bit rate would be even lower since the redirection delay is longer then the delay used here. Redirecting service requests that are blocked on serving frequency to the second frequency is as pooling the radio resources on frequency f1 and f2 together. This leads to trunking gains on blocking ratios and, hence, increased capacity.

The trunking gain on service requests that are not queued can be estimated by using the Erlang B formula. This formula gives the probability that all servers, N , are busy for the traffic intensity ρ . Hence, this is the probability of blocking a user that arrives to the cell if no queuing line is used. The Erlang B formula is

$$P(N) = \frac{\frac{\rho^N}{N!}}{\sum_{n=0}^N \frac{\rho^n}{n!}} \quad (6)$$

For speech services a UMTS cell implies a large amount of servers, since a speech service requires a fairly small amount of radio resources (power and OVFS codes). A PS radio link of 64 kbps or higher requires much more radio resources than speech. Thus a UMTS cell offers much less amount of servers for HTTP than for speech.

Assume that the UMTS access in a cell is a queuing system without waiting line and with 24 servers on frequency f1, and 24 servers on frequency f2. At 19 Erlang on each frequency we get 5% blocking according to the Erlang B formula. If the two frequencies are pooled together

into 48 servers we get 5% blocking at 42 Erlang. Hence, the load can be 11% higher at 5% grades of service when using redirection. If the 48 (2×19) servers are loaded with 38 Erlang, the blocking would be around 2%.

Assume now instead that each UMTS cell on frequency f1 and f2 consists of 10 servers. Then 6 Erlang on each frequency gives 5% blocking. By pooling the servers together the aggregated load on frequency f1 and f2 can be 15 Erlangs at 5% blocking ratio. Hence, at this lower amount of servers the load can increase 25% when the resources are pooled together without degraded blocking ratio. If the 20 servers are loaded with 12 (2×6) Erlangs the blocking would be around 1%.

The conclusion of this analysis is the HTTP blocking should reduce more than speech blocking ratio when redirection is used. This is in agreement with the simulation results even though one might expect that the speech service would gain more than the simulations shows. Note that when pooling more and more resources together the trunking gain will be less and less.

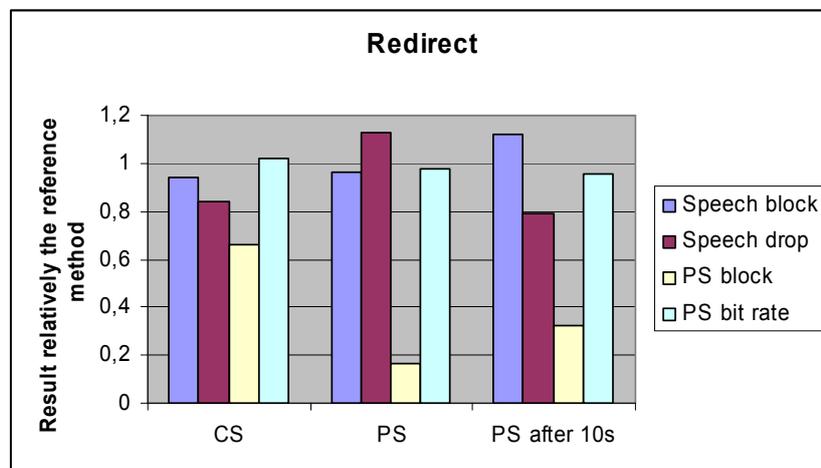


Figure 12. result obtain when redirecting service requests blocked on frequency f1 to f2, and vice versa

Services are redirected to the opposite frequency when a service request is blocked on the current frequency. "CS" means that only speech services are redirected, "PS" means that only HTTP services are redirected. "PS after 10 s" means that PS services are redirected after the service request has been queued 10 seconds and hence registered a PS block on the current frequency.

When all speech services and some fraction of the HTTP load is allocated to frequency f1 the obtained result is as shown in Figure 13. Frequency f1 and f2 gets 50% of the offered load in Bytes each when 33% of the HTTP load is allocated to frequency f1. At 45% HTTP load on f1, the terminals using HTTP are camping with almost the same probability on frequency f1 as on frequency f2. The blocking ratio for both PS and speech are very high in this case, but the speech dropping and PS bit rate are similar to the reference method.

The lowest PS blocking is at 33% HTTP load on f1, i.e., it is approximately here that HTTP reaches its maximum capacity. In this case, the PS blocking is almost half of the PS blocking for the reference method, but the speech QoS is worse than the reference method. The main reason why the speech blocking and dropping is worse in this case than the reference method is that speech uses more radio power per OVSF code tree usage than a PS radio link. On frequency f1 the radio power is a more scarce resource than it is on f2, which mainly block service requests due to high OVSF code usage. Allocating all speech to frequency f1 makes the power resource even more scarce on f1 and the power usage versus OVSF code

usage even more unbalanced on f2. Hence, it is more likely that a speech service will be blocked on frequency f1 than on f2.

The speech blocking becomes similar to the reference method first at 20% HTTP load on f1. At this load all other QoS measures are worse than for the reference method.

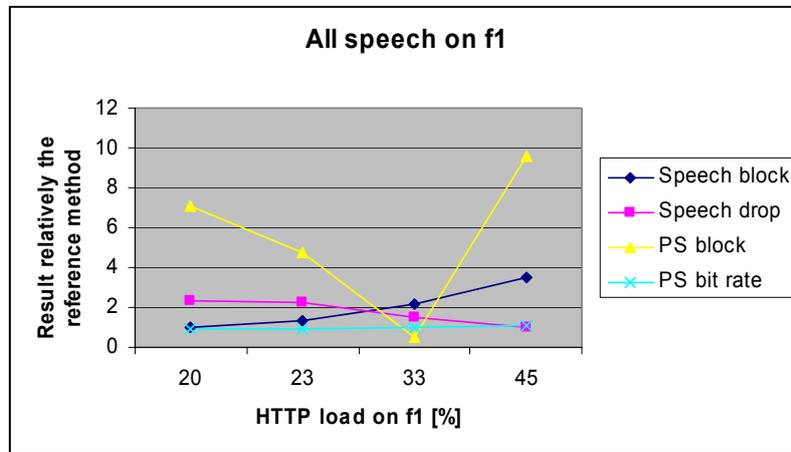


Figure 13. All speech is allocated to frequency f1. 20 to 45 percent of the HTTP load is allocated to frequency f1 as well.

A way to partly correct the unbalance of power usage versus OVFS code usage on frequency f2 is to allocate all speech on f2. Note that this may have the drawback of dropped calls when performing inter-frequency handover at the cell edge, which is not simulated here. In Figure 14 it is possible to compare the results of allocating all speech to f2 or all speech to f1. 33% of the offered HTTP load is allocated to the frequency with allocated speech service. The speech blocking is reduced approximately 6% compared to the reference method when all speech is allocated to frequency f2. The speech dropping is zero in this case and the HTTP QoS is similar to the reference method.

In conclusion, if an operator wants to improve HTTP blocking with the drawback of worsen the speech QoS, the strategy to allocate all speech to frequency f1 could be used. If an operator wants to improve speech QoS then all speech could be allocated to frequency f2 (assuming that inter-frequency handover does not cause dropped calls).

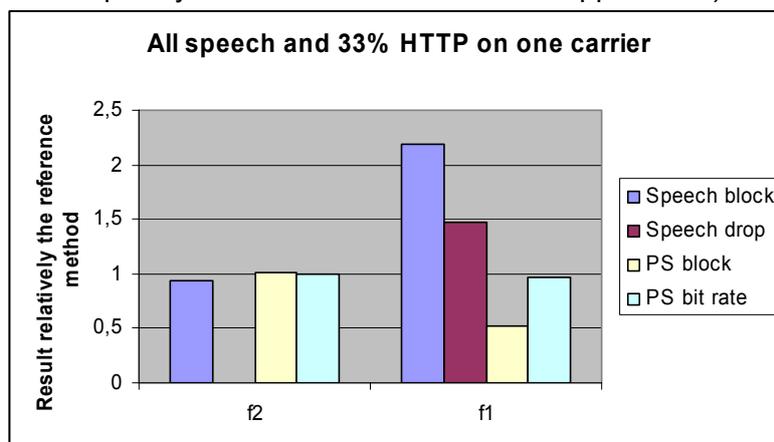


Figure 14. All speech load is allocated to frequency f1 or f2. 33% of the HTTP load is allocated to the frequency as all speech traffic uses.

If soft handover is tuned off on frequency f2, the PS bit rate is increased approximately 9%, see Figure 15. Of the methods we test here, this is the best way of increasing the PS bit rate. Unfortunately, the PS blocking is also increased around 10%. The speech blocking is not affected much when turning off soft handover on f2. The speech dropping is reduced 20%, which seems strange since turning off soft handover should increase the power usage and a speech service may only drop if congestion or coverage control finds out that the power usage is too high.

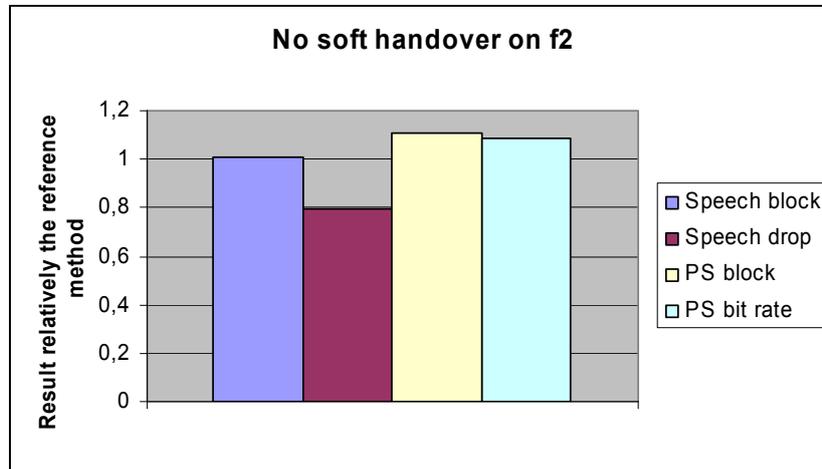


Figure 15. Soft handover is turned off on frequency f2.

2.8.6 Conclusions

We have simulated an UMTS operator using two frequencies in an indoor traffic Hot Spot within an urban area, where frequency f2 only consists of a three sector site and frequency f1 is a macro layer of cells. Different methods affecting on blocking ratio, dropping ratio and bit rate have been studied. The methods considered are unsymmetrical allocation of load to frequency f1 and f2, redirection of blocked service requests on frequency f1 to frequency f2 (or vice versa), allocation of all speech to one frequency and turning off soft handover on frequency f2.

The results indicate that PS blocking is sensitive to the load allocation on the frequencies, whereas speech blocking and dropping, and PS bit rate, is less sensitive to different load distributions. HTTP blocking is improved and the speech QoS is worsening when all speech is allocated to frequency f1. Speech QoS is improved then all speech is allocated to frequency f2 (assuming that inter-frequency handover does not cause dropped calls). If soft handover is tuned off on frequency f2, the PS bit rate is increased, but, unfortunately, the PS blocking is also increased.

Both speech and HTTP gain when speech services are redirected. The speech blocking is however reduced marginally. When redirecting PS services, the PS blocking is reduced significantly but the PS bit rate is not improved. The speech service does not improve its QoS when HTTP requests are redirected.

Redirecting service requests is as pooling together the radio resources of the frequencies. A service that requires a small amount of radio resources will have a smaller trunking gain of redirection than a service which requires a large amount of radio resources. When pooling more and more resources together the trunking gain will be less and less.

2.9 HIERARCHICAL CELL STRUCTURES

2.9.1 Cell selection/reselection criteria in HCS

The 3GPP standard describes in [18] the cell-selection and cell-reselection procedures taking into account the presence of a Hierarchical Cell Structure (HCS) layout.

In general, the selection depends on propagation conditions, user speed and network parameters. In the followings, the 3GPP algorithm will be explained, with the aim to highlight the most important parameters to set from an operator point of view.

2.9.1.1 Generalities

First of all, each terminal receives all the information about the measures to do for cell-reselection in the SYSTEM INFORMATION messages broadcast in the serving cell. The possible measures are divided in three groups:

- Intra-frequency measures: all the cells with the same frequency of the serving cell
- Inter-frequency measures: all the cells with frequency different from the serving cell
- Inter-RAT measures: all the cells of other RAT (e.g. GSM or CDMA2000)

In this description the Inter-RAT measures will not be considered; for further clarifications refer to [18].

When the terminal is camped on a cell (i.e. it performed the cell-selection algorithm founding a suitable cell), it continuously carries out measures of the strength and of the quality of the CPICH channel of the serving cell, according to:

$$S_{\text{qual}} = Q_{\text{qualmeas}} - Q_{\text{qualmin}}$$

$$S_{\text{rxlev}} = Q_{\text{rxlevmeas}} - Q_{\text{rxlevmin}} - P_{\text{compensation}}$$

Where:

Table 5 Parameters involved in the S criterion

S_{qual}	Cell Selection quality value (dB) Applicable only for FDD cells.
S_{rxlev}	Cell Selection RX level value (dB)
Q_{qualmeas}	Measured cell quality value. The quality of the received signal expressed in CPICH E_c/N_0 (dB) for FDD cells. CPICH E_c/N_0 shall be averaged. Applicable only for FDD cells.
$Q_{\text{rxlevmeas}}$	Measured cell RX level value. This is received signal, CPICH RSCP for FDD cells (dBm) and P-CCPCH RSCP for TDD cells (dBm).
Q_{qualmin}	Minimum required quality level in the cell (dB). Applicable only for FDD cells.
Q_{rxlevmin}	Minimum required RX level in the cell (dBm)
$P_{\text{compensation}}$	$\max(\text{UE_TXPWR_MAX_RACH} - P_{\text{MAX}}, 0)$ (dB)
UE_TXPWR_MAX_RACH	Maximum TX power level an UE may use when accessing the cell on RACH (read in system information) (dBm)
P_MAX	Maximum RF output power of the UE (dBm)

According to 3GPP standard, when the HCS is used each cell is assigned a *HCS priority level*, in order to differentiate cells belonging to different layers: the possible values of the HCS priority level are the integers between 0 and 7, for a maximum of eight layers. In general, the lowest HCS priority level is used for the widest range cells, while the highest HCS priority level is used for the shortest range cells. Then, in a dual-layer deployment with macro and micro layers, the macro layer will have a HCS priority level lower than the micro layer.

2.9.1.2 The mobility class evaluation

When HCS is used, it is important to assess the *mobility class* of the terminal since the cell-reselection algorithm is affected by the speed of the terminal. To evaluate its mobility class, the terminal will count the number of cell reselections performed in a fixed time period specified by the T_{CRmax} parameter. If the number of cell reselections during time period T_{CRmax} exceeds N_{CR} , *high-mobility* has been detected. When the number of cell reselections during time period T_{CRmax} no longer exceeds N_{CR} , the terminal shall continue these measurements during time period $T_{CRmaxHyst}$, and revert to measurements according to the threshold based measurement rules. The parameters involved in the mobility class evaluation are reported in the following table.

Table 6 Parameters involved in the mobility class evaluation

T_{CRmax}	This specifies the duration for evaluating allowed amount of cell reselection(s).
N_{CR}	This specifies the maximum number of cell reselections.
$T_{CRmaxHyst}$	This specifies the additional time period before the UE can revert to low-mobility measurements.

A mobility class evaluation example is depicted in Figure 16. In a two layers scenario, a user riding a bicycle has his terminal camped on a micro cell. Another user has his terminal camped on the micro layer, but he is moving with a sport car. Moving along the streets, the terminal of the first user performs one cell-reselection on the micro layer in 60 seconds, thus remaining in the low-mobility state since the threshold N_{CR} is set at 2 cell-reselection every $T_{CRmax} = 60$ seconds. The terminal of the latter user, instead, performs 3 cell-reselections on micro cells in 60 seconds, exceeding the threshold N_{CR} and then moving to high-mobility state. In the high-mobility state the terminal will prioritize the cell-reselection to macro layer cells.

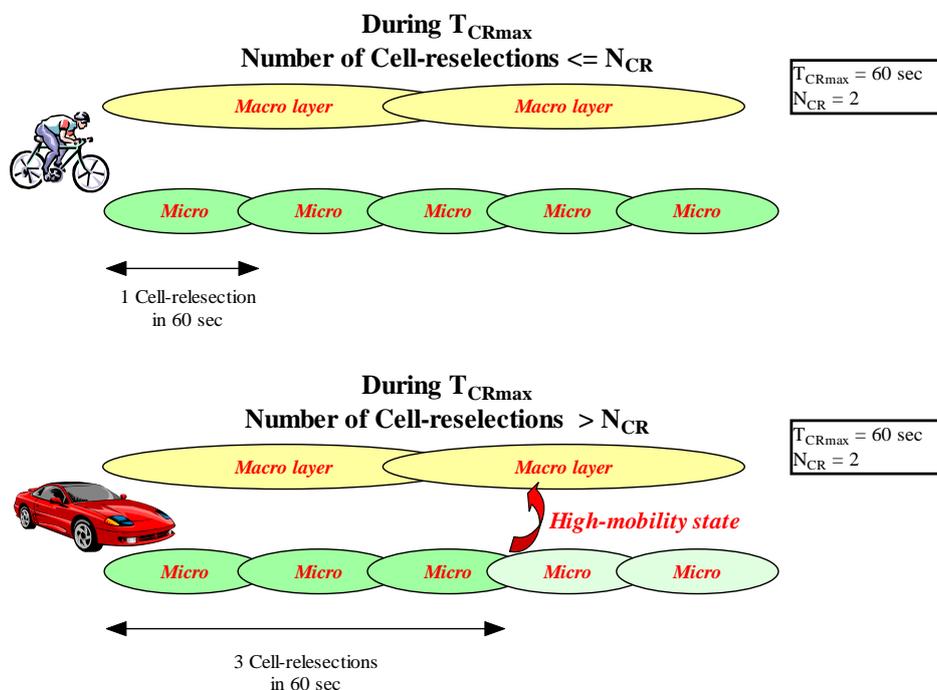


Figure 16 Mobility class evaluation

2.9.1.3 The cell-reselection algorithm

The cell-reselection algorithm is based on the thresholds reported in the following table.

table 7 Thresholds used in the cell-reselection algorithm

$S_{\text{searchHCS}}$	This threshold is used in the measurement rules for cell re-selection when HCS is used. It specifies the limit for S_{rxlev} in the serving cell below which the UE shall initiate measurements of all neighboring cells of the serving cell.
$S_{\text{intrasearch}}$	This specifies the threshold (in dB) for intra frequency measurements and for the HCS measurement rules.
$S_{\text{intersearch}}$	This specifies the threshold (in dB) for inter-frequency measurements and for the HCS measurement rules.

If the SYSTEM INFORMATION messages in the serving cell report that HCS is used, the terminal will behave according to the following **algorithm**:

```

IF ( $S_{\text{rxlev},s} \leq S_{\text{searchHCS}}$ ) or (if FDD and  $S_{\text{qual},s} \leq S_{\text{intersearch}}$ ) THEN
    measure on all intra-frequency and inter-frequency cells
ELSE
    IF ( $S_{\text{qual},s} > S_{\text{intrasearch}}$ ) THEN
        measure on all intra-frequency and inter-frequency cells, which have higher HCS priority level than the serving cell unless measurement rules for fast-moving UEs are triggered
    ELSE
        measure on all intra-frequency and inter-frequency cells, which have equal or higher HCS priority level than the serving cell unless measurement rules for fast-moving UEs are triggered
    ENDF
ENDIF

```

In case HCS is used and $S_{\text{intrasearch}}$ or $S_{\text{searchHCS}}$ or $S_{\text{intersearch}}$ (in FDD) are not sent in the SYSTEM INFORMATION messages for the serving cell, the terminal shall measure on all intra-frequency and inter-frequency cells.

In this high-mobility state, the terminal shall perform measurements on intra-frequency and inter-frequency neighboring cells, which have equal or lower HCS priority than serving cell and it shall prioritise re-selection of intra-frequency and inter-frequency neighboring cells on lower HCS priority level before neighboring cells on same HCS priority level.

The following figure represents the different behaviors of the terminal according to the algorithm reported above. The assumption of the picture is that the threshold $S_{\text{intrasearch}}$ is higher than $S_{\text{intersearch}}$ and the terminal is camped on the macro layer. When the terminal is in the position 1, i.e. when the measured S_{qual} of the serving cell is higher or equal to $S_{\text{intrasearch}}$, it performs measurements on intra-frequency and inter-frequency cells with the HCS priority level higher than the HCS priority level of the serving cell. When the user moves and terminal is in the position 2, i.e. when the measured S_{qual} of the serving cell is lower than $S_{\text{intrasearch}}$ and greater or equal than $S_{\text{intersearch}}$ and the measure S_{rxlev} of the serving cell is higher or equal than $S_{\text{searchHCS}}$, the measurements performed by the terminal are on intra-frequency and inter-frequency cells with the HCS priority level higher or equal than the HCS priority level of the serving cell. Finally, when the user moves and the terminal is in the position 3, i.e. when the measured S_{qual} of the serving cell is lower than $S_{\text{intersearch}}$ or the measure S_{rxlev} of the serving cell is lower than $S_{\text{searchHCS}}$, the measurements performed by the terminal are on all the intra-frequency and inter-frequency cells.

UMTS cell-reselection with HCS
no fast-moving UE

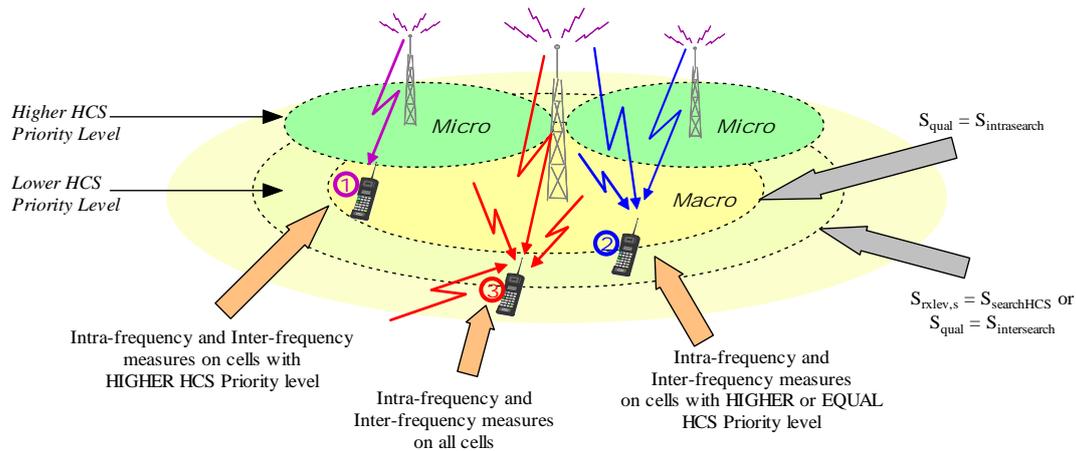


Figure 17 Cell reselection with HCS and low-mobility terminal

The measures carried out by the terminal are used to evaluate if a cell-reselection is needed. First of all, a cell must satisfy the so called **S criterion** that fixes the minimum level of quality to consider a cell suitable. This criterion is based on the measures of S_{qual} and S_{rxlev} of the serving cell; the criterion is satisfied if:

$$S_{qual} > 0 \text{ and } S_{rxlev} > 0$$

After selecting a list of suitable cells, the terminal has to evaluate the second criterion, called **H criterion**. This criterion is valid only for low-mobility terminals and it is based on the following relationships:

$$H_s = Q_{meas,s} - Q_{hcs,s}$$

$$H_n = Q_{meas,n} - Q_{hcs,n} - T_{on} * L_n$$

$$T_{on} = TEMP_OFFSET_n * W(PENALTY_TIME_n - T_n)$$

$$L_n = 0 \quad \text{if } (HCS_PRIO_n = HCS_PRIO_s)$$

$$L_n = 1 \quad \text{if } HCS_PRIO_n \neq HCS_PRIO_s$$

$$W(x) = 0 \quad \text{for } x < 0$$

$$W(x) = 1 \quad \text{for } x \geq 0$$

Table 8 Parameters involved in the H criterion

Q_{meas}	Quality value. The quality value of the received signal derived from the averaged CPICH E_c/N_0 or CPICH RSCP for FDD cells, from the averaged P-CCPCH RSCP for TDD cells and from the averaged received signal level for GSM cells. For FDD cells, the measurement that is used to derive the quality value is set by the Cell selection and reselection quality measure information element.
HCS_PRIO_s, HCS_PRIO_n	This specifies the HCS priority level (0-7) for serving cell and neighboring cells. HCS priority level 0 means lowest priority and HCS priority level 7 means highest priority.
Q_{hcs_s}, Q_{hcs_n}	This specifies the quality threshold levels for applying prioritised hierarchical cell re-selection to serving and neighbor cells.
$PENALTY_TIME_n$	This specifies the time duration for which the $TEMPORARY_OFFSET_n$ is applied for a neighboring cell.
$TEMPORARY_OFFSET_{1n}$	This specifies the offset applied to the H and R criteria for a neighboring

	cell for the duration of PENALTY_TIME _n . It is used for TDD and GSM cells and for FDD cells in case the quality measure for cell selection and re-selection is set to CPICH RSCP.
TEMPORARY_OFFSET2 _n	This specifies the offset applied to the H and R criteria for a neighboring cell for the duration of PENALTY_TIME _n . It is used for FDD cells in case the quality measure for cell selection and re-selection is set to CPICH Ec/No.

The H criterion is satisfied for a neighbor cell when it has:

$$H_n > H_s \quad (7)$$

The cells that passed the H criterion have to be ranked with the ***R criterion***. In case the H criterion has not been applied due to high-mobility state of the terminal, just the R criterion is used. The R criterion is based on the following relationships:

$$R_s = Q_{\text{meas},s} + Q_{\text{hyst}} \quad (8)$$

$$R_n = Q_{\text{meas},n} - Q_{\text{offset},s,n} - TO_n * (1 - L_n)$$

Where TO_n and L_n have been defined above, the others parameters are:

Table 9 Parameters involved in the R criterion

Q_{meas}	Quality value. The quality value of the received signal derived from the averaged CPICH Ec/No or CPICH RSCP for FDD cells, from the averaged P-CCPCH RSCP for TDD cells and from the averaged received signal level for GSM cells. For FDD cells, the measurement that is used to derive the quality value is set by the <u>Cell_selection_and_reselection_quality_measure</u> information element.
$Q_{\text{offset}1_{s,n}}$	This specifies the offset between the serving cell and neighbor cell. It is used for TDD and GSM cells and for FDD cells in case the quality measure for cell selection and re-selection is set to CPICH RSCP.
$Q_{\text{offset}2_{s,n}}$	This specifies the offset between the serving cell and neighbor cell. It is used for FDD cells in case the quality measure for cell selection and re-selection is set to CPICH Ec/No.
$Q_{\text{hyst}1_s}$	This specifies the hysteresis value (Q_{hyst}) of the serving cell. It is used for TDD and GSM cells and for FDD cells in case the quality measure for cell selection and re-selection is set to CPICH RSCP.
$Q_{\text{hyst}2_s}$	This specifies the hysteresis value (Q_{hyst}) of the serving cell. It is used for FDD cells if the quality measure for cell selection and re-selection is set to CPICH Ec/No.

The R criterion is satisfied for a neighbor cell when it has:

$$R_n > R_s \quad (9)$$

After the ranking with the R criterion, the first cell in the ranked list, if it exists, is the most suitable cell to camp on.

2.9.1.4 Some guide-lines for setting the cell-reselection parameters

The criteria reported above are based on two main categories of settable parameters: offset and threshold parameters. By means of these parameters, it is possible for a network operator to guide the cell reselection performed by the terminals, according to some specific rules:

- it is possible to favour one layer or another one
- it is possible to move “fast moving” (i.e. high mobility) traffic on lower hierarchical layers (e.g. macro cells) and the “slow moving” (i.e. low mobility) traffic on higher hierarchical layers (e.g. micro cells), thus avoiding the “ping-pong” effect amongst the layers.

The settable offset parameters involved in the cell-reselection process are grouped below with the indication of the affected criterion (R or H):

Table 10 Offset and threshold parameters involved in the cell-reselection

Qoffset1 _{s,n}	This specifies the offset between the serving cell and neighbor cell. It is used for TDD and GSM cells and for FDD cells in case the quality measure for cell selection and re-selection is set to CPICH RSCP.	<i>R criterion</i>
Qoffset2 _{s,n}	This specifies the offset between the serving cell and neighbor cell. It is used for FDD cells in case the quality measure for cell selection and re-selection is set to CPICH Ec/No.	<i>R criterion</i>
Qhyst1 _s	This specifies the hysteresis value (Qhyst) of the serving cell. It is used for TDD and GSM cells and for FDD cells in case the quality measure for cell selection and re-selection is set to CPICH RSCP.	<i>R criterion</i>
Qhyst2 _s	This specifies the hysteresis value (Qhyst) of the serving cell. It is used for FDD cells if the quality measure for cell selection and re-selection is set to CPICH Ec/No.	<i>R criterion</i>
Qhcss, Qhcsn	This specifies the quality threshold levels for applying prioritised hierarchical cell re-selection to serving and neighbor cells.	<i>H criterion</i>

Remembering the *R criterion* explained above, in order to favour a cell during the cell-reselection the R_n value of this cell should be increased. To do this, it is possible to set the Qoffset1/2_{s,n} parameters with negative values: more the Qoffset1/2_{s,n} is high (in absolute value) more the R_n value increases.

Instead, in order to favour the serving cell, the R_s value of this cell should be increased. To do this, it is possible to set the Qhyst1/2_s parameters with positive values: more the Qhyst1/2_s is high (in absolute value) more the R_s value increases.

Finally, for low-mobility users, it is possible to favour the cell-reselection of a layer or another using the Qhcs_{s/n} parameters used in the *H criterion*. The Qhcs_{s/n} is the quality threshold level of the serving cell or the neighbor cell. From the H criterion explained above, it is clear that to favour a cell-reselection H_n value has to be greater than H_s . Using a low value for the Qhcs_n parameter H_n will have a high value, promoting the cell-reselection to the neighbor cells. On the contrary, using high values for the Qhcs_n parameter (i.e. using a high quality threshold for the neighbor cells), the H_n value will have a low value, thwarting the cell-reselection to the neighbor cells.

2.9.1.5 Simulated layout

The study aims to present some guidelines in order to segregate the traffic between layers, according to the mobility class of the users: the macrocells should be used for the high-mobility traffic, while the microcell should be used for the low-mobility traffic. The goal of the work is to show the behavior of the system with different values of the cell-reselection parameters.

The evaluations have been carried out using an event-driven simulator, in which the HCS layout showed in the figure below has been implemented. The simulated layout foresees two layers:

- Macrocells, with 1 km of cell radius
- Microcells, with 0.5 km of cell radius.

The area covered by the simulated layout has a rectangular shape with a surface of about 24 square kilometres.

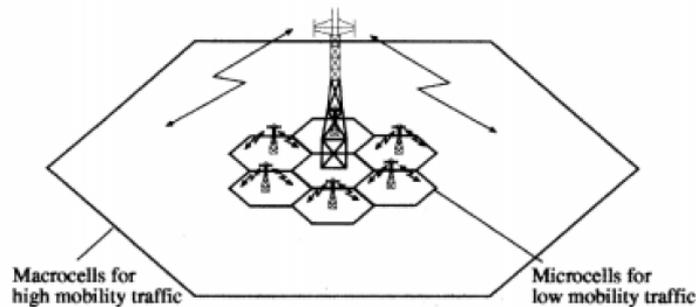


Figure 18 Considered hierarchical layout

In the considered situation, both macrocellular and microcellular layers use the same radio frequency; thus, in this case the layer discrimination is only *power-based*.

In the simulator two classes of users have been considered:

- Low-mobility users with an average speed of 3 km/h
- High-mobility users with an average speed of 50 km/h

The total amount of users in the system is equally divided between the two mobility-classed defined above, i.e. half of the users are low-mobility users and half of the users are high-mobility users. All the users require voice services.

The macrocellular layer is a 12 NodeBs deployment, each with tri-sectorial antennas, obtaining 36 cells in total (see figure below).



Figure 19 Macrocellular layer deployment

The microcellular layer is a 48 NodeBs deployment. In the evaluations, two different antennas for the microcellular layer have been considered:

- omnidirectional antenna, as depicted in Figure 20, obtaining 48 cells in total
- tri-sectorial antenna, as depicted in Figure 21, obtaining 144 cells in total.

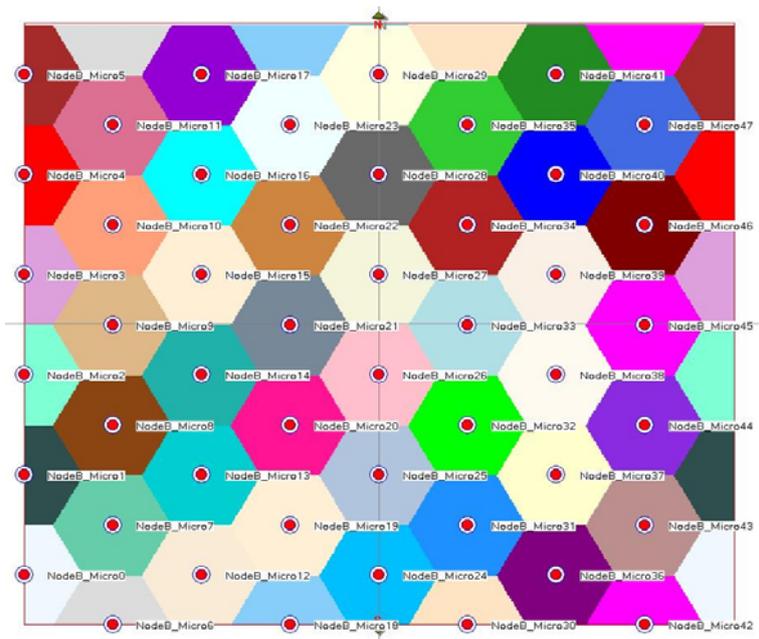


Figure 20 Microcellular layer deployment with omni-directional antennas

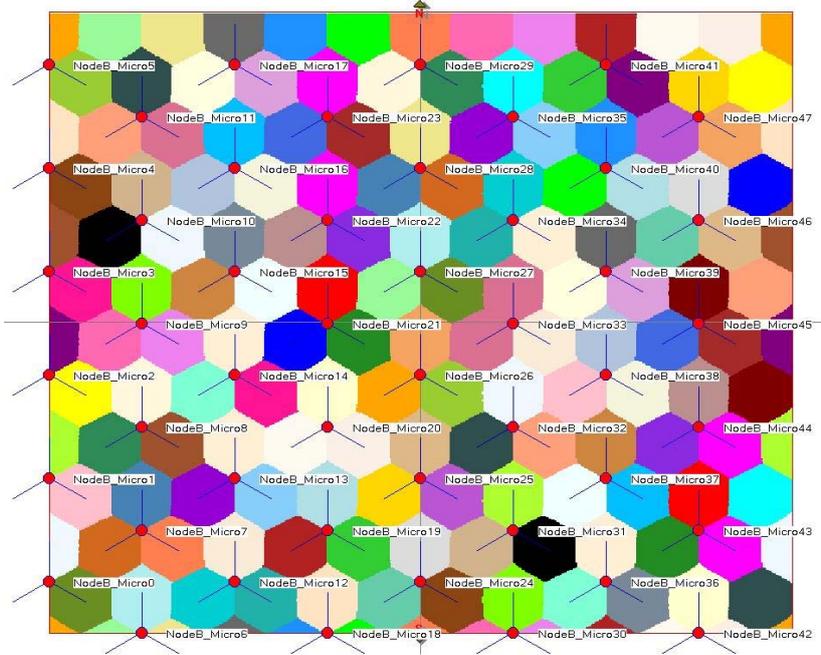


Figure 21 Microcellular layer deployment with tri-sectorial antennas

Both the macrocellular and the microcellular layers have been implemented using the *wrap-around technique*, by which it is possible to collect statistics from all the cells in the scenario, reducing the simulation time needed to obtain statistically acceptable results.

The propagation parameters used in the simulations are reported in the following table. It has to be noted that, with the aim to better catch the system behavior, the shadowing has not been considered.

Table 11 Propagation parameters

Parameter	Macrocellular layer	Microcellular layer with omni-directional antennas	Microcellular layer with tri-sectorial antennas

Maximum Coupling Loss (MCL)	70 dB	53 dB	53 dB
Antenna Gain	14 dB	2 dB	6 dB
CPICH Transmission power	27 dBm	17 dBm	17 dBm
Maximum NodeB transmission power	43 dBm	38 dBm	38 dBm
Noise figure at the receiver in the NodeB	2 dB	14 dB	14 dB

The propagation models used in the simulations for each type of cellular deployment are reported in the following table:

Table 12 Propagation models

<i>Propagation models</i>		
$L = L_0 + 10 \cdot \alpha \cdot \text{LOG}_{10}(d)$		
<i>Macrocellular layer</i>	<i>Microcellular layer with omni-directional antennas</i>	<i>Microcellular layer with tri-sectorial antennas</i>
$L_0 = 142$ $\alpha = 3.7$	$L_0 = 123$ $\alpha = 4.5$	$L_0 = 128$ $\alpha = 3.0$

According to Table 12, different parameter values were considered for the propagation rule of omni-directional and tri-sectorial micro-cellular layers. In fact, the most common micro-cell deployment foresees to use omni-directional antennas at medium or low heights [19] the resulting propagation behavior can be modeled by the parameter values shown in the corresponding column of Table 12. Alternately, also a micro-cellular tri-sectorial deployment can be assumed, locating the antennas at the same height of roofs [20]. In this last scenario, the parameter values of the propagation rule are different, as shown in the corresponding row of Table 12.

From a propagation point of view, the best-server map of the macrocellular layout is reported in next figure

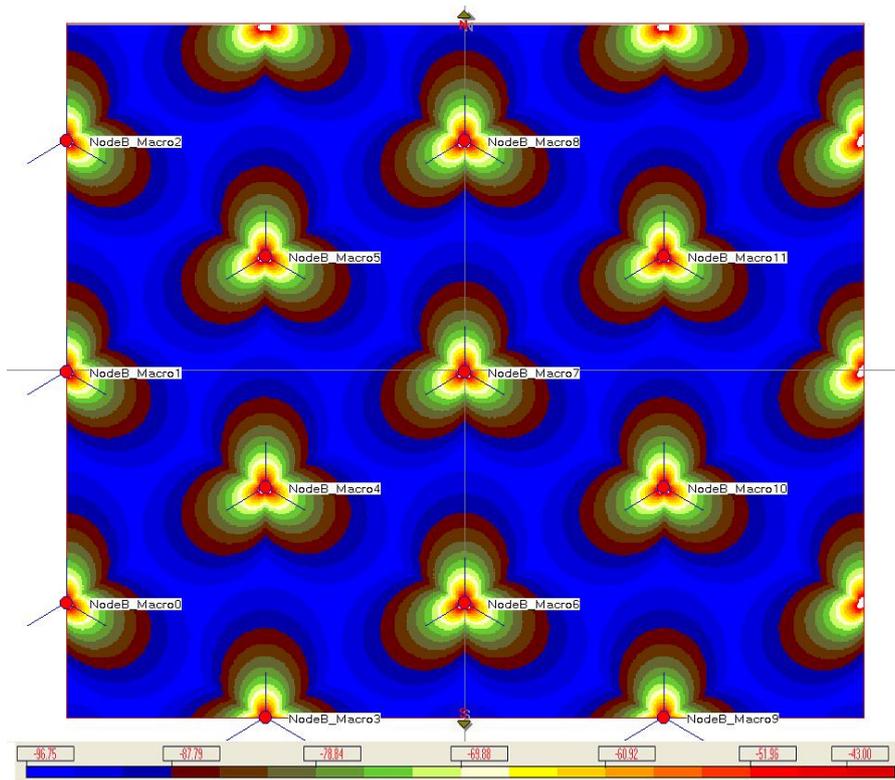


Figure 22 Best-server propagation loss in the macrocellular layer

The best-server map of the microcellular layout with omni-directional antennas is reported in next figure.

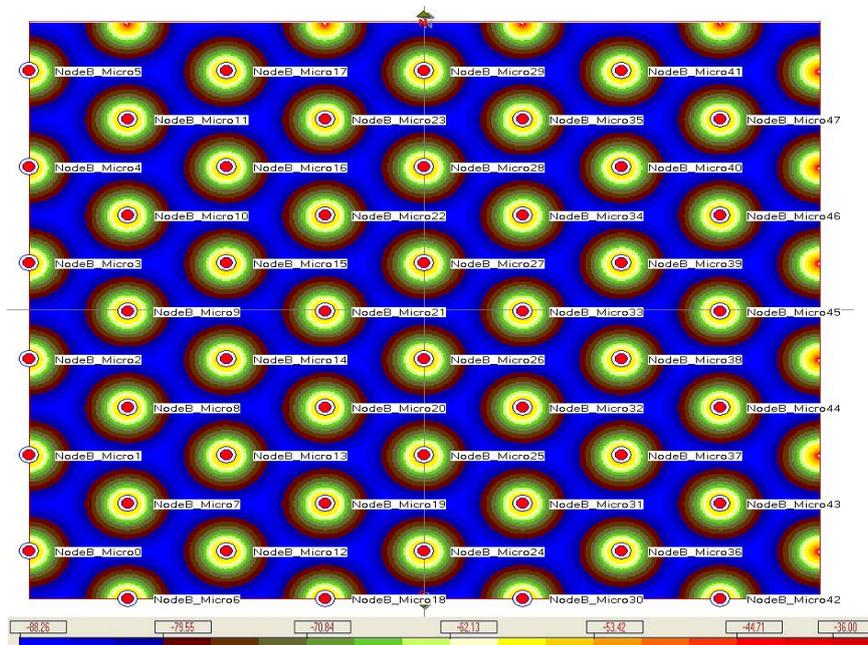


Figure 23 Best-server propagation loss in the microcellular layout

The best-server map of the microcellular layout with tri-sectorial antennas is reported in the figure below:

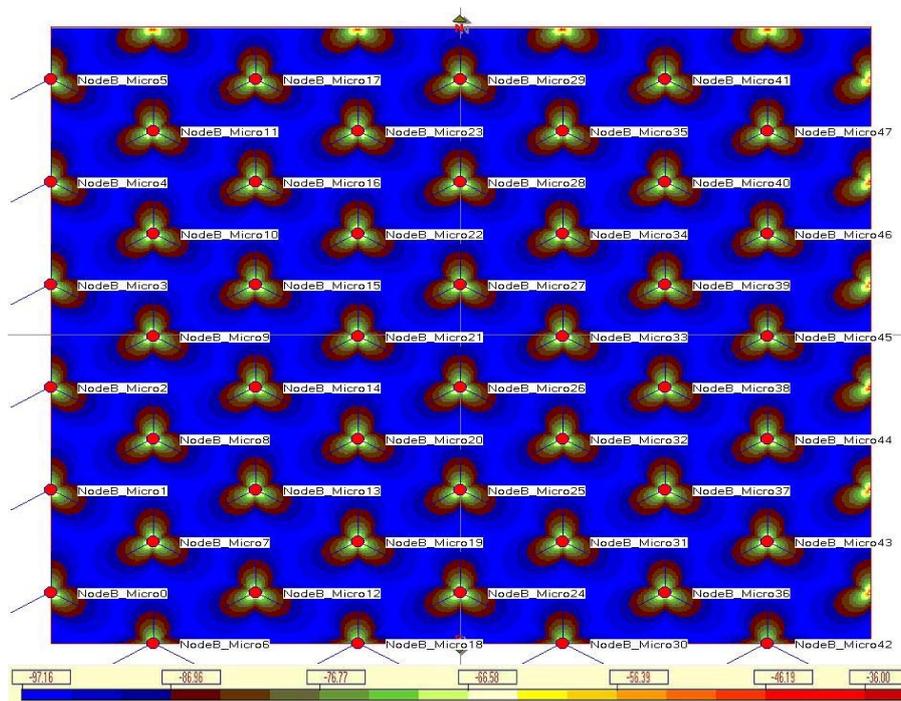


Figure 24 Best-server propagation loss in the microcellular layout

Comparing Figure 23 and Figure 24, it can be argued that the coverage offered by the considered tri-sectorial micro-cellular layout is more restricted than the omni-directional case, due to the two different cases that these deployment scenarios represent. The resulting propagation areas are in accordance with the propagation models shown in Table 12.

2.9.2 Capacity enhancement with HCS

2.9.2.1 Analysed layout

The considered HCS layout is showed in [1].

The study was aimed to assess the performance in a multi-layer deployment versus a single-layer layout, in terms of capacity, taking into account two situations have been considered:

- A) both macrocellular and microcellular layers use the same radio frequency; in this case the layer discrimination is only *power-based*.
- B) macrocellular and microcellular layers use different radio frequencies; in this case the layer discrimination is the radio frequency.

In the situation in which the same radio frequency is used for both layers, the boundary between the two layers may be defined in spatial terms, considering the position of the *break-point*, in which the CPICH received power from one layer is equal to the CPICH received power from the other layer (see Figure 25)¹. The break-point determines the coverage area of the layer. In the considered situation, the microcellular layer is “spatially isolated” from the macrocellular layer.

The break-point position depends on the antenna height, the distance between microcell and macrocell, the propagation environment.

¹ The break-point position depends on the antenna height, the distance between microcell and macrocell, the propagation environment.

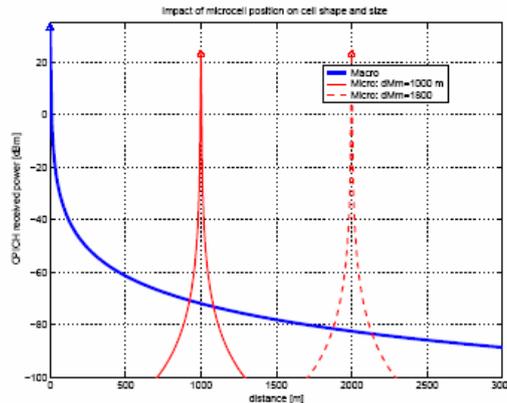


Figure 25 – Hierarchical layout with single carrier: the break-point concept

In the situation, instead, in which the two layers use different radio frequencies, the management of this system is easier since the inter-layer interference is not so strong as in the case of single frequency. In this case the main cause of interference is the energy “overflow” from the adjacent radio channel.

The analysis performed has been based on the segregation of users according to the level of mobility, taking into account two admitted values: HI (for high mobility) and LO (for low mobility). Two services have been considered: voice traffic and 64 kbps data traffic. The study has been performed by means of Montecarlo simulations and the comparison between the situation with HCS and the situation without HCS has been done. The reference condition to obtain statistics is when the percentage of outage is 2-3% for voice and 4-5% for data.

2.9.2.2 Simulation results

Analysing the simulation results, it has been noted that the main blocking reason in the multi-layer layout with both one and two carriers is the availability of channelisation codes.

The expected results about the availability of channelisation codes in case of two carriers should be more or less analogous to the results in case of single carrier. Instead, with two carriers the performance are better than the case with a single carrier. The reason of this improvement is due to the soft-handover: in case of single carrier, some User Equipments could be in a inter-layer macrodiversity situation, using channelisation codes both from macro and micro layers; instead, in case of two carriers, the User Equipments cannot be in an inter-layer macrodiversity situation, and the channelisation codes used in the situation A can be freed and used for other connections.

More in detail, considering only the single macrocellular layer, the capacity was of 35 users/cell, corresponding to 52 users/km². Introducing also the micro-layer, the capacity was increased up to 28 users/cell, corresponding to 98 users/km² for the case A (same frequency for both layers) and up to 35 users/cell, corresponding to 122 users/km² for the case B (two frequencies). Comparing the values of capacity without and with the microcellular layer, the improvement was about 86% using only one frequency (case A) and 133% using two frequencies (case B).

2.9.3 RRM and mobility issues in HCS idle mode

2.9.3.1 Introduction

This section of the document reports the dynamic evaluations performed on HCS, taking into account macro and micro layers. The activity's goals have been to identify how to set the thresholds controlling the algorithm adopted by the terminal for the identification of the high/low mobility status by means of the number of cell reselections performed in a fixed period of time.

The first section describes the 3GPP cell-reselection management in case of HCS. In the second section the simulation layout is presented and in the third section the main simulation results and conclusions are reported.

2.9.3.2 Cell reselection criteria with HCS

This section of the document reports the dynamic evaluations performed on HCS, taking into account macro and micro layers. The activity's goals have been to identify how to set the thresholds controlling the algorithm adopted by the terminal for the identification of the high/low mobility status by means of the number of cell reselections performed in a fixed period of time.

The first section describes the 3GPP cell-reselection management in case of HCS. In the second section the simulation layout is presented and in the third section the main simulation results and conclusions are reported.

The 3GPP standard describes in [18] the cell-selection and cell-reselection procedures taking into account the presence of a Hierarchical Cell Structure (HCS) layout.

In general, the selection depends on propagation conditions, user speed and network parameters. In the followings, the 3GPP algorithm will be explained, with the aim to highlight the most important parameters to set from an operator point of view.

The evaluations have been divided in two groups:

- Setting the offsets and the thresholds of the cell-reselection algorithm in order to segregate the traffic according to the mobility-class of the terminals
- Capacity analysis using different values of offsets and thresholds of the cell-reselection algorithm.

2.9.3.2.1 Traffic segregation

This analysis has been carried out in both the HCS layouts introduced in the previous section.

First of all, some preliminary simulations without all the offsets and hysteresis parameters set to 0 have been done. The results of these evaluations showed that *the microcellular layer is predominant in the system*, since most of the terminals are camped on it.

The aim of this first analysis is to segregate the traffic according to the mobility class of the users, then it is needed to find the right setting of values of the cell-reselection parameters. Since the microcellular layer is prevalent, such settings should move traffic from the microcellular layer to the macrocellular layer, above all *the high-mobility traffic should be segregated in the macrocellular layer*.

The parameters have been set as follows:

- $S_{\text{intrasearch}} = 8 \text{ dB}$

- $S_{\text{intersearch}} = 0$ dB
- $S_{\text{searchHCS}} = 25$ dB in case of microcellular layer with omni-directional antennas;
 $S_{\text{searchHCS}} = 15$ dB in case of microcellular layer with tri-sectorial antennas

Afterwards, different simulations with different values of $Q_{\text{offset}2_{s,n}}$ for the macrocellular cells have been carried out: the values of $Q_{\text{offset}2_{s,n}}$ have been swept from 0 to -14 dB with a step of 2 dB. It has to be noted that the usage of $Q_{\text{offset}2_{s,n}}$ means that the measures on the neighbor cells performed by the terminals is the CPICH E_c/N_0 ; in case of measures of CPICH RSCP the affected parameters should have been $Q_{\text{offset}1_{s,n}}$.

Finally, the last parameters to set are related to the mobility class evaluation. Since the cell radius of the macrocellular layer is about 1 km and the speed of the high mobility users is 50 km/h, the average cell-reselection period should be 72 seconds. Then, the used settings are the followings:

- $T_{\text{CRmax}} = 180$ s
- $N_{\text{CR}} = 2$
- $T_{\text{Crmaxhyst}} = 60$ s

2.9.3.2.2 Microcellular layer with omni-directional antennas

The results obtained in case of the microcellular layer with omni-directional antennas are reported in Figure 26, Figure 27, Figure 28 and Figure 29.

From Figure 26 and Figure 27 it could be noted that, decreasing the Q_{offset} values of macro cells, the percentage of time during which the high-mobility users stay on the macrocellular layer increases while the staying of low-mobility users are not affected. Without the setting of the Q_{offset} value (i.e. with Q_{offset} set to 0), the predominance of the microcellular layer is really evident: only 30% of high-mobility users is on the macro layer, and nearly 95% of low-mobility users is on the micro layer. With the lowest value of Q_{offset} , instead, the percentage of high-mobility users on the macro layer pass 65%.

Moreover, in Figure 26 it has also to be noted that with Q_{offset} values lower than -10 dB the percentage does not increase more, as for Q_{offset} values higher than -4 dB the percentage does not decrease significantly. It could be concluded that in the simulated layout, the most affecting values of Q_{offset} are between -4 dB and -10 dB.

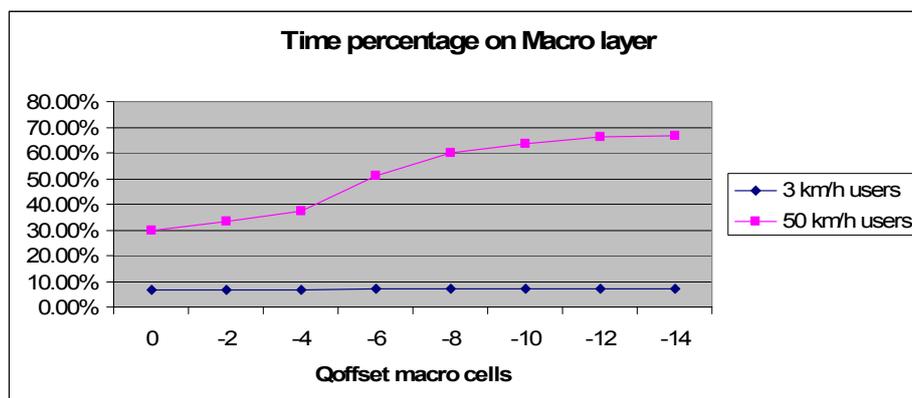


Figure 26 Percentage of time of stay on the Macro layer – Micro with omni-directional antennas

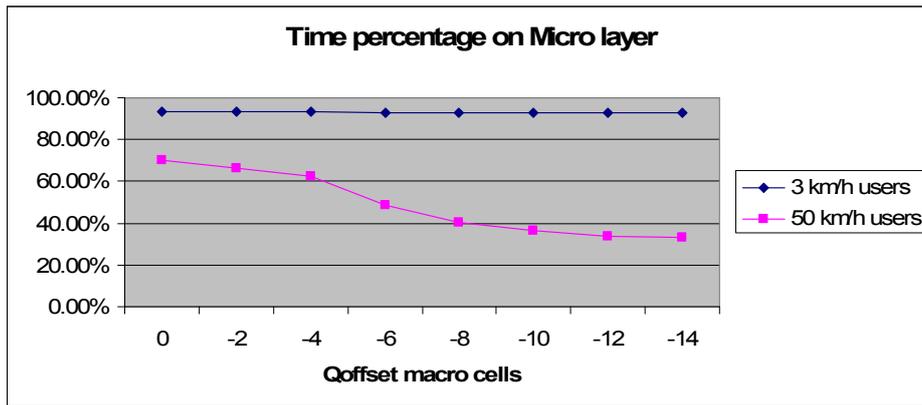


Figure 27 Percentage of time of stay on the Micro layer – Micro with omni-directional antennas

In Figure 28 and Figure 29 the time between two consecutive cell-reselections versus the Qoffset values of macro cells has been depicted, respectively for high-mobility users and low-mobility users. It should be noted that, decreasing the values of Qoffset, the time between cell-reselections decreases. In fact, decreasing the Qoffset value, the cell-reselection to macro cell is more and more favoured, thus increasing the total number of cell-reselections and reducing the time between two consecutive reselections. This phenomenon is very clear in the case of high-mobility users, while it is less marked in the case of low-mobility users.

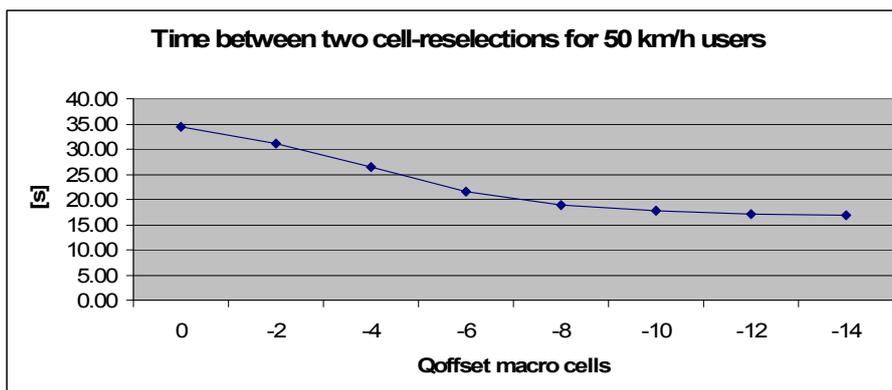


Figure 28 Time between two consecutive cell-reselections for high-mobility users – Micro with omni-directional antennas

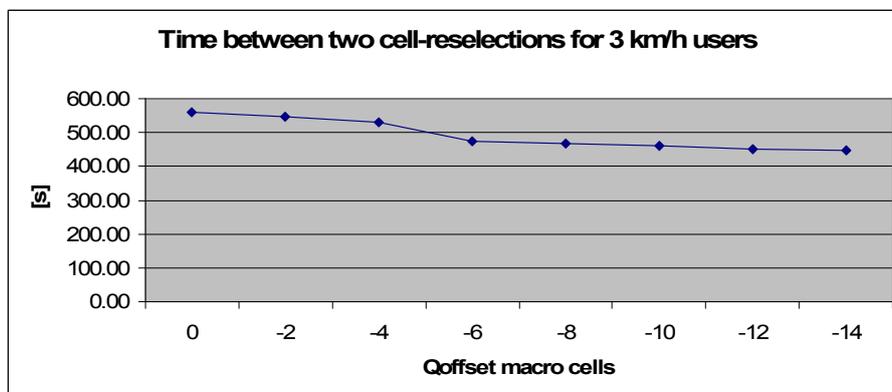


Figure 29 Time between two consecutive cell-reselection for low-mobility users – Micro with omni-directional antennas

Another analysis performed is the evaluation of the percentage of time in which the terminals are in the high-mobility state or in the low-mobility state. The values chosen for the mobility-class evaluation parameters are good, as shown in next figure: nearly 100% of the high-mobility users are always in fast-mobility state, while less than 1% of the low-mobility users are in the fast-mobility state.

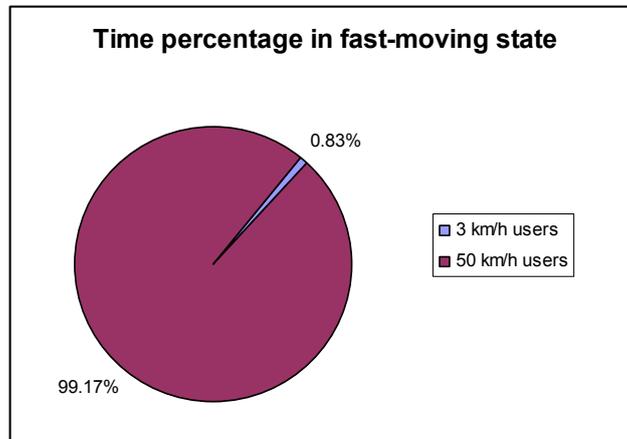


Figure 30 Percentage of time of staying in the fast-mobility state for both the simulated mobility classes of the users

2.9.3.2.3 Microcellular layer with tri-sectorial antennas

The results obtained in case of the microcellular layer with tri-sectorial antennas are reported in Figure 31, Figure 32, Figure 33 and Figure 34.

From Figure 31 and Figure 32 it could be noted that, decreasing the Qoffset values of macro cells, the percentage of time during which the high-mobility users stay on the macrocellular layer increases while the staying of low-mobility users are not affected. Without the setting of the Qoffset value (i.e. with Qoffset set to 0), the predominance of the microcellular layer is less evident than in the case of microcellular layer with omni-sectorial antennas: 65% of high-mobility users is on the macro layer, and nearly 90% of low-mobility users is on the micro layer. With the lowest value of Qoffset, instead, the percentage of high-mobility users on the macro layer pass 90%.

Moreover, in Figure 31 it has also to be noted that with Qoffset values lower than -8 dB the percentage does not increase more significantly. It could be concluded that in the simulated layout, the most effecting values of Qoffset are between higher than -8 dB.

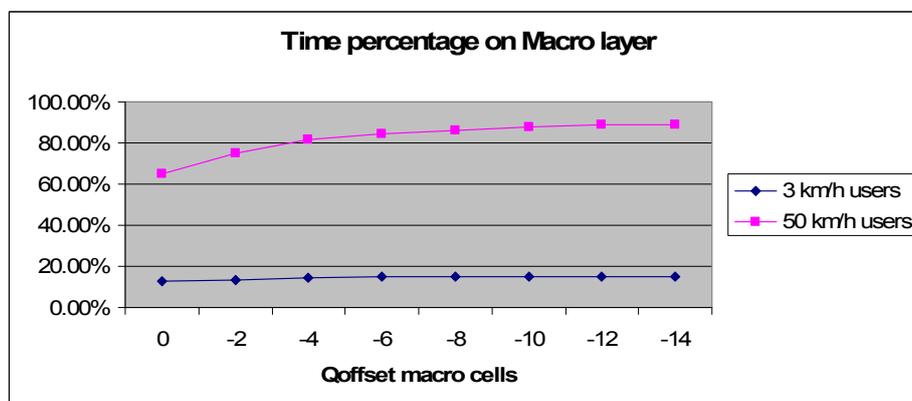


Figure 31 Percentage of time of stay on the Macro layer – Micro with tri-sectorial antennas

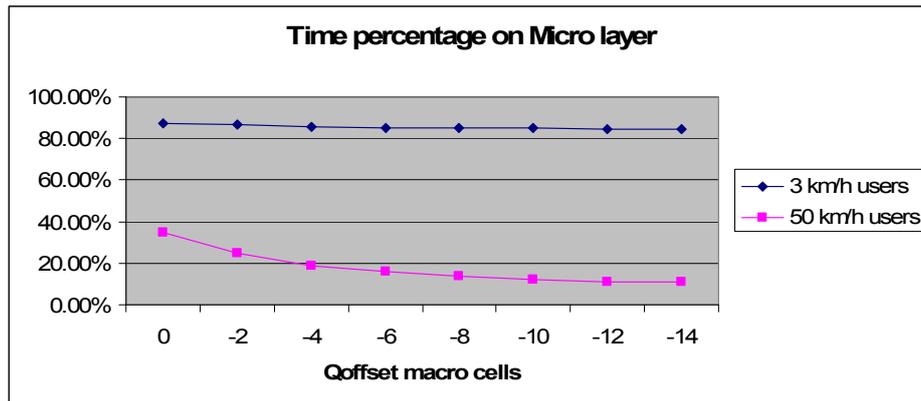


Figure 32 Percentage of time of stay on the Micro layer – Micro with tri-sectorial antennas

In Figure 33 and Figure 34 the time between two consecutive cell-reselections versus the Qoffset values of macro cells has been depicted, respectively for high-mobility users and low-mobility users. It should be noted that, as in the omni-directional layout, decreasing the values of Qoffset, the time between cell-reselections decreases.

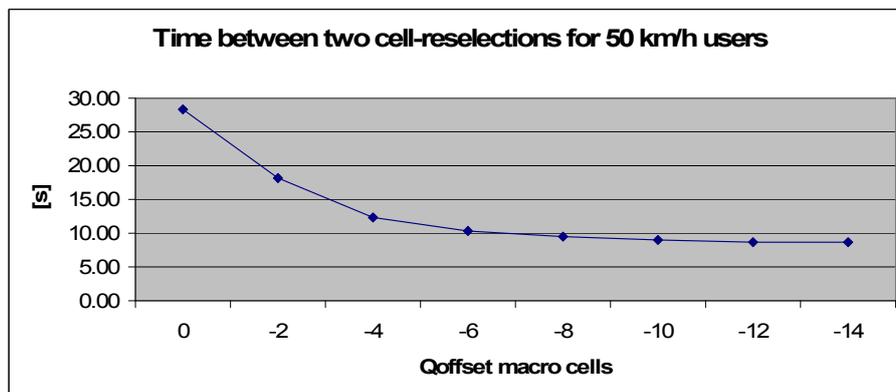


Figure 33 Time between two consecutive cell-reselections for high-mobility users – Micro with tri-sectorial antennas

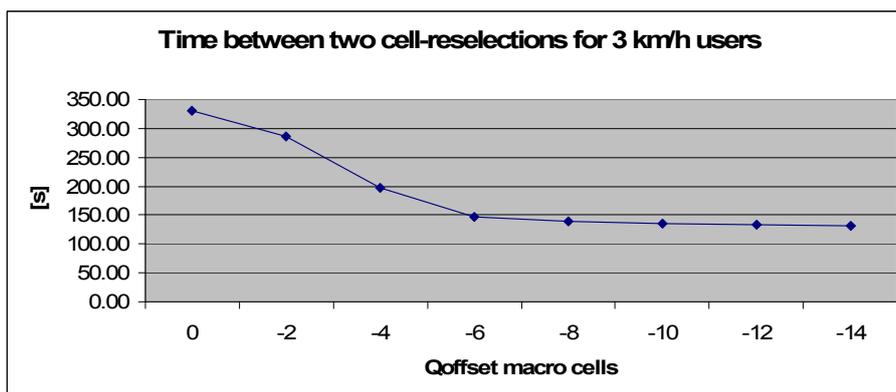


Figure 34 Time between two consecutive cell-reselection for low-mobility users – Micro with tri-sectorial antennas

Also in this case, the evaluation of the percentage of time in which the terminals are in the high-mobility state or in the low-mobility state shows that the values chosen for the mobility-class evaluation parameters are good, as shown in next figure: more than 99% of the high-mobility users are always in fast-mobility state, while less than 4% of the low-mobility users is

in the fast-mobility state. It has to be noted that in this case the number of micro cells is really bigger than in the case of microcellular layer with omni-directional antennas. Then, the number of cell-reselections performed by the low-mobility users is higher in this case than in the other case with omni-directional antennas. Hence, the time between two consecutive cell-reselections is lower in this case than in the other and the percentage of low-mobility users in fast-mobility state is higher.

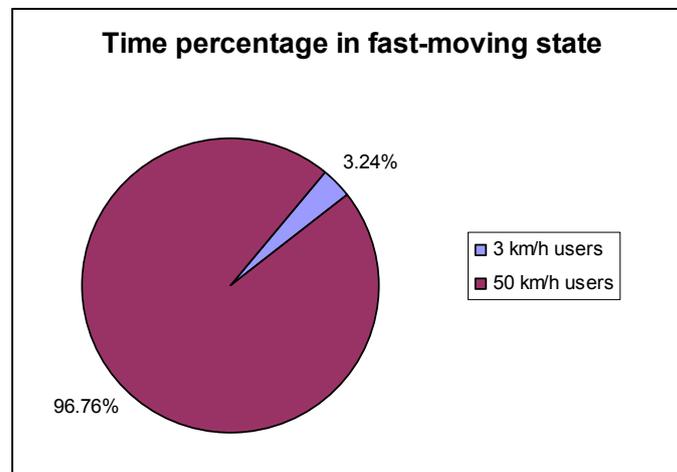


Figure 35 Percentage of time of staying in the fast-mobility state for both the simulated mobility classes of the users

2.9.3.3 Capacity analysis

The capacity analysis reported in this section has been carried out using the HCS layout with omni-directional antennas in the microcellular layer.

In order to maintain the users segregation also in connected-mode, when the call is present the terminal is admitted to measure only the cells belonging to the same layer of its serving cells, i.e. if the terminal starts a connection when it is camped on a macro-cellular cell, during the call it will perform measurements only of macro-cellular neighbor cells. This situation will lead to a lower number of Active Set Update in connected mode than the number of cell-reselections in idle-mode.

The evaluations have been performed with three different values of the Qoffset parameters of macro cells: 0 dB, -6 dB and -14 dB.

In Figure 36 the traffic results have been reported, considering:

- Input load: the traffic offered to the system
- System load: the traffic carried by the system
- Blocked load: the blocked traffic (blocked when entering in the system)
- Dropped load: the dropped traffic (during the call).

It has to be noted that, decreasing the value of Qoffset, the blocked and dropped loads increase and then the system load decreases. The same phenomenon could be seen analyzing the blocking percentage in case of new RAB setup, as depicted in Figure 65: both the uplink and the downlink blocking percentages increase with the increase of the segregation of the users on the layers according to their mobility class. As explained in the previous section, decreasing the value of Qoffset means to move users from the micro layer to the macro layer, above all the high-speed users are moved from the micro to the macro layer. The segregation of users leads to increase the "near-far" effect: in fact, considering a user camped on a macro cell, he could be closer to a micro cell antenna than a user camped on this micro cell, thus increasing the uplink interference. Furthermore, in downlink a similar

situation is present: a user camped on a macro cell could be closer to a micro cell antenna than the antenna of its macro cell, thus increasing the downlink interference. Both this phenomena have been depicted in a simplified way in Figure 38 and Figure 39. The main simplification is that in these figures it has been supposed that the problem is only related to the distance between terminals and antennas, neglecting the different transmitted powers used in micro and macro layers. In Figure 40 the average power transmitted by the terminal is depicted, taking into account the different mobility classes: as a consequence of the traffic segregation, the power transmitted by the high-mobility terminals increases, while the power transmitted by the low-mobility terminals remains about stable. In fact, since the traffic segregation moves above all the high-mobility users, the probability to have a situation of near-far effect increases above all for users with high speed.

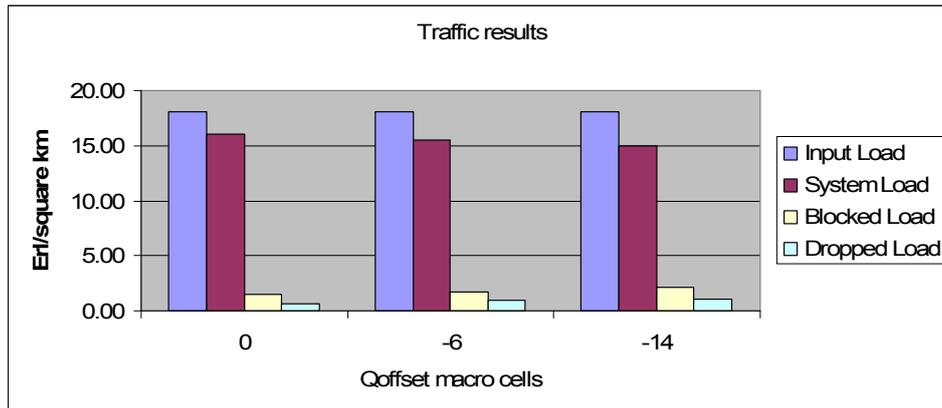


Figure 36 Input load, System load, Blocked load and Dropped load results

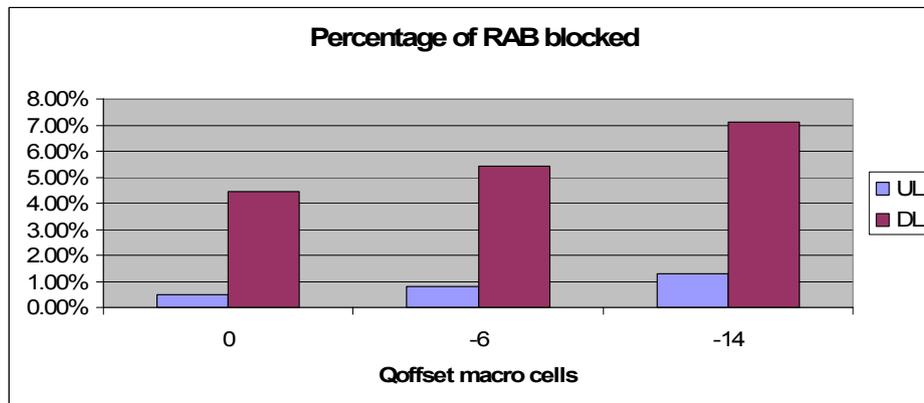


Figure 37 RAB blocking percentages due to uplink and downlink congestion

Uplink interference in a multi-layer layout

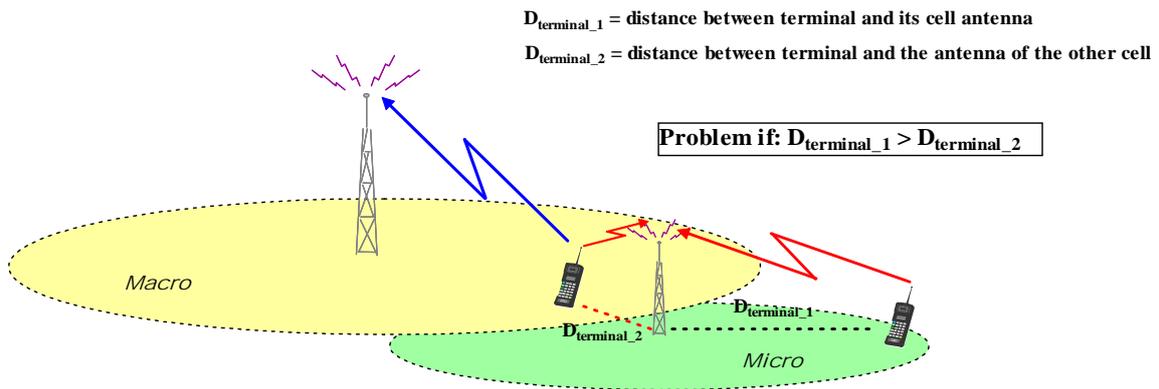


Figure 38 Simplified representation of the uplink interference problem in a multi-layer layout

Downlink interference in a multi-layer layout

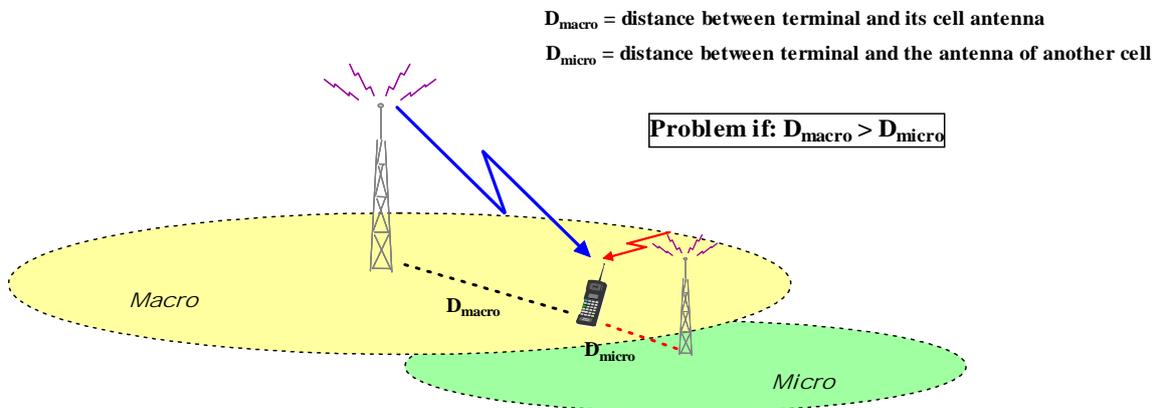


Figure 39 Simplified representation of the downlink interference problem in a multi-layer layout

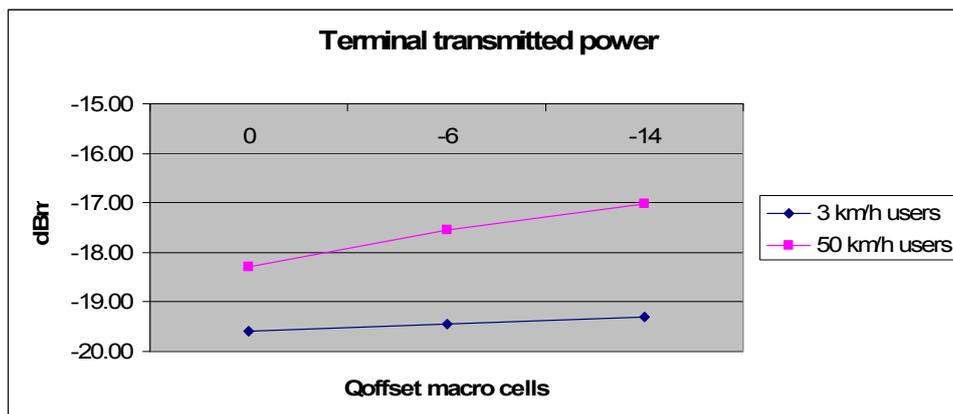


Figure 40 UE transmitted power according to the mobility class of the terminal

Summarizing, the capacity analysis in the presented HCS layout with the segregation of traffic between layers points out an increase of the blocking probability and then of the blocked traffic. It has to be noted that adopting the traffic segregation could be useful for an operator. In fact, there is a trade-off of the system performance, between the blocked traffic and the amount of signaling. In the performed simulations it has been analyzed the amount of signaling for Active Set Updates in the case without segregation of traffic and in the case with segregation. As reported in next figure, the number of Active Set Updates in the system

decreases while the value of Qoffset of the macro cells decreases, that is the amount of signaling for ASU decreases while the traffic segregation increases. Moreover, in the micro cells this behavior is linear decreasing the Qoffset values, while in the macro cells the number of ASUs remains approximately the same. This result is important from an operator point of view, since reducing the signaling leads to have a more efficient network and then to low the operational costs.

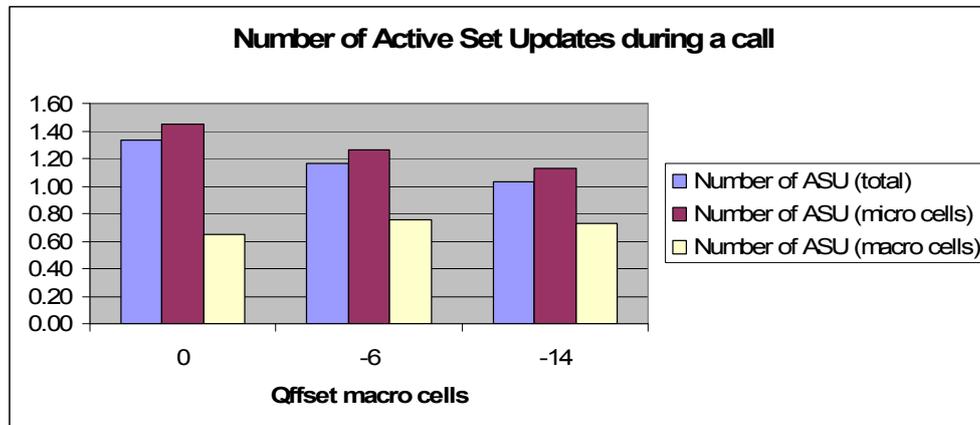


Figure 41 Number of Active Set Updates during a voice call in the different groups of cells

2.9.3.4 Conclusions

In this chapter some dynamic evaluations on Hierarchical Cell Structure (HCS), taking into account macro and micro layers, have been reported. The activity's goals have been to identify how to set the thresholds that control the algorithm applied by the terminal for the identification of the high/low mobility state by means of the number of cell reselections performed in a fixed period of time.

The results of this analysis highlight that a network operator could manage the settings of a plethora of parameters in order to segregate the traffic according to its necessities. In the situation described in this document, the segregation has been based on the mobility class of the users. In addition, it has been also showed that an operator may set the trade-off between the negative effects of the segregation (e.g. impacts on the interference level due to the increase of the "near-far" effect) and the benefits of it, according to its objectives. The decision of which solution has to be used is left to be assessed case by case.

2.9.4 Application of the derivatives framework to HCS

The complete frequency reuse in WCDMA simplifies in the frequency domain the planning exercise. Nevertheless, in the usual case that more than one WCDMA carrier is available for a given operator, frequency assignment to cells also plays a key role.

A typical radio network rollout in early deployment phases is based in outdoor macrocells and a progressive increase in the number of sites, in order to provide the desired coverage area as well as sufficient network capacity for the traffic demand. With increasing capacity demand another carrier may be added to the macrocell layer. Further, non homogeneous traffic distributions and non-homogeneous mobility patterns are eventually targeted by deploying microcells in high traffic areas, leading to hierarchical cell structures (HCS). In that sense, it is usual to operate the different cell layers with different carrier frequencies, although depending on the interference levels, this condition may be broken [21]. Also in [21] the planning of HCS with a single carrier is addressed.

One key issue in wireless cellular networks is that traffic distribution along time and space is inherently dynamic and subject to sudden and unexpected changes. Then, it is fairly common that real traffic distributions are substantially different from those considered in the planning phase.

In order to cope with these situations, dynamic planning mechanisms can be envisaged, so that frequency assignments to cells may vary along time and space. Clearly, from a practical implementation point of view, the adjective “dynamic” here stands for a rather long-term change (i.e. once or twice in a day, several times in a week, etc.), so that dynamic planning mechanisms will be able to cope with sudden but, at the same time, significant and long lasting traffic variations. Furthermore, since the number of carriers per operator in WCDMA systems uses to be rather low, coping with dynamic planning seems to be feasible from a practical point of view.

Traffic variations may be of different nature. One type is traffic variation on the average traffic level (for example, an average traffic increase because e.g. tickets for a unique concert are starting to be sold in a given shop at a given time and day). Another possibility is a variation on the spatial traffic distribution while keeping the average traffic level (for example, an entrance to an underground station is closed for maintenance for some days, so that people needs to get into the underground through a different entrance).

In this section, the analytical framework described in [22][23] is used as the basis for a dynamic planning algorithm. This framework is based on the computation of the derivatives of the load factors among the different cells. The main characteristic of the proposed framework is that, in contrast to other alternative methodologies, it is able to detect and capture all types of traffic variations (i.e. either on traffic averages values or on spatial distribution).

2.9.4.1 Frequency allocation schemes

In the following, a method is proposed to detect the cells having the highest influence over the rest of the cells in the scenario, according to the cell load factor derivative computation, thus being able to decide which cells should operate with a different carrier frequency, whenever a carrier frequency should be changed in the scenario.

In particular, the following scheme, denoted as derivatives-based algorithm, is proposed. It selects the base station to change as the one having the highest influence over the rest of the cells in the scenario.

For comparison purposes, let also define the so-called load-based algorithm, which changes the carrier frequency of the base station experiencing the highest uplink load factor.

Case study 1: In this case, there is a single hot spot located close to the microcell BS0.

This section discusses some of the results obtained in the analysed scenarios. In particular, focusing on the case study 1, Figure 42 presents the histogram of the base station that is selected by the frequency planning algorithm in order to change the carrier, obtained in the different snapshots. It can be observed how the selection is very different for both algorithms: while the load-based algorithm selects most of the time the microcell BS0, which has the highest load factor, the selection of the derivative-based algorithm is done mainly between BS0 and BS2, depending on how the users of the hot spot are distributed with respect to the other base stations.

In terms of performance, Table 13 provides the outage in the scenario when the carrier is the same and when one carrier is changed according to the two algorithms. Notice that the

reduction achieved with the derivative-based algorithm is much higher than the reduction obtained with the load-based algorithm. For more details on the considered case studies the reader is referred to [17], section 2.1.1.

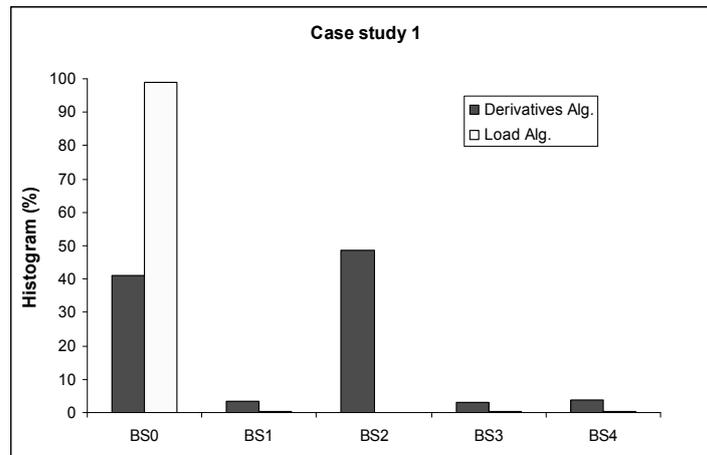


Figure 42 Histogram of the selected BS in case study 1

Table 13 Outage probability in the different scenarios

	Equal Freq.	Deriv. Alg.	Load. Alg.
Case 1	2.52 %	1.06 %	1.62 %
Case 2	28.92 %	18.28 %	27.99 %
Case 3	6.89 %	3.45 %	4.87 %

2.9.5 Non-Real Time Packet Transmission for a Microcell (Hotspot) Embedded in CDMA Macrocell Systems

The objective of this study is to obtain extra capacity for W-CDMA systems by adding hotspot base stations to an already present continuous macrocell layer as shown in Figure 43 with using the same frequency band in both hotspot and macrocell. The idea here is to take advantage of the time-varying macrocell interference profile, using adaptive Link Quality Control (LQC) in MAC layer to send and receive as many packets as possible from the hotspot bs, whenever macrocell interference allows. In the proposed scheme, as the users in macrocell always have the first priority to be served, the impact of hotspot on the macrocell performance has been reduced to its minimum or zero. With the emergence of asymmetric wireless data services, the downlink is widely regarded as the 'bottleneck' in providing these services. Furthermore, the demand for such asymmetric traffic is forecast to increase. Therefore this study focuses on the downlink. We show how, by controlling the transmitted power at the hotspot bs, the effect of the hotspot on the macrocell can be minimised, and how the packet transmission makes it possible to offer non-real time services at the hotspot, without sacrificing the macrocell performance.

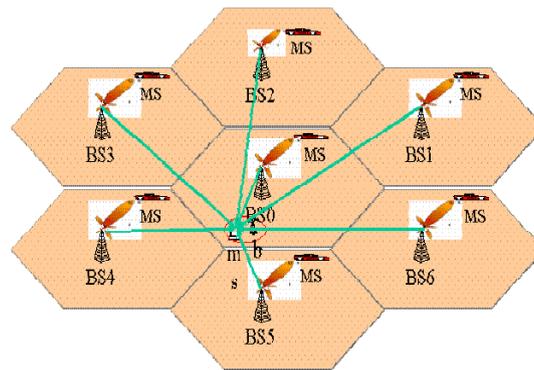


Figure 43 HCS Scenario

Based on the temporal situation of macrocell layer (MS’s locations, power control, shadowing, etc.) and related position of bs to BS, the time varying interference from macrocell to hotspot is estimated for each packet time interval. The maximum allowable bs power for that time interval is also dictated to the hotspot bs, in order to guarantee that the QoS of MS’s do not suffer. Then, the bs makes a decision to allocate the power to each mobile station at hotspot (ms), based on the “macrocell to hotspot interference” and “maximum allowable bs power”, which come from the macrocell layer, and “queuing policy priority” and “scheduling method”, which come from hotspot requirements, to maximise the hotspot throughput. Figure 44 shows the interaction between layers. Detailed algorithm implementation please refer as to [17] section 2.1.2.

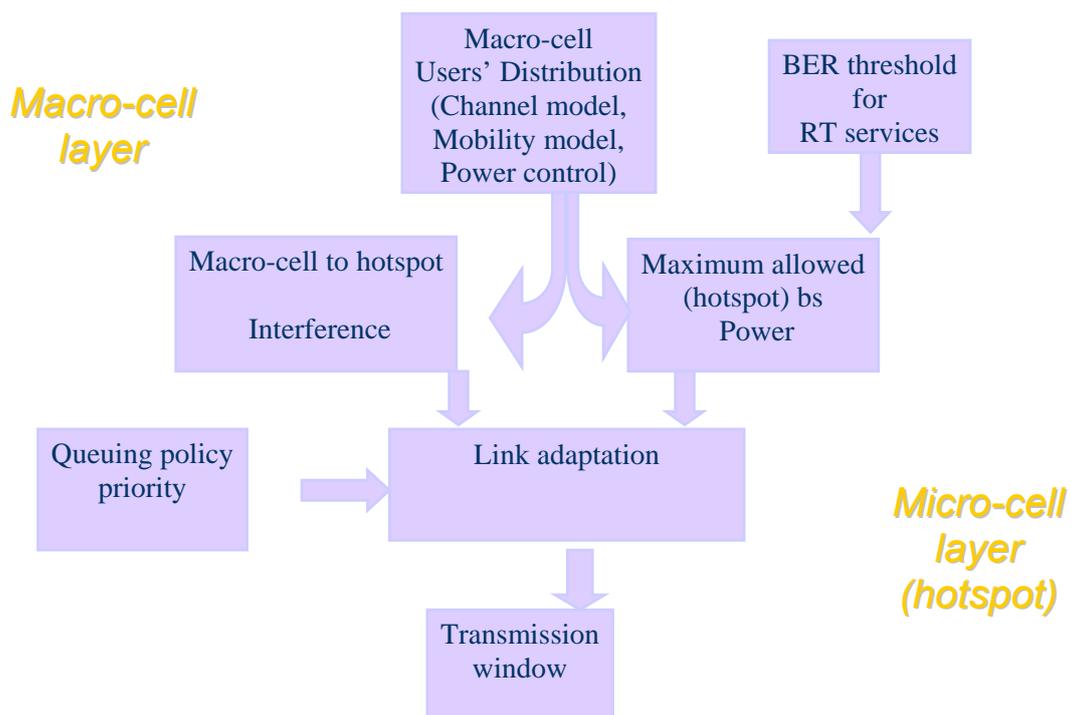


Figure 44 Interaction between layers

2.9.5.1 Results and dicussion

In our evaluation, assuming that the cross-layer interference is the dominant component of interference on hotspot (in comparison with intra and inter cell interference and noise), the data rate for the m^{th} hotspot user is a function of maximum allowable bs power and macrocell interference, which are both changing with time:

$$R_m(t) = \frac{W}{(E_b/(I_0 + N_0))_m} \cdot \frac{\phi_m}{L_m^t} \cdot \frac{P_{\max}^h(t)}{I_m^{\text{Inter-layer}}(t)} \quad (10)$$

This time-varying affordable rate means that by scheduling in hotspot bs, non-real time services can be delivered to hotspot users.

As a first order approximation, and assuming that the inter-layer interference is the dominant component of interference on the hotspot and $\phi_m = 1$, the mean of the affordable data rate at the hotspot is used to examine the hotspot performance on the downlink (see (10)). A typical example of this affordable data rate is shown in Figure 45. Based on(10), the desired $E_b/(I_0+N_0)$ for MS's has a major impact on the affordable data rate at the hotspot. This is illustrated in Table 14. This reduction of minimum acceptable $E_b/(I_0+N_0)$ can be translated to using different techniques such as multi-user detection, advanced coding and various diversity methods, at macrocell layer. This observation is proven through our hotspot simulation results in Figure 46. The figure shows the delay as a function of the number of WWW links is illustrated for different MS desired $E_b/(I_0+N_0)$, which can shown the extra capacity gain in terms of the number of WWW links, e.g with a delay of 2s for one http page, 7, 12, and 12 WWW links can be achieved corresponding to the MS desired $E_b/(I_0+N_0)$ of 7dB, 5dB and 4dB respectively.

Table 14 Averaged affordable data rate at hotspot

Required $E_b/(I_0+N_0)$ for MSs [dB]	Affordable data rate at hotspot (averaged) [kbps]
7	284
5.5	460
4	601

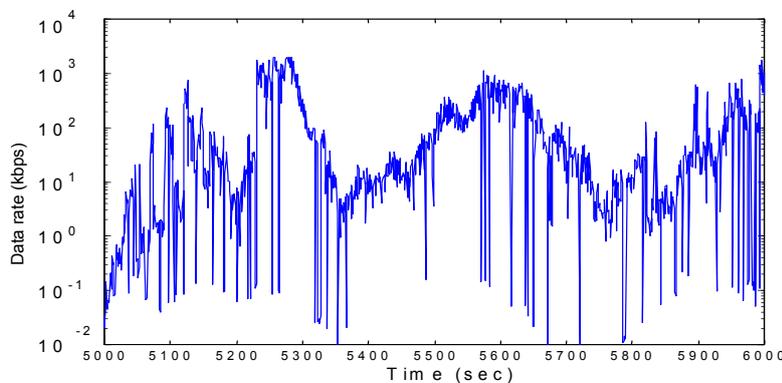


Figure 45 Affordable data rate at hotspot, as a function of time

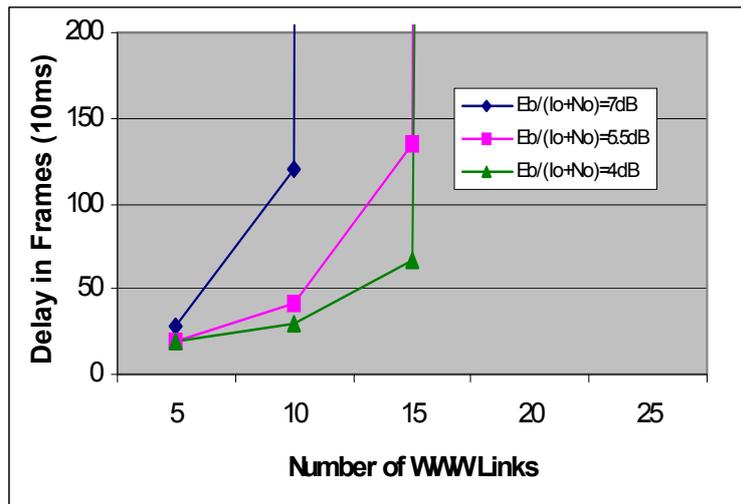


Figure 46 Delay as a function of WWW user numbers, for different required $E_b/(I_o+N_o)$, result from hotspot simulation

2.10 TRANSPORT CHANNEL TYPE SWITCHING

2.10.1 Introduction

This section summarizes the work carried out with the aim to investigate the system-level performances of UTRAN when a Transport Channel Type Switching (TCTS) algorithm is taken into account.

Within the context of the RRM strategies for the UMTS system, TCTS algorithms can have a very important role when users request VBR (Variable Bit Rate) services. This is the case, for example, of the World Wide Web browsing: this service implies a very discontinuous data transfer due to reasons like the reading time spent by the user to analyze the WEB page or the specific mechanism for requesting to the host a specific WEB page.

In the case of a discontinuous data transfer, it is very important to be able to optimize the usage of dedicated transport channel (DCH), preventing channelization code shortage in the downlink without degrading the end-to-end quality of service experienced by the user in an appreciable manner.

2.10.2 Bidirectional traffic model for the WWW service

The WWW bidirectional model used in the simulations has been implemented starting from what is specified in ETSI 30.03 for the non-real time service models. According to this statistical model, a typical WWW *browsing session* consists of a sequence of *packet calls*, which corresponds to the downloading of a single web page. Each packet call collects a burst of *packets*, constituting the objects of the page, like images, applet, banner or pop-up.

When the download is completed, the user consumes some time for reading the information, called *reading time*.

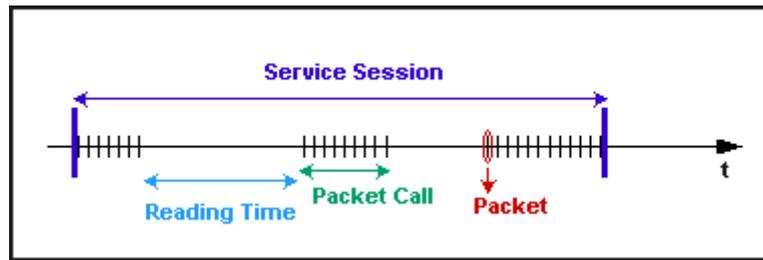


Figure 47 Characteristics of a WWW browsing session

In the case for a file transfer by means FTP protocol, the session contains only one packet call.

The typical model [30] is characterized by the following parameters:

- *Session arrival process*: it is a Poisson distribution process used to model the average number of session occurring within a given time interval, by means the input parameter. □The session is also characterized by the average session dimension (byte); consequentially the session length is known by means RAB bit-rate. These parameters define the offered traffic.
- *Number of packet calls per session*: it is a random variable with geometric distribution with a mean N_{pc} .
- *Reading time T_r* : it is a random variable with negative exponential distribution that indicates the time during the user read the information. The reading time starts when the download is completed and ends when the user sends a new request.
- *The inter-arrival time between two packets inside a packet call*: It is a geometrically distributed random variable.
- *Number of packets within a packet call*: it is a random variable with geometric distribution with a mean N_d .
- *Packet size*: it is a random variable with Pareto distribution with cut-off on a maximum value M .

$$\begin{cases} f_x(p) = \frac{\alpha k^\alpha}{p^{\alpha+1}} & k \leq p < M \\ \beta & p = M \end{cases} \quad (11)$$

Hence the packet dimension is defined by $S_p = \min(P, M)$, where P is the Pareto distribution variable and β is the probability that $p > M$.

It is worth to note that the model described so far is able to characterize in a statistical way the data transfer in the downlink. The model assumes implicitly that the network is able to carry out a certain amount of data per second (on average) and, consequently, it is possible to set a mean value for the inter-arrival time between two packets inside a packet call. Differently from the typical model, the implemented bidirectional WWW model used during the simulations, does not explicitly assume a fixed mean value for the packet inter-arrival time, because this quantity is implicitly defined by means of the behavior of the HTTP protocol, which is the mechanism used to transport Web documents.

According to this, the implemented model is able to capture the main behavior of the HTTP protocol: when the user (client) sends a fixed length request message (GET) to the server and when the server receives it (after T_{req} seconds), it sends back the data to the client (T_{send}), as depicted in the following figure.

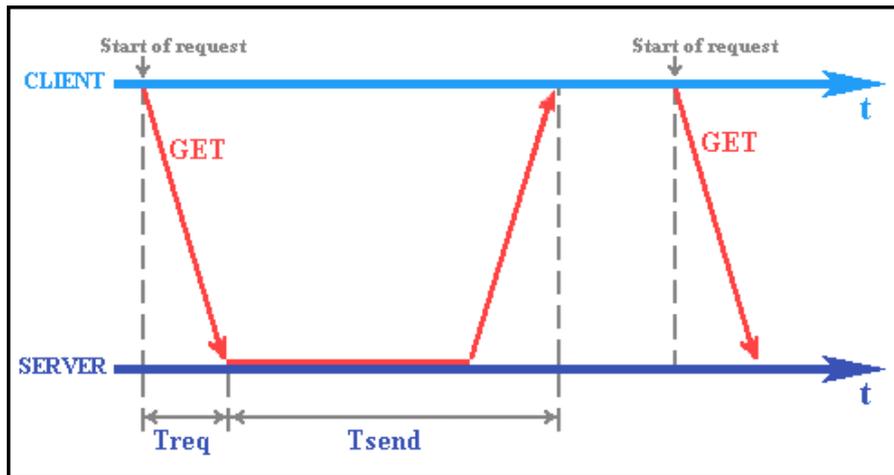


Figure 48 HTTP messages between client and server.

When a client starts a browsing session, sending the first GET REQUEST, a single connection to the server has established and only the page's body (html file) has retrieved. A conventional web page contains several objects: images, links to other files, applet, pop-up, banner, and so on. To retrieve these objects the client sends other *GET* requests, but, at this time, it is possible to improve the bandwidth by opening multiple TCP connections; each connection can be persistent, that is can manage different request/response. In this way, the inter-arrival between two consecutive packet transmissions within a packet call depends on the load conditions of the network and the characterization of the traffic in the uplink is also derived².

More in detail, the implemented traffic model is consistent with the release 1.1 of the HTTP protocol and consents to take into account several parameters to manage the methods to improve the protocol performance:

- *Multiple TCP connection*: the same client could open multiple connections to the same server to improve response time performance, but this method creates additional network load.
- *Persistent connections*: which reduces the amount of redundant information that is transmitted, by using the same TCP connection to send and receive multiple HTTP requests/responses, instead to opening a new one for every single request/response pair. Using persistent connections is very important for improving HTTP performance. The model consents to set if using the persistent connections.
- *Document compression*: by considering that several documents transferred are text based, an improvement to compress the file has been introduced. This consents to conserve the bandwidth, consuming only a few milliseconds of CPU-time. The input parameter "Compression Factor" is available to manage this feature.
- *GET time out*: If the response to a GET request is retrieved over this time, a new GET request is sent.
- *GET dimension*: this is the dimension in byte of the GET request that the client sends in uplink to the server.

According to the model, also a very accurate characterization of the internal structure of the available WEB pages is needed, in terms of overall size and number of objects inside each page. For a proper setting of these parameters it has been considered an analysis of the actual composition of the web pages and the recommendations of the W3C Consortium.

² The correlation between the data traffic in uplink and downlink is obtained by means of the characterization of the HTTP protocol behavior.

2.10.3 Traffic measurements and TCTS algorithm

The Transport Channel Type Switching (TCTS) procedure is a RRC procedure which is used to optimize the radio resources usage for data links in connected mode. The general idea behind this procedure is that, during the data connection, different types of transport channels, respectively dedicated and common ones, may be used in a switched mode according to the traffic volume. When the traffic is high, a dedicated transport channel (DCH) is to be used, whereas for small traffic values, common transport channels will be enough (in the CELL_FACH state, the user takes advantage of the RACH channel in the uplink and the FACH channel in the downlink). Of course, the switching between the two types of channel is not free of charge, so it should be done only if the system is confident that the traffic volume is not going to change for a reasonable long time.

The state diagram of RRC TCTS procedure is depicted in Figure 49: When a data connection is established between a User Equipment and the UTRAN, it is possible to switch from the dedicated channel RRC state (i.e. CELL_DCH state) to the common one (i.e. CELL_FACH state), or viceversa.

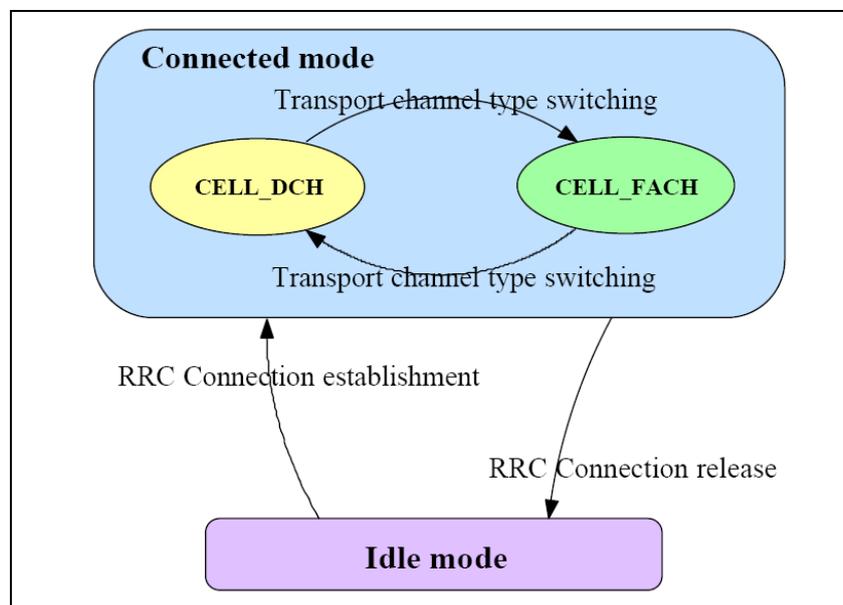


Figure 49 RRC states [11]

The criteria (i.e. the transport channel switching algorithm) which is used to decide when it is time to switch, clearly depends on different network implementations, but of course the measurements of traffic volume made by User Equipment or by UTRAN are the most important parameters for the definition of the above mentioned criteria. In a typical situation, these criteria may be expressed as follows:

- When the data volume which is to be sent is OVER a predefined threshold, dedicated channels (DCHs) are chosen;
- When the data volume which is to be sent is UNDER a predefined threshold, common channels (FACH for downlink and RACH for uplink) are chosen;

The two switching thresholds may be of course quite different; typically the first threshold may also be set to zero, which means that the switch procedure is started as soon as data has to be transmitted. This does not mean that these transitions are immediate but some delays may be introduced (in both directions) in order to guarantee that a switch procedure is done only when the traffic conditions really change for a “long” period of time (“long” in this context means comparable to the introduced delay). The delay introduced when the channel switches from dedicated channel to common channel (because there are no data to be

transmitted) is managed by the so called “inactivity timer” and this is a parameter which may be configured by the network operator in order to optimize the network performance.

Going into details, in connected mode, the switching algorithm between the two RRC states depicted in Figure 49 may be described by the 4-state diagram depicted in Figure 50 keeping into account the fact that either the uplink or the downlink Transmission Volume may be over a certain threshold. As a consequence, the CELL_DCH RRC state is splitted in 3 different states. The transition between these different states of the transport channel type switching algorithm depends on two different events which are defined as follows:

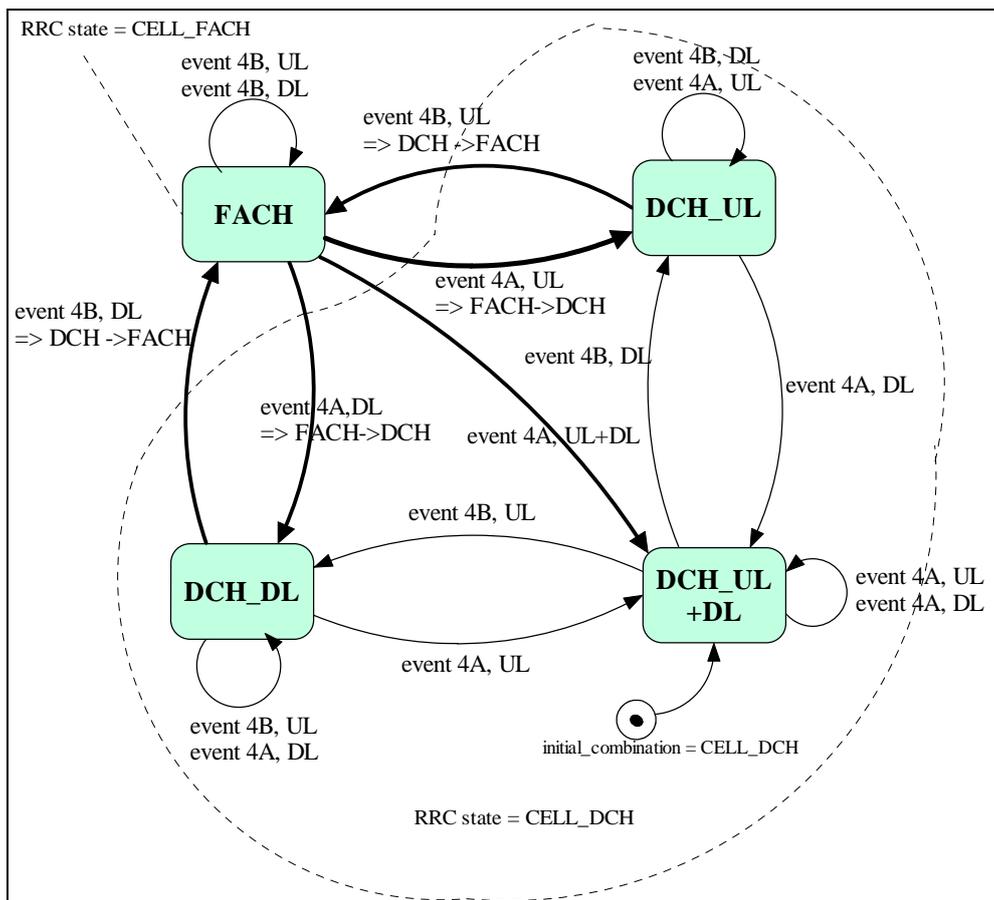


Figure 50 Transport channel type switching algorithm.

- 4a event: The traffic volume is OVER a defined threshold
- 4b event: The traffic volume is UNDER a defined threshold

From an abstract point of view, these events may concern either uplink or downlink direction and the corresponding channel transition is specified in Figure 50.

It is worth to note that it is not possible to use different types of transport channel for the uplink and the downlink. So the comparison between the buffer occupancies and the related thresholds must be done that way: to switch from shared to dedicated channel the buffers must be greater than an absolute threshold EITHER in uplink direction OR in the downlink direction, whereas to switch from dedicated to shared channel, the buffers must be lower than an absolute threshold BOTH in uplink direction AND in the downlink direction.

In the uplink case, according to the 3GPP 25.331 technical specifications, the User Equipment is in charge to notify the above described events to the RNC, so the 4a-type and the 4b-type events are reported according to the specified measurements report criteria.

The traffic volume for each transport channel is defined as the sum of all buffer occupancies for each logical channel which is multiplied on it:

$$TCTV = \sum_i^n BO_i \quad n = \text{number of Logical Channels multiplied on a Transport Channel}$$

It is not mandatory to have traffic measurements for each transport channel but a small subset of them may be defined as an input parameter of TCTS procedure (such as for example giving a list of transport channels for which some traffic measurements are defined). Of course, if NO measurements for some channel are foreseen, the corresponding switching function is inhibited.

Usually, with the context of the RRM framework belonging to each manufacturer, it is also possible to inhibit the switching function in an explicit way, even when measurements are done, in order e.g. to guarantee that specific data services always have a dedicated transport channel (e.g. the streaming data services).

The reporting events 4A and 4B are triggered each time that a fixed threshold is respectively reached from lower values of transport channel traffic volumes (4A event) or from higher values of transport channel traffic volumes (4A event). The two situations are depicted in the following figures

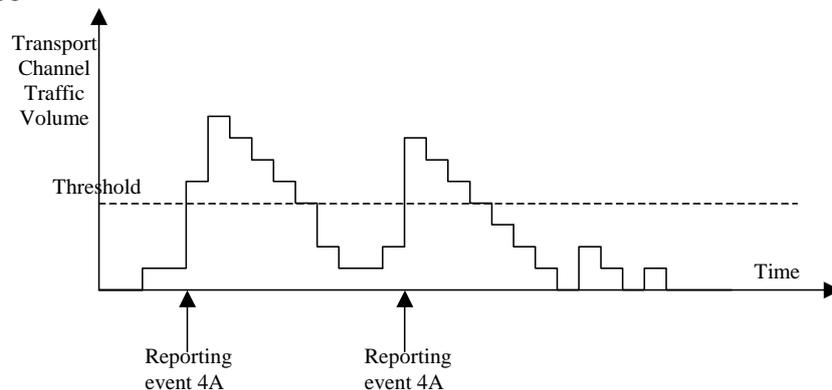


Figure 51 Event triggered report when Transport Channel Traffic Volume becomes larger than a certain threshold

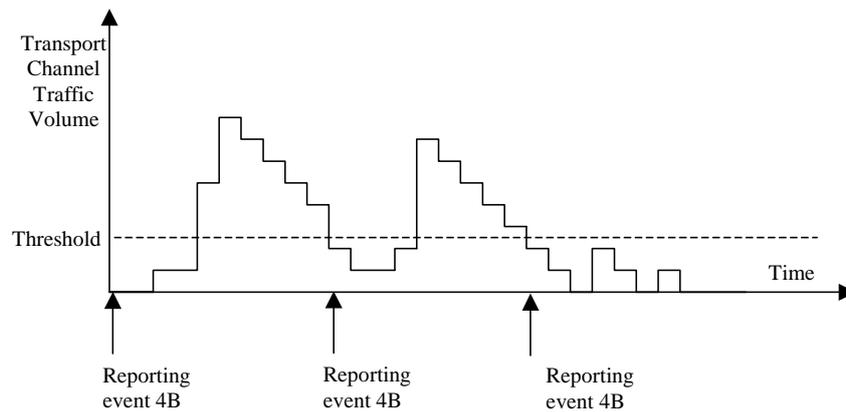


Figure 52 Event triggered report when Transport Channel Traffic Volume becomes smaller than certain threshold

A time to trigger parameter may be defined. In that case the timer is started in the User Equipment when the Transport Channel Traffic Volume triggers the event. If the Transport Channel Traffic Volume crosses the threshold before the timer expires, the timer is stopped. If the timer expires then a report is triggered, as it is shown in the following figure:

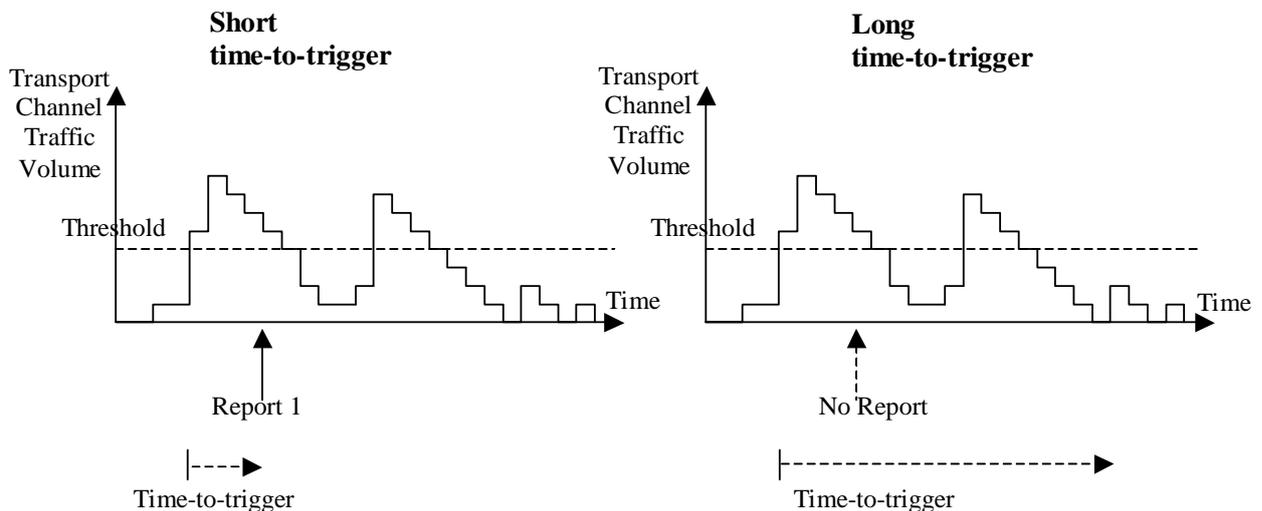


Figure 53 Time-to-trigger is used to achieve time hysteresis

2.10.4 Main simulation inputs

In the overall sets of carried out simulations, only PS WWW users have been considered: they are uniformly distributed in the entire scenario, already described in [1]. It has been also assumed that these users move at 3km/h.

The radio access bearer (RAB) for the user data transfer between UE and UTRAN, used in these simulations, belongs to the INTERACTIVE class and foresees a maximum bandwidth equal to 64 kbps both in uplink and downlink. It is worth to note that the RAB profile specifies a value only as far as the maximum bandwidth is concerned but it does not specify any value for the minimum bandwidth. This is in line with the fact that the WWW browsing service belongs to the INTERACTIVE class which is a non-real-time (NRT) type of service. According to this, this type of service can be managed also taking advantage of common transport channel transmission, when useful.

In the case of usage of dedicated transport channel, the spreading factor for the channelization code in downlink is equal to 32.

The carried out simulations are classifiable in two different groups:

1. "DCH-ONLY": the transport channel type switching algorithm has been inhibited by forcing the use of a dedicated transport channel, also in a low traffic conditions;
2. "TCTS active": the transport channel type switching is active, and two different settings were taken into account as far as the switching algorithm is concerned.

The overall set of simulations has been carried out in order to investigate the following main issues:

- to emphasize the differences on the system-level performances of UTRAN with respect to the use of DCH only;
- to optimize the input parameters of the TCTS algorithm, investigating the impacts of different configurations.

Table 15 and Table 16 show the setting values for some WWW traffic model parameters that are maintained invariant for both simulation groups.

Table 15 Input HTTP parameters

Input Parameter	Value
Multiple TCP connection	4
Persistent connections	YES
Compression Factor	1
GET time out	60 s.
GET dimension	350 byte

Table 16 Input WWW session parameters

Random Variable	Input Parameter
Session Inter Arrival Time	$\lambda^{-1} = \frac{PS \text{ Traffic per User}}{Mean \text{ Session Dimension}}$ [s.]
MeanSessionDim	636570 [bytes]
PS Traffic per User	1625 [byte/s]
Reading Time	$\lambda^{-1} = 1/12.7$ [s.]
Packet Dimension	M = 5120 [bytes] , k = 1024 [bytes] , $\alpha = 1.1$, $\mu = 2546.28$ [bytes]
Packet Calls Number	p = 1/10
Packets Number	p = 1/25

Other input parameters have been varied according to the aim of the simulation analysis, so they values have been reported in the specific paragraph.

The mean values for the packet dimension, number of packet calls and number of packets within a packet call have been set according to the results coming from statistics collected during the real operation of the Telecom Italia mobile network.

Also the value chosen for the reading time parameter is representative of a typical behavior of the users and it is such that the duty-cycle of the service is about 50%³.

³ Taking into account the RLC throughput offered by the RAB (64 kbps) and the mean web page size (63650 bytes), it can be assumed that the mean transfer delay for a web page is about 10 seconds.

The MeanSessionDim and PS_Traffic_per_User parameters have been set in such a way that each user produces 0.6 Erl-equivalent data traffic. As a consequence, 30 users per cell are able to produce about 600 kbit/s of traffic data.

2.10.5 Simulation results

According to what stated in 2.10.4, in the following section (ref. 2.10.5.1) the DCH-only family of simulations (where the transport channel type switching is not activated and every user has a dedicated channel) and the TCTS families of simulations (where the transport channel type switching is activated) are compared in order to underline the differences between the two cases in terms of the main aspects, like UTRAN capacity, interference, code occupancy, etc. The configuration set of the switching algorithm considered for this type of analysis is reported in Table 22 (row “Set 2”) and corresponds to the most optimized one, as described in details in section 2.10.5.2. In fact, in section 2.10.5.2, the two TCTS families of simulations (with different values of some input parameters, as reported in Table 22) are compared between them, investigating the most appropriate settings in order to maximize the performance of UTRAN cells.

2.10.5.1 DCH-only versus Transport Channel Type Switching comparison

2.10.5.1.1 Capacity analysis

Next figure reports the total amount of traffic carried by the radio access network (“traffic load”) in the two different cases, together with the corresponding amount of blocked traffic (“blocked load”) ⁴:

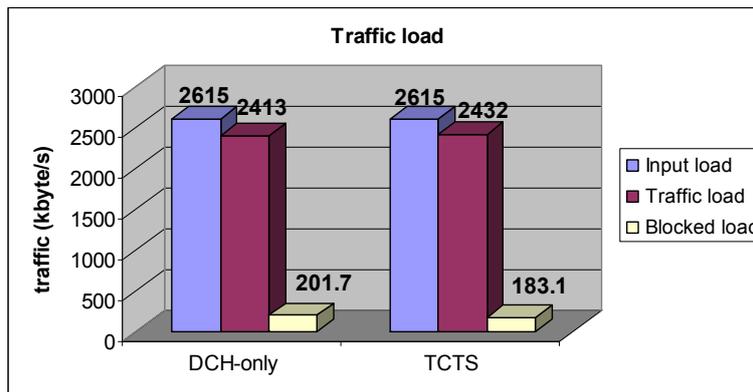


Figure 54 Traffic results (DCH-only versus TCTS)

The achieved results show that the UTRAN capacity is higher when the transport channel type switching algorithm is active; the following figure represent this fact in terms of percentage of traffic blocked with respect to the traffic offered in input by the users:

⁴ Taking into account that the users have been distributed uniformly in all the 36 cells of the simulated scenarios, the mean amount of traffic in input per each cell is approximately equal to 600 kbit/s in both cases, as reported in 3.4.

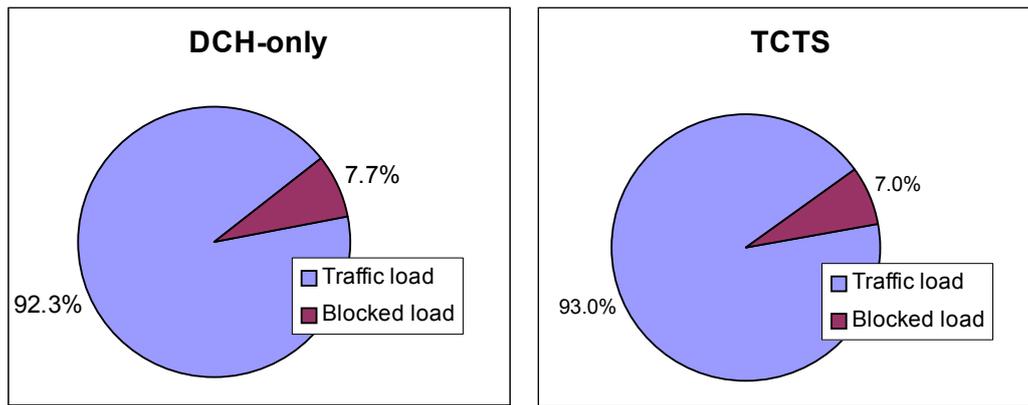


Figure 55 Percentage of blocked traffic (DCH-only versus TCTS)

As represented by Figure 55, using the transport channel type switching algorithm the percentage of blocked traffic decreases from 7.7% to 7.0% (-9.2%).

The capacity gain achieved by the switching algorithm can be justified taking into account the admission and congestion control procedures operated by UTRAN. With respect the admission control, the simulation results show that 3.8% of radio access bearer setup requests are blocked due to shortage of available downlink codes in the DCH-only case, whereas these events do not occur in the TCTS case, as represented in next figure:

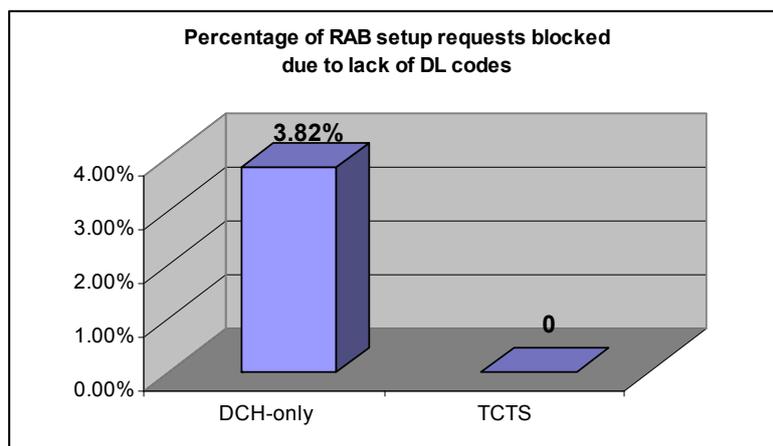


Figure 56 Percentage of RAB setup requests blocked due to lack of DL codes (DCH-only versus TCTS)

The difference in the achieved results between the two cases can be understood taking into account that in DCHonly mode, each DCH has its own specific OSVF code corresponding to an appropriate spreading factor assigned for the whole duration of the connection. If the number of users requiring DCHs becomes high, the code shortage cause the block of some users. As a matter of fact, the total number of codes is finite and depends on the data rate of RAB which is used. In the considered simulations, an ideal code allocation policy is taken into account, so that each cell can make available 32 codes with SF equals to 32 (this number depends on the assumption that 64 kbit/sec RAB are used uplink/downlink). Taking into account the codes allocated for the common channels (i.e. CPICH, SCH, AICH, Primary-CCPCH, Secondary-CPCCH), only 31 codes with SF equals to 32 are available for the traffic DCHs. According to this, the assignment of a DCH for the whole duration of the session, especially for packet data users with ON/OFF traffic, results in inefficient utilization of the resource. We must also keep in mind that if a user does not use the channel (for example during the reading time when no data must be downloaded) the code is anyway considered

as occupied and cannot be released and used by other users. Only when the session is over, the channel is released.

In accordance with the above considerations, the following table shows the difference between the two cases in terms of the mean number of contemporary radiolink active per cell:

Table 17 contemporary active radiolink per cell

RL_Capacity	DCH only	TCTS
TCP persistent	26,98642	13,80578
TCP volatile	25,29767	12,65389

We must consider that the input parameter for these simulations are fixed in such a way that “reading time” (users read the Internet pages which have been downloaded) is more or less equal to the “downloading time” (users download the pages to be read). The first parameter is directly given as a simulation parameter Table 16 whereas the second one derives from other input parameters (average length of a page expressed in packets, average length of a packet, etc.). This “duty cycle” (related to reading and downloading) nearly equal to 50% has so a great relevance in making a great difference between DCHonly and TCTS cases. For example if we consider the 30 users case, in DCHonly this number of users clearly saturates the capacity for RAB which is fixed, as mentioned before, in 31 channel for every cell. In that case the number of blocked RAB is quite great. If we use the TCTS, assuming a 50% duty cycle we may release half of the channels and use them for other users so that a virtually null value for this RAB blocking is shown.

To go into details, the following diagram shows the Radio Link capacity distribution in the two different cases (DCHonly and TCTS); the differences are due to the fact that with DCHonly we have a higher number of DCH which are active at the same time, so blocks are also caused by the lack of codes (the diagram also shows the percentage difference of DCH channels usage in both cases). In that case, the curve has a peak for 30 users and rapidly slows down because no more than 31 codes are available.

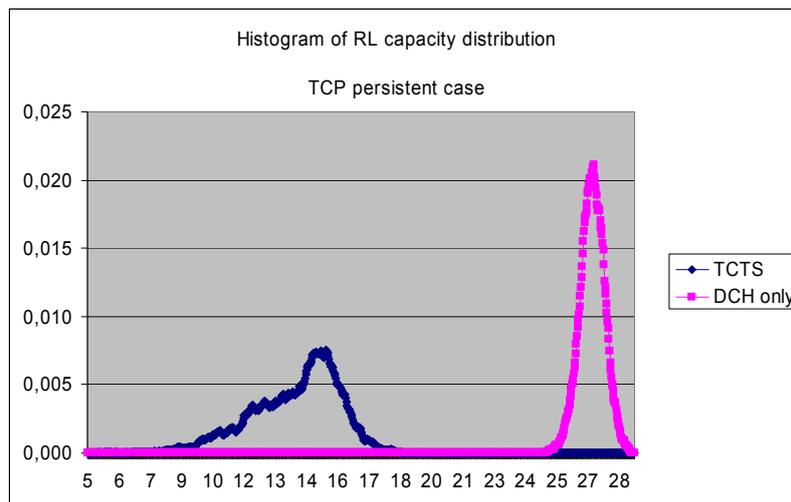


Figure 57 Radiolink capacity PDF.

If we have instead a look at the TCTS distribution of Radio Link Capacity we see a smoother curve which has a peak for 15 users (due to 50% duty cycle). The virtually null area of this curve which has more than 31 offered channels (with respect to the overall area under the curve) represents the percentage of blocked RAB as seen in previous graphs.

Another difference able to justify the capacity gain achieved in the TCTS case comes from the behavior of the congestion control algorithm that drops the radio access bearer when the level of total interference in uplink is above the threshold (8 dB). Figure 58 shows the percentage of RAB dropped due to uplink interference in the two cases:

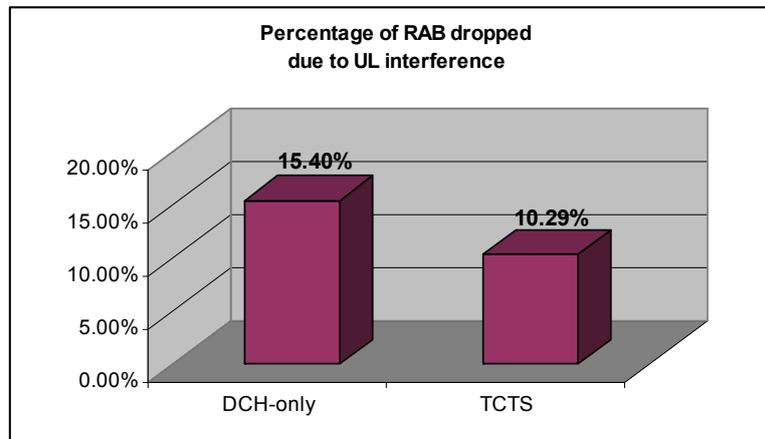


Figure 58 Percentage of RAB dropped due to UL interference (DCH-only versus TCTS)

In the DCH-only scenario, the percentage of RAB dropped is higher than the TCTS case and, as a consequence of this, a higher percentage of traffic is not carried by the network due to the session interruption. These results are in accordance with the values of RTWP reported in the section 2.10.5.1.2 that demonstrate a higher interference of the DCH-only scenario.

2.10.5.1.2 Uplink and downlink interference

In DCHonly mode a DCH channel is allocated for each user; more precisely a DTCH channel is used to transmit data whereas a DCCH channel is used for signalling purposes. The signalling messages are continuously exchanged between the network and the terminal in order to refresh the information which is used by the power control procedures. This means that some data (even if with a very low throughput) are anyway sent and received even during reading time. The advantage in this case is that the channel is always available and no setup procedures are needed; furthermore the power used to transmit is always the best possible choice due to power control mechanism, and this fact guarantees a very high QoS with respect to TCTS case. The effects on interference due to this channel will not be so evident when the channel is not busy because only signalling data is being transmitted but the forementioned reduction of capacity is an important issue to take into account.

A preliminary consideration concerns the fact that, when taking into account the uplink direction, the DCHonly simulations show a greater interference than the TCTS cases. As a matter of fact, in the first case all the users have a dedicated channel (which is always ON, also for simple signalling purposes) whereas in the second case the users share the RACH channel which is managed by a queue so the interference is reduced.

In fact the RACH acts as a bottleneck for every user which aims to use it, but there is an advantage of having a single point of concentration of uplink messages, because the overall interference is reduced. Moreover, the RACH channel is able to fix the most appropriate transmission power by means of the ramp-up procedure [39].

The following table shows also the value of Noise Rise and received total wideband power (RTWP) in DCHonly and TCTS cases:

Table 18 Values for NR and RTWP

UPLINK interference	TCP persistent	
	DCH only	TCTS
Noise Rise [dB]	6,77	6,07
Rx Total WB Power [dBm]	-98,38	-99,05

Viceversa, when we concentrate on downlink direction, the DCHonly case is less interfered than TCTS case because in the latter case, when the FACH channel is used, a fixed transmission power equal to -33 dBm (adjustable as an input parameter) is considered. On the contrary, when DCH is used, the transmission power is set to the most appropriate value by the power control procedure with a frequency 1500 Hz.

Table 19 Values for NodeB TX power and CPICH Ec/N0

DOWNLINK interference	TCP persistent	
	DCH only	TCTS
UMTS_NodeB_TX_power [dBm]	17,3848	19,239
UMTS_CPICH_Ec/N0 [dB]	-7,02097	-8,74176

Another parameter which may indicate a higher quality in DCHonly case (due to a reduced interference) with respect to the TCTS case is CPICH_Ec/N0. This is a direct way to measure interference because it represents the interference which is measured on the pilot channel. As we said before, both values are greater for TCTS case with respect to DCHonly case as we already estimated in advance.

2.10.5.1.3 QoS analysis

The simulation results reported in section 2.10.5.1.1 show the capacity gain of the TCTS case with respect the DCH-only case, specifying the reasons for what this event occurs. Besides this, it is easy to guess that a trade-off exists when the common transport channels are used instead of the DCH-only solution.

As a general rule, even if the common channels are used only when a very small amount of data (or nothing at all) have to be transferred, the QoS is better for DCH-only than for TCTS because of the switching time. DCH offers higher transfer speeds (throughput) but requires a significant setup time, whereas shared channels have a low throughput but also a low setup time. Due to these considerations the usage of common channels is more efficient only when the traffic is sporadic or for short time, otherwise for long and frequent data sessions, the usage of a dedicated channel is highly recommended.

In this section some simulations results very well correlated with the quality of the WWW service are reported and discussed.

2.10.5.1.4 Throughput and delay analysis

The following table and the corresponding figure show the mean throughput and the page download delay in the two observed scenarios:

Table 20 Mean values of throughput and page downlad delay (DCH-only versus TCTS case)

	mean throughput [kbit/s]	page download time [s]
DCH-only	49.4	9.9
TCTS	26.6	20.3

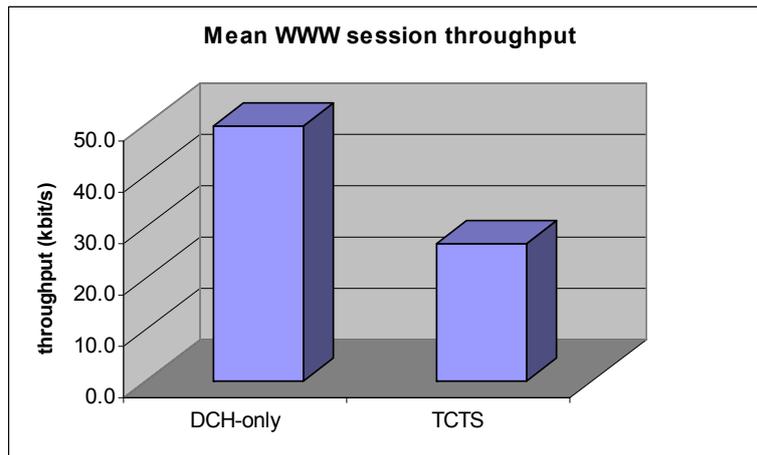


Figure 59 Mean values of throughput (DCH-only versus TCTS case)

Since the same traffic profile was considered in the two analyzed cases, the higher page download time perceived when TCTS is used, produces also a higher duration of the web browsing session: 239.7 seconds in DCH-only versus 396.6 in the TCTS case.

The above simulation results clearly point out that in the DCH-only case a higher throughput and consequently a higher level of quality of service perceived by the users is achieved. These results were obtained even if the usage of common channels in the TCTS case occurs only when users are not requesting any data (during the “reading time” period, Table 16), therefore the observed throughput degradation can be ascribed to the necessary time to switch from common to dedicated channels when users restart to download new data. Figure 60 shows the time requested by the transport channel reconfiguration procedure used by UTRAN:

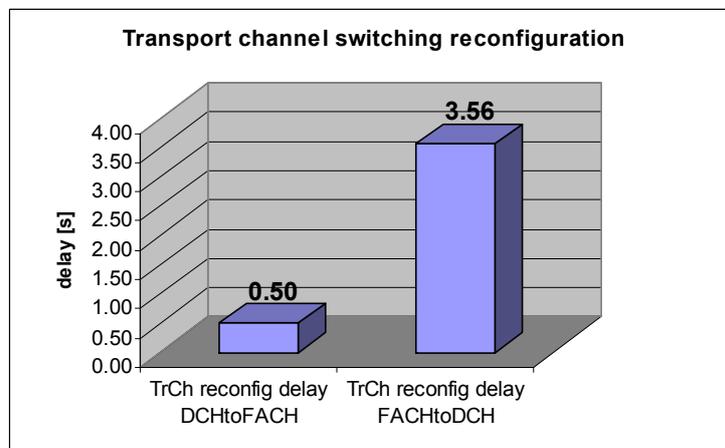


Figure 60 Performance of the transport channel type switching procedure.

Figure 60 point out the fact that the switch from common to dedicated channel requires about 3.5 seconds, reducing the mean throughput of the web browsing session. On the contrary, it is also worth to note that the switch from dedicated to common channel is much less time consuming, requiring only 0.5 seconds. This reason of this fact is that when common channels are used, the first RRC message of the switching procedure (“transport channel type switching request goes from the RNC to the users via the FACH common channel.

Besides the switching time, also other issues contribute to the decrease of throughput. For instance, the following table shows the RLC PDU retransmission rate performed by the RLC protocol in the two cases:

Table 21 Retransmission rate performed by RLC protocol (DCH-only versus TCTS)

	RLC PDU RTX uplink	RLC PDU RTX downlink
DCH-only	3.3%	10.6%
TCTS	5.7%	22.5%

The retransmission rates for the DCH-only case are the results of the C/I values chosen in uplink and downlink for the considered radio access bearer (ref. 2.10.4). In the TCTS case, the retransmission percentages increase: from 3% to 10% in uplink and from 6% to 22% in downlink. The observed increase is due to the queuing induced by the common channels that causes an increase in the transfer delay of the RLC PDUs⁵ with respect to the DCH-only case. In order to mitigate at least partially this effect, proper values of the parameters controlling the RLC retransmission frequency have to be set; in the TCTS simulations, the values of “Timer_Poll” and “Status_Periodic” parameters [11] were increased from 0.5 to 2 seconds.

2.10.5.1.5 Transport reliability analysis

Another very important element to take into account in the estimation of the level of quality of the web browsing service is related to the level of reliability offered by UTRAN cells.

Simulation statistics covering this aspect are:

- mean percentage of web pages downloaded (respect to the total number of web pages requested during each session);
- mean percentage of objects downloaded successfully within a page.

In fact, during the downloading phase of a web page, congestion can occur in the serving cell, resulting in a download break of the requested data. In order to catch these events, the appropriate timers between get request and response of the http protocol where simulated, as well as all the timers controlling the behavior of TCP protocol in case of data-transfer inactivity. Figure 61 shows the simulation results concerning the above mentioned two items:

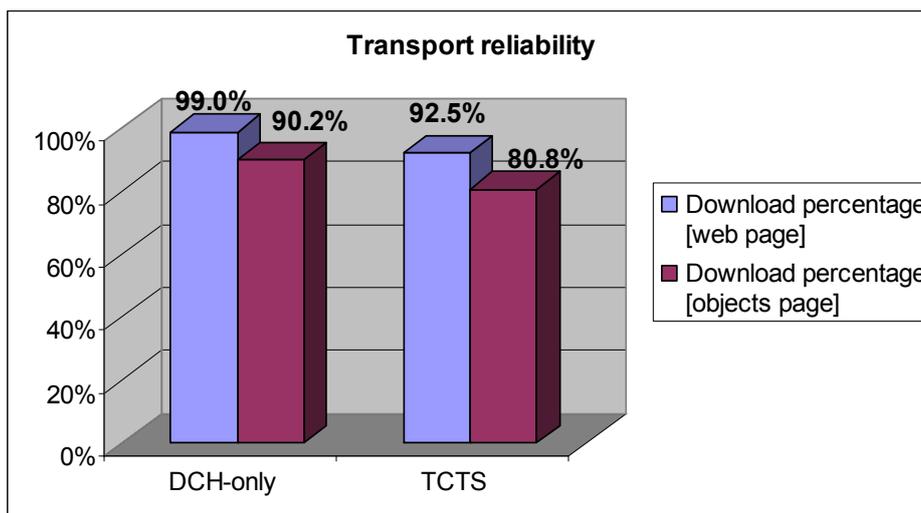


Figure 61 Download percentages (web pages and objects within a web page); DCH-only versus TCTS

⁵ When RACH/FACH are used in uplink/downlink, the transfer delay of a radio blocks is afflicted with the waiting time inside the buffers of these common transport channels.

From the analysis of the results summarized in Figure 61, it follows that in the TCTS case the download percentages are only slightly lower than the DCH-only case. Therefore, with respect to the level of QoS, it is possible to conclude that the main drawback coming from the TCTS usage is only the decrease of throughput with respect the DCH-only case.

2.10.5.2 TCTS active

In this section, we will investigate quantitative performance of the system with respect to some characteristic parameters of the channel switching algorithm. Since the channel switching is costly in time and in signaling and during the switching time, no packets can be transmitted, it is very important, in order to obtain improvements from the channel type change, set very well these parameters.

By setting correctly the TCTS parameters it may be possible to follow the ON/OFF characteristics of the WWW traffic, where the OFF time represents the time during which a user is reading or browsing a Web page, the ON time represents the time for the objects download.

In general terms, it may be desirable to offer a DCH to the user during objects download to guarantee a good throughput, and to switch to RACH/FACH during the inactivity period (reading time) to improve the radio interface capacity.

2.10.5.2.1 Input parameters

This simulation group considered in this section has been run with a low traffic load, that is only 10 users per cell have been set. The WWW traffic model parameters, shown in Table 16, create web sessions consisting on an average of 10 packet calls, each one with 25 packets of 2546.28 bytes of size, totally 63657 bytes. After each packet call there is averagely 12.7 seconds of reading time.

The radio access bearer (RAB) used in these simulations foresees a maximum bandwidth equal to 64 Kbit/s, but considering the protocol headers (TCP, IP, PDCP), remains nearly 44 Kbit/s. So to download a 63657 bytes packet call nearly 11.5 seconds are necessary.

Under these assumptions it is expected an On/Off traffic profile with nearly a 48% duty cycle, being:

$$\frac{time_{ON}}{time_{ON} + time_{OFF}} \cong 0.48 \quad (12)$$

Other parameters TCTS algorithm specific, have been varied in order to control the critical item before mentioned, like the signaling and the profile pursuit.

Table 22 reports the different simulations' sets, and shows how many parameters have been changed each run.

Table 22 "Set 1" and "Set 2" parameters.

Event	Parameter	Set 1	Set 2
4A event	Time to trigger event [sec]	0	0
	Pending time after trigger [sec]	0.25 →	1
	Reporting Threshold [byte]	512 →	256
4B event	Time to trigger event [sec]	0 →	1
	Pending time after trigger [sec]	0.25 →	5
	Reporting Threshold [byte]	8	8

2.10.5.2.2 Simulation results

First simulation has been run with “Set 1” TCTS parameters which have been chosen without particular care.

As a matter of fact the results analysis gives evidence of a non-optimized behavior under different aspects:

- the switching frequency, that does not consent to reply the WWW traffic course with respect of the duty cycle (nearly 48%);
- the signalling, that is too much repeated, being costly in time;

First aspect has been emphasized by the comparison between the DCH time and the FACH time, as shown in Figure 62 and by the comparison between the number of switching from DCH to FACH and vice versa with respect to the packet calls number (expected value), as shown in graph Figure 63.

This behavior depends on the 4A-reporting threshold that has been set higher than the GET dimension, so a user request does not imply the allocation of a dedicated channel, but only the first packet in down-link consents the switching on DCH.

Consequently the DCH time is lower than the FACH time, generating a switching duty cycle of:

$$\frac{time_{DCH}}{time_{DCH} + time_{FACH}} \cong 41\% \quad (13)$$

as shown in Figure 62.

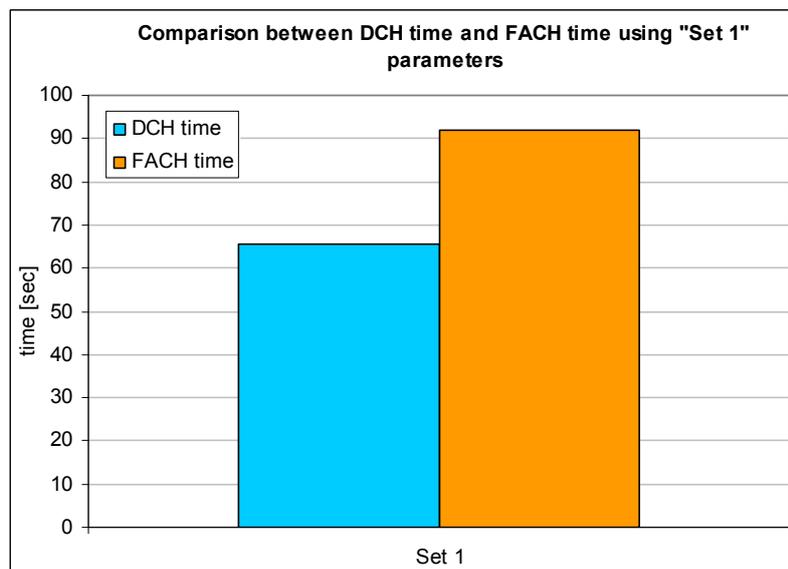


Figure 62 DCH time versus FACH time using “Set 1” parameters.

Since the mean number of packet calls is equal to 10, the number of switching from DCH to FACH and vice versa should be the same, because it is desirable that during a packet call DCH is used, while during the reading time a FACH is carry on.

Instead the number of switching from DCH to FACH and vice versa is higher than the expected value, being more or less twice the quantity, as shown in the graph Figure 63. The number of switching from FACH to DCH is always smaller of 1 unit than the number of switching from DCH to FACH, because the simulation allocates initially a DCH and finishes in FACH.

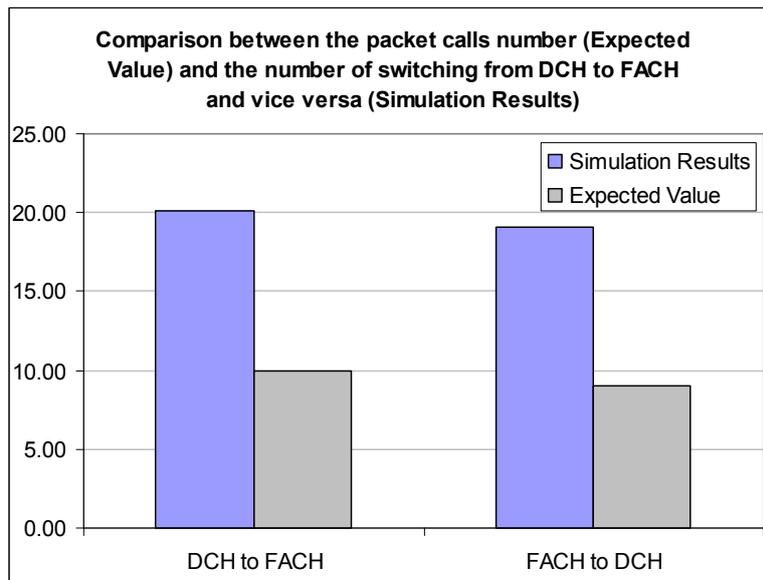


Figure 63 DCH number of packet calls versus number of switching from DCH to FACH and vice versa.

These considerations can be summarised on the time behavior shown in the Figure 64, where a switching from DCH to FACH and vice versa is triggered also during a packet call. The red line represents the WWW traffic profile, where the packet calls alternates the reading time. The blue line represents the switching from DCH to FACH and vice versa.

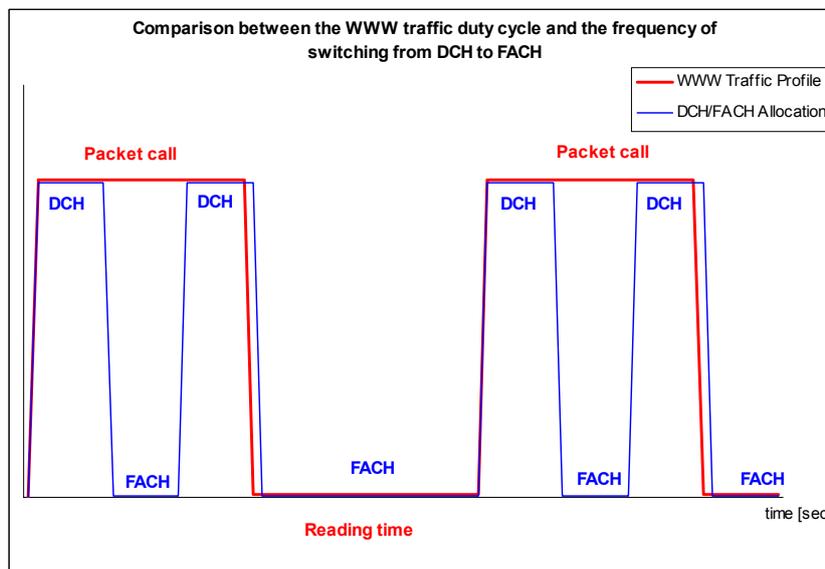


Figure 64 WWW traffic duty cycle versus frequency of switching from DCH to FACH and vice versa.

The second aspect, related to the signalling, has been emphasized by the number of 4A and 4B events tracked during the simulation. Since both 4A and 4B times to trigger are set equal to 0, a measurement report is immediately sent to the RLC when the threshold is exceeded. More over each pending time after trigger (250 milliseconds in the considered case), other reports are sent. This causes the triggering several times either for the 4B event, as shown in Figure 65, or for the 4A event, as shown in Figure 66.

These graphs show the number of tracked events and the ignored⁶ ones both in uplink and in downlink.

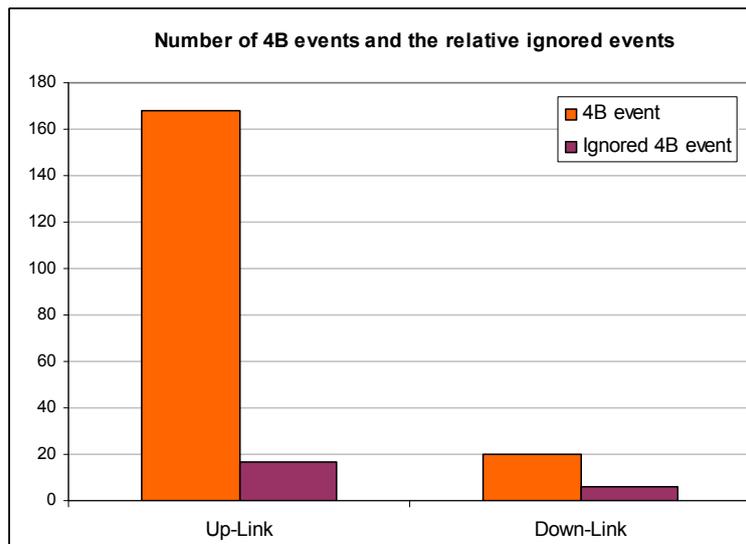


Figure 65 Number of 4B events in both uplink and downlink (executed and ignored ones).

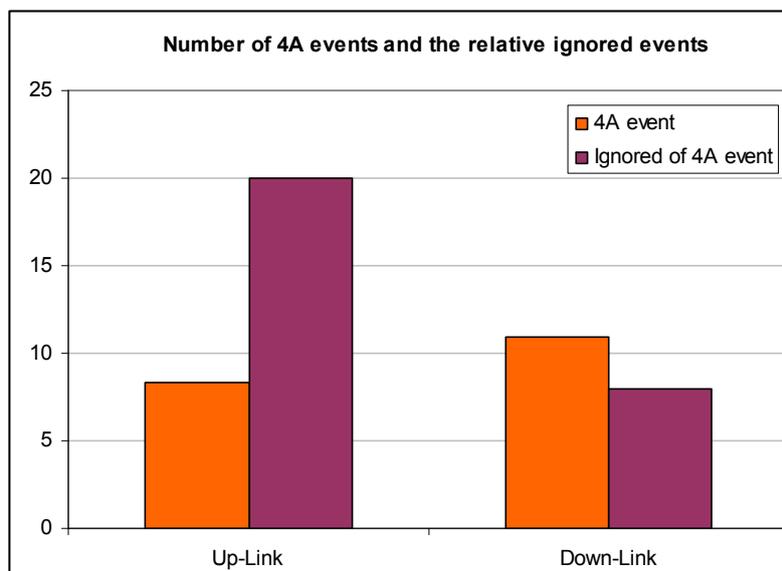


Figure 66 Number of 4A events in both uplink and downlink (executed and ignored ones).

Second simulation has been run with “Set 2” TCTS parameters and the same traffic load characteristics.

The main differences with respect the “Set 1” parameters are:

- the 4A-reporting threshold which is lower than the GET dimension;
- the time to trigger for 4B event that is higher than before;
- both 4A and 4B pending times after trigger, which are higher than before.

4A-reporting threshold change balances the ratio between the DCH time and the FACH time, by growing the first one and generates a switching duty cycle of 47.9%, as shown in Figure 67.

⁶ When the transport channel type switching procedure is in progress, UTRAN ignores other reports coming from the user.

This value consents to pursue better the WWW traffic profile.

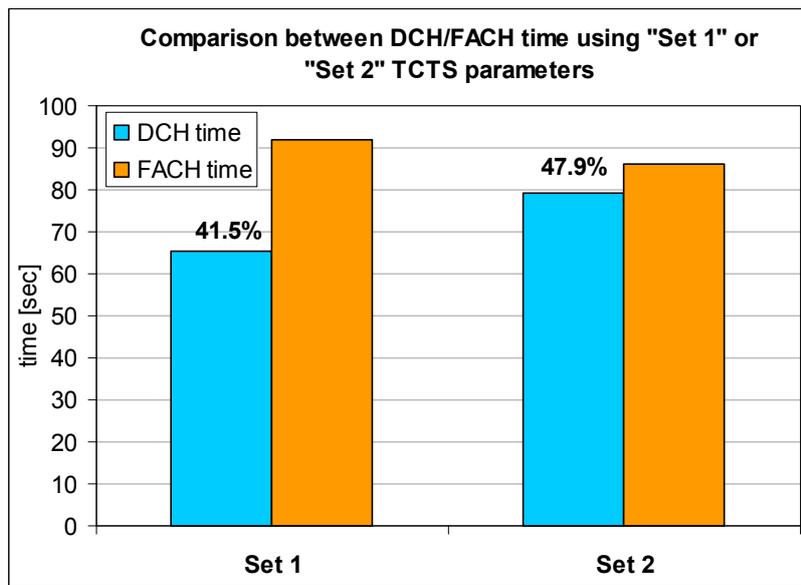


Figure 67 DCH time and FACH time (“Set 1” versus “Set 2” TCTS parameters)

To confirm this behavior it can be observed that the number of switching from DCH to FACH and vice versa is lower than before and closer than the packet calls number, as shown in the Figure 68.

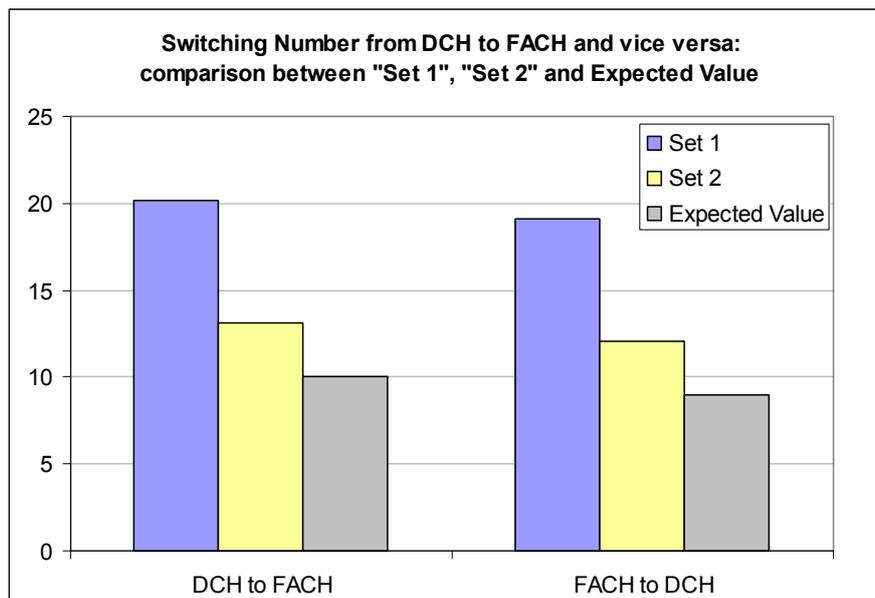


Figure 68 Comparison between the switching number using “Set 1” and “Set 2” TCTS parameters

The other changes, related on the time to trigger for 4B event and both 4A and 4B pending times after trigger, reduce the number of measurements reports sent to the RLC.

To trigger an event the condition must be fulfilled during the *time to trigger*, and consequently reports are sent. By growing the 4B time to trigger if a threshold is only temporary exceeded, because the traffic load fluctuates around the threshold, any report is sent to RLC and DCH channel isn’t released. More over also the 4B pending time after trigger has been gone down, hence even if the condition to report a 4B event remain true, only each 5 seconds a

report is sent to RLC. This improvement is shown in the Figure 69, where the comparison between the 4B event number using respectively “Set 1” and “Set 2” TCTS parameters is presented: it can be observed how the 4B event number is drastically decreased with respect the previous conditions.

Also the 4A event is less triggered, because its pending time after trigger has been set to 1 second with respect the previous 250 milliseconds. The effect is smaller than for the 4B event but it can be all the same observed in the Figure 70.

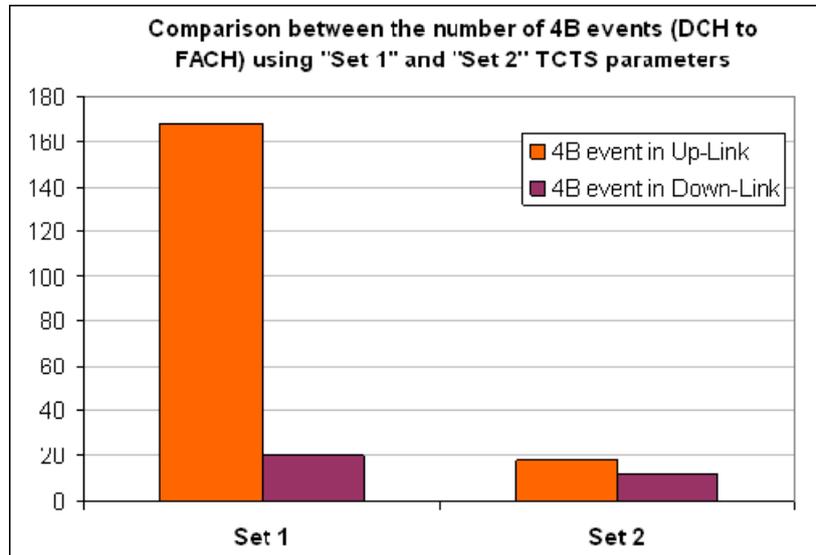


Figure 69 Comparison between the number of 4B event in uplink and downlink using respectively “Set 1” and “Set 2” TCTS parameters

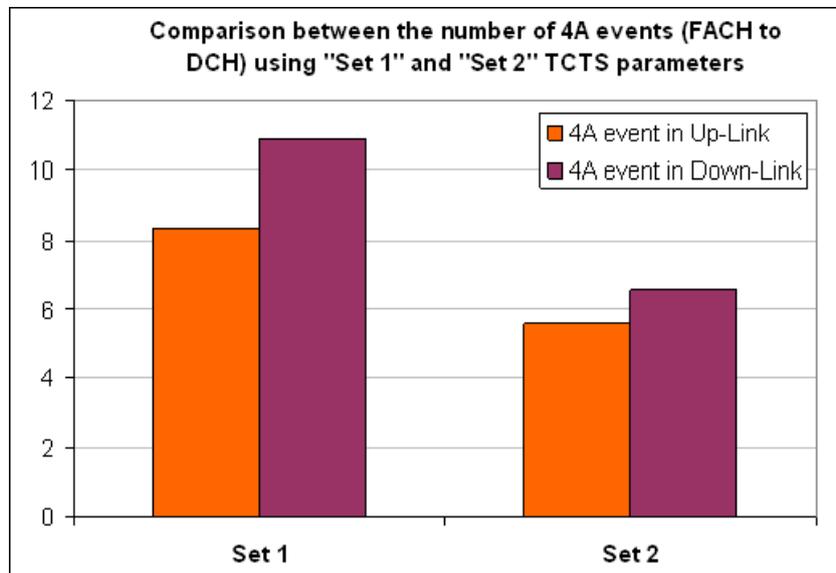


Figure 70 Comparison between the number of 4A event in uplink and downlink using respectively “Set 1” and “Set 2” TCTS parameters

The last analysis has been carried out with “Set 2” TCTS parameters, by varying only the 4B time to trigger from 1 second to 3 seconds. By growing the 4B-time to trigger, the condition to switch in shared channel has been fulfilled for longer time; hence each user does not release its dedicated channel so often. In this contest it is obvious that the average download throughput increases, as shown in Figure 71.

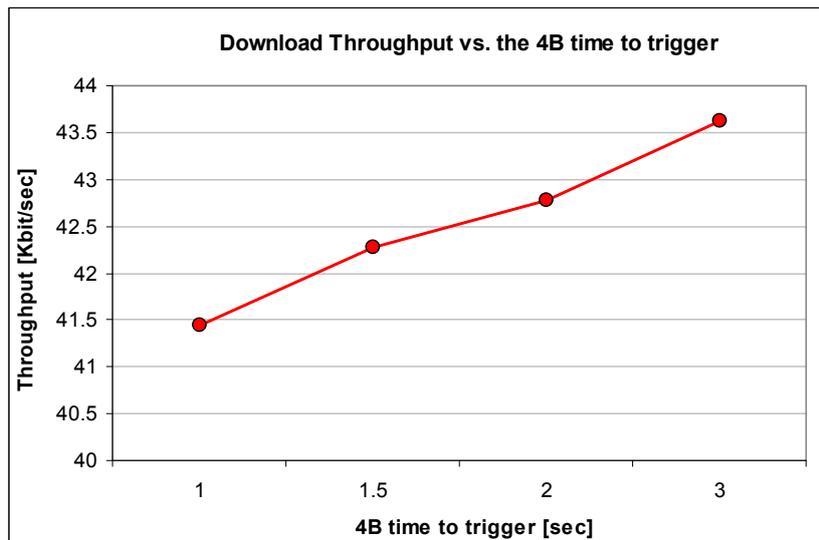


Figure 71 Download throughput versus the 4B time to trigger.

However in a performance evaluation it is necessary to consider other aspects, except the throughput, above all the radio interface capacity. If a dedicated channel is held by a user for longer time, lower number of users can access to the network, as empathized in the previous section. For example, Figure 72 shows how much increase the DCH time with respect the 4B time to trigger:

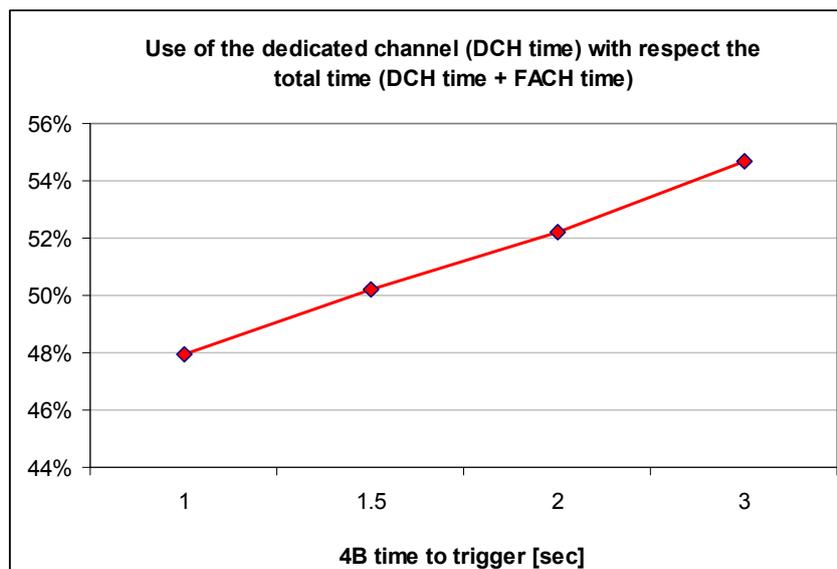


Figure 72 Use of the dedicated channel with respect the total time.

2.10.6 Conclusions

The RRM algorithms controlling the transport channel type switching procedure can have a very important role when users request VBR (Variable Bit Rate) services. This is the case, for example, of the World Wide Web browsing: this service implies a very discontinuous data transfer due to the reading time spent by each user to analyze the received WEB page. In cases like this, it is very important to be able to optimize the usage of dedicated transport channel (DCH), preventing channelization code shortage in the downlink without degrading the end-to-end quality of service experienced by the user in an appreciable manner.

As a general rule, even if the common channels are used only when a very small amount of data (or nothing at all) have to be transferred, the QoS is better for DCH-only than for TCTS because of the switching time. DCH offers higher transfer speeds (throughput) but requires a significant setup time, whereas shared channels have a low throughput but also a low setup time. Due to these considerations the usage of common channels is more efficient only when the traffic is sporadic or for short time, otherwise for long and frequent data sessions, the usage of a dedicated channel is highly recommended.

Another very important issue consists in the proper setting of the parameters controlling the reporting of the traffic measurements performed by the users, in order to prevent an overhead of useless signalling traffic.

2.11 ADMISSION CONTROL

This section summarizes the most relevant results obtained during the analysis of how different prioritization strategies for admission control impact on the UMTS GoS.

2.11.1 Scenario description

The chosen scenario is based on one of the proposed in [24]. This scenario is inspired in a business area of one of the main cities of Spain; high office buildings, residential blocks and a principal avenue characterize it.

Parameters detailed in [24] have been used for NodeB and users. The RABs selected for the services were also defined in [24]:

The traffic density used for the simulations follows the pattern obtained from the current traffic distribution in the GSM network. The traffic pattern is similar to a hot spot traffic model, where users are mostly distributed in the main avenues and buildings.

Two classes of users [24] have been defined; consumer and business users with different service usage. These two user classes and their service mixes were selected to study in the EVEREST framework and they are described in detail in [24]. A 50% / 50% distribution between the two users classes has been chosen in the scenario, following the indications for a dense urban scenario given in [24]. The service mix ratio used for simulations is depicted in the next figures for each user class.

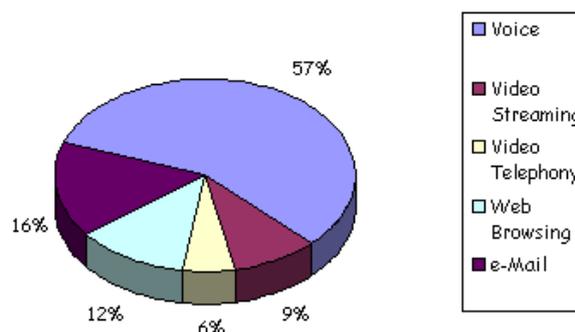


Figure 73. Service mix for the business user class

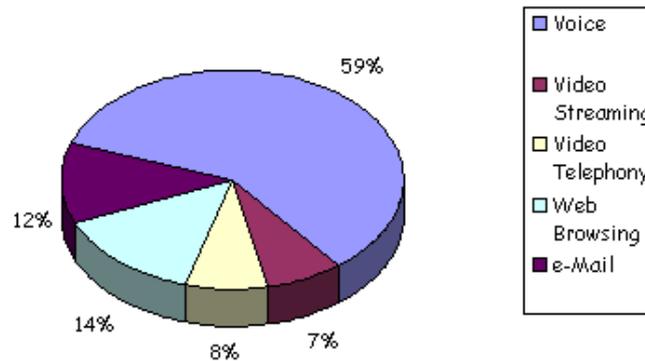


Figure 74. Service mix for the consumer user class

It is also important to remark that the simulations carried out are static, i.e, Montecarlo simulations. Each simulation is equivalent to a set of snapshots of the mobile positions. Due to the random behaviour of the mobile positions, it has been necessary to repeat the process a number of times (in this case 20 times) enough to be sure that the system operation is adequate even in the most unfavourable cases. Also, it has to be noticed that all users are active in the right instant of each snap shot.

2.11.2 Services, Users and Services&Users prioritization

Several services, with their corresponding QoS, and two kind of users, business and consumer, have been defined in EVEREST project [24]. In the previous WP3 deliverables, [17] and [1], the capacity obtained with the following strategies was presented:

- Business users were given more priority when connecting than consumer ones.
- A priority list for services was followed.
- A joint study of both strategies. Both, type of user and requested service are considered to establish a priority list that the network follows when connecting users.

The priority list considered is:

- Voice for business users.
- Rest of services for business users.
- Voice for consumer users.
- Rest of services for consumer users.

This strategy is compared with another one that consists on giving priority just to the type of user.

The main conclusions were that these strategies had sense in highly loaded networks, when NodeBs were saturated and their mean transmission power was similar to the maximum. Then the number of connected users of the prioritized services or type of user got increased or at least decreased softer than the rest of services or type of user.

2.11.3 Services prioritization study in an indoor scenario.

In this section the results on the prioritization strategies in the indoor scenario will be presented.

The scenario used is the 2c scenario described in [1] but some other outdoor Nodes-B have been included.

First of all a brief description of the 2c scenario will be done. Then the strategies to be followed will be detailed. Basically the following stragies are being probed:

- No prioritization strategy is followed when connecting the users.

- Data services (all of them with the same priority) will have more priority when connecting than voice users.

The number of data users connected to indoor cells and the percentage of voice traffic attended by the outdoor stations are the key parameters that are being studied.

2.11.3.1 Scenario description

The employed scenario is a modification of the 2c scenario described in [1]. It is a real three-storey office building with meeting rooms and office pools. The number and position of indoor UMTS microcells have been defined following the actual WLAN planning. The reason is that UMTS indoor deployment strategies are not very clear yet and this WLAN planning has been considered as a good reference.

The EIRP of the UMTS stations is 20dBm and the calculation of propagation losses has been done using the model In3 detailed in section 5.4.3.1 of [1].

Three outdoor UMTS macrostations have been added to this 2c scenario. The objective is to observe the influence of them over the services offered in the indoor building.

The location of all the UMTS stations (macro/micro) is shown in the images below.

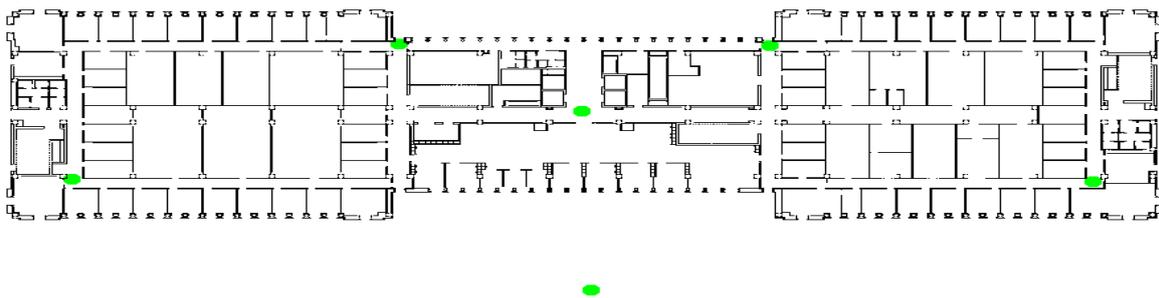


Figure 75 UMTS stations in lower floor.

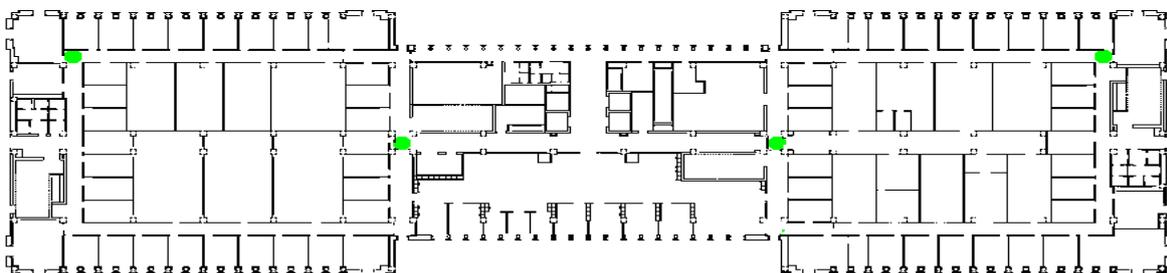


Figure 76 – UMTS stations in simulated floor.

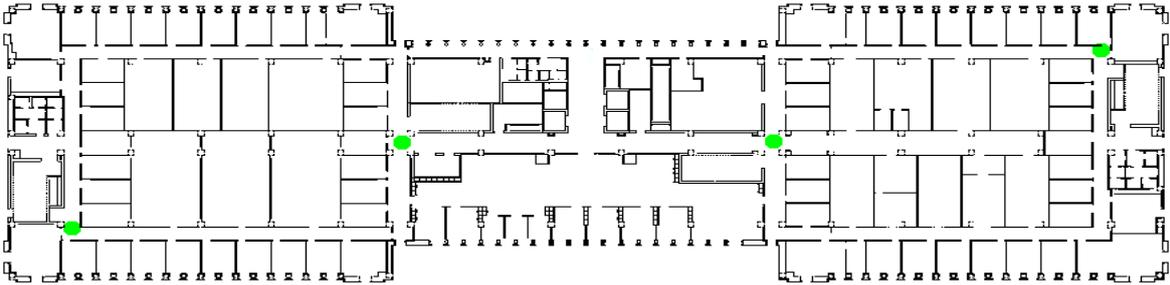


Figure 77 – UMTS stations in upper floor.

Next figures show the calculation of propagation losses in each floor.

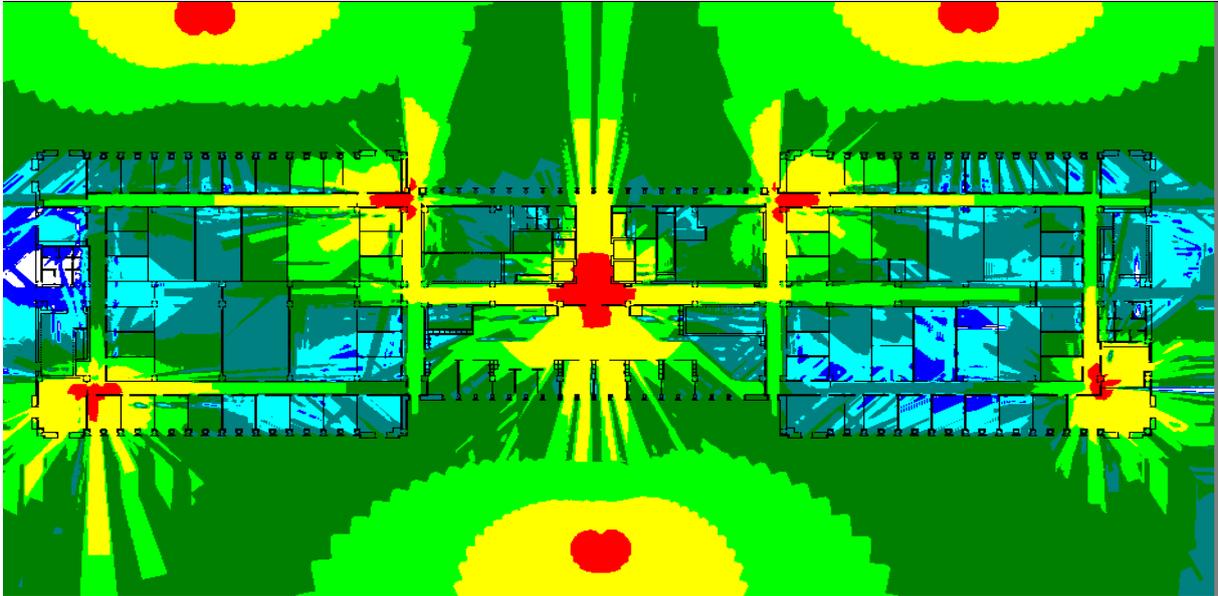


Figure 78 – Coverage calculation in lower floor.

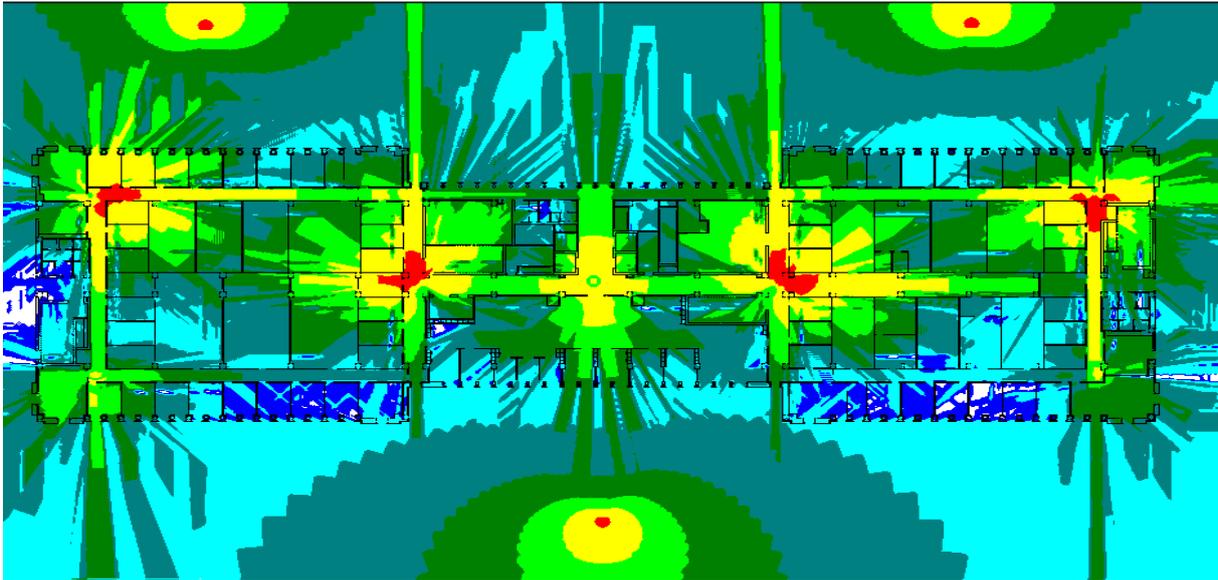
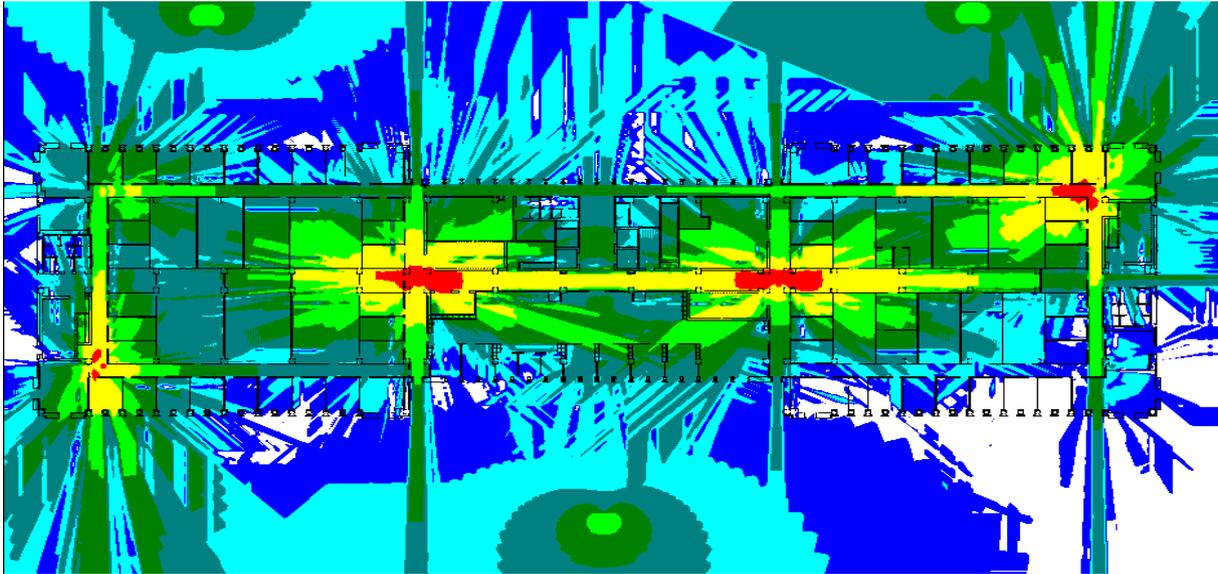


Figure 79 – Coverage calculation in simulated floor.**Figure 80 – Coverage calculation in upper floor.**

The kind of users ‘business’ has been employed, with the same service mix defined also in [24]. The results are obtained using the simulation tool URANO, as in previous studies ([37],[87]).

2.11.3.2 Admission control strategies

The objective of this study is to make an analysis from the operator point of view of about the influence of service prioritization in an in-building scenario.

The study is focused on the intermediate floor, and the results are generated for this floor. Nevertheless, stations in upper and lower floors are considered in the simulation in the situation in which they are in their respective floors (i.e, taking into account that in those floors there are also users requesting services), so they have a certain percentage of remaining capacity.

As in previous simulations, results on percentage of connected users per service will be shown, and also, the percentage of users attended by the upper floor stations, percentage of users attended by the simulated floor stations, percentage of users attended by the lower floor stations and percentage of users attended by macro stations will be provided.

In this type of scenario the users are expected to request mainly data services, as it would be in a future ‘mobile office’ where all data services that actually are provided by fixed network would be requested to mobile networks.

Two strategies have been mainly simulated:

- No priority. All services have the same priority when the users try to connect to the network.
- Data services have more priority than voice service when the users try to connect to the network.

In the graphics below the results for each service are shown:

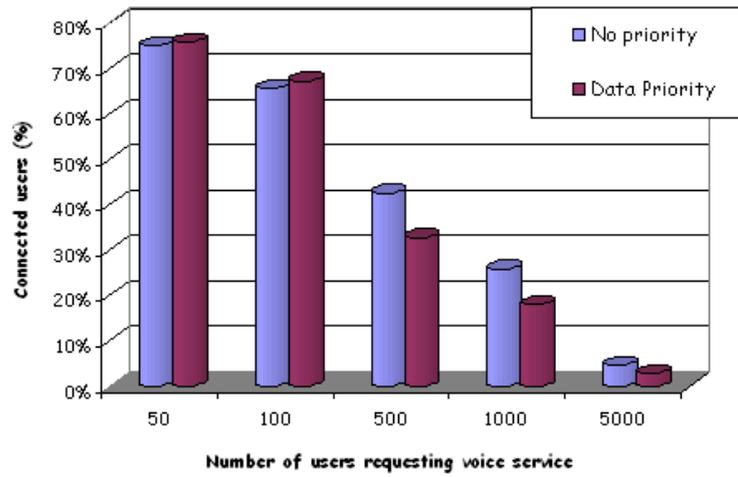


Figure 81 – Connected users for voice service

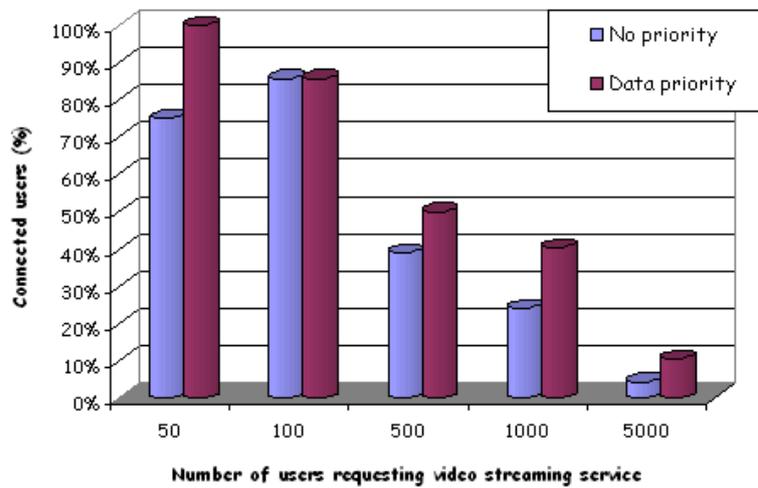


Figure 82 – Connected users for video streaming service

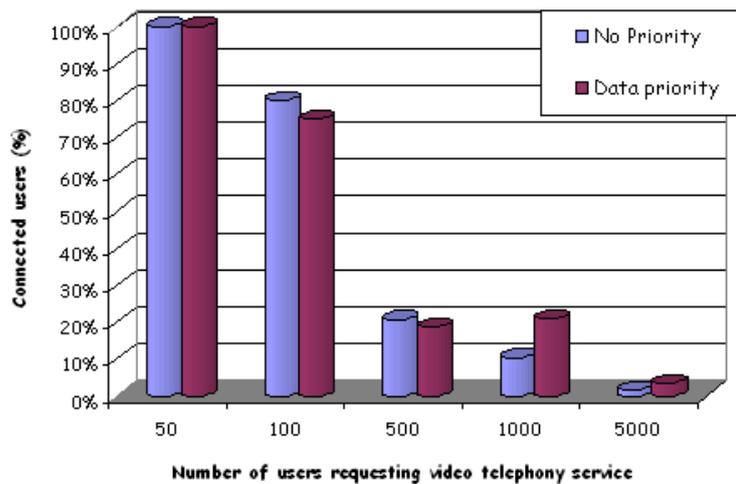


Figure 83 – Connected users for video telephony service

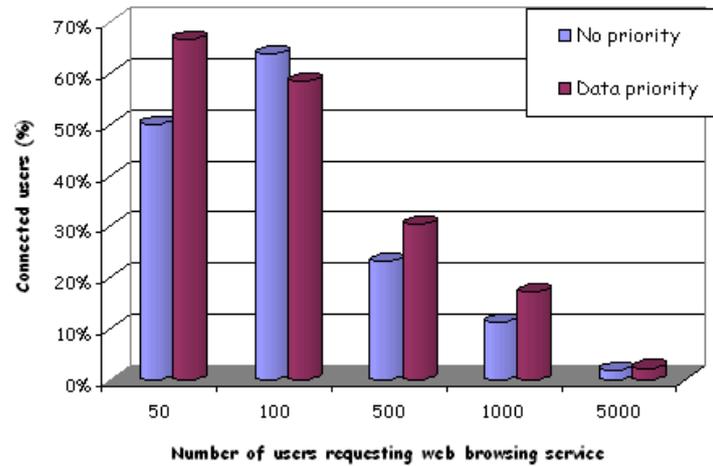


Figure 84 – Connected users for web browsing service

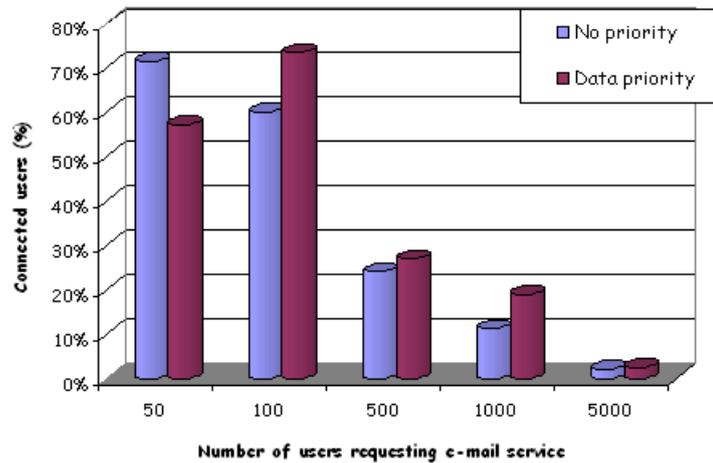


Figure 85 – Connected users for e-mail service

In the graphics below the percentage of attended users in each floor is shown:

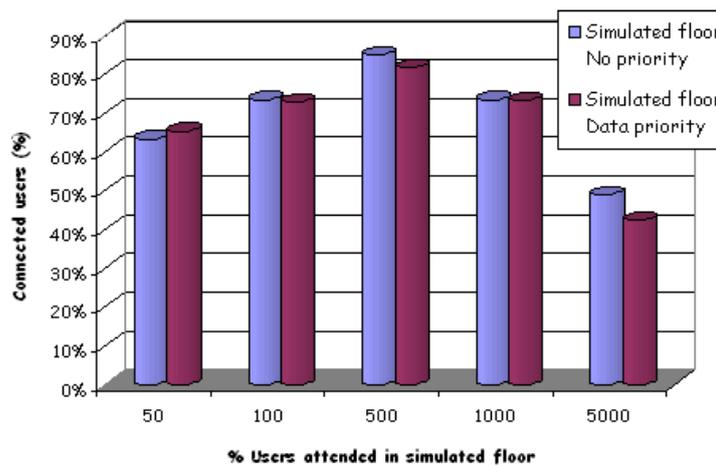


Figure 86 – Percentage of attended users by the stations situated in simulated floor

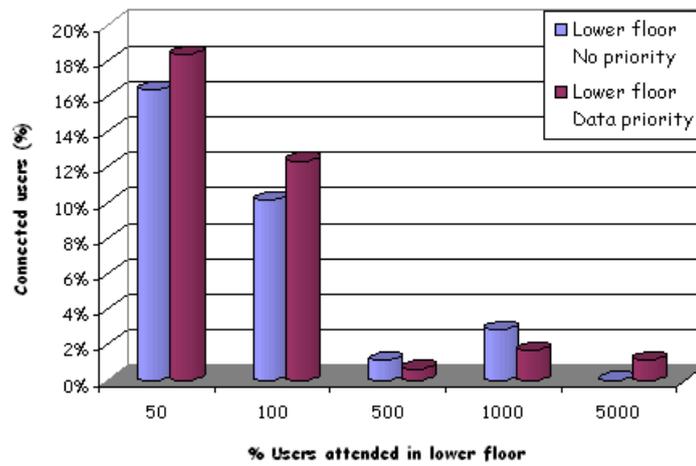


Figure 87 – Percentage of attended users by the stations situated in lower floor

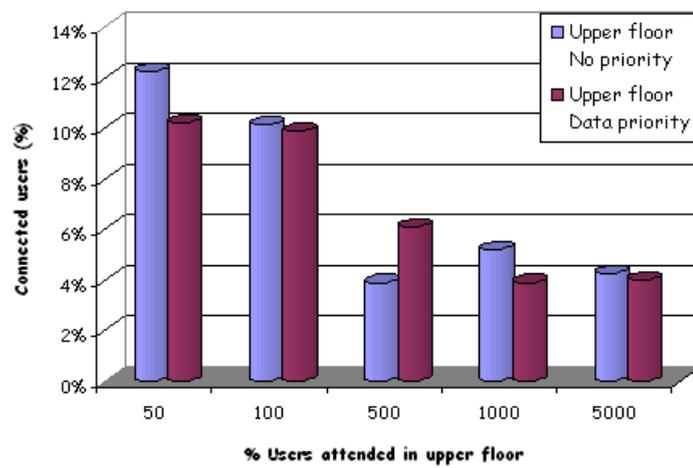


Figure 88 – Percentage of attended users by the stations situated in upper floor

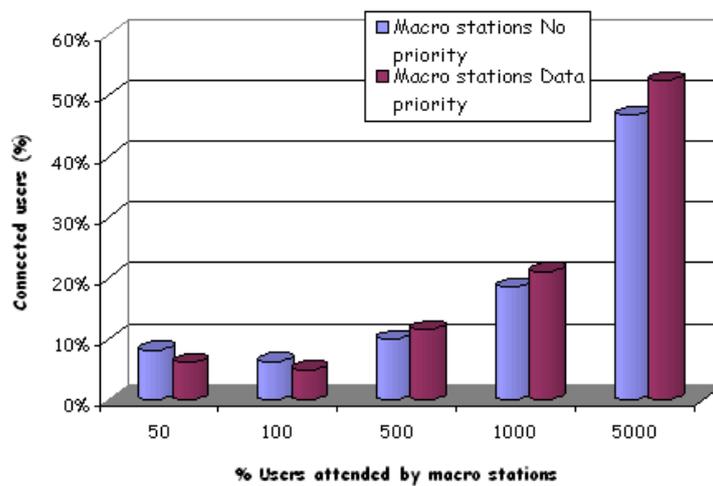


Figure 89 – Percentage of attended users by the macro stations

2.11.3.3 Conclusions

The following conclusions may be inferred from the graphics above:

- For low load situations, the effect of prioritizing services is faintly noticed. It is the same effect appreciated in simulations for previous deliverables.
- As the network load grows up the percentage of connected users for data services considerably increases. The percentage of attended voice service users decrease, but not dramatically (as in past simulations). On the contrary, the lost is almost 10% with respect to no-prioritization strategy. Perhaps, this lost may be compensated with another macro station in the outdoor.

The simulations have allowed corroborating what was being observed by the mobile operator by means of measurements: the percentage of users located in a certain floor attended by stations situated in a different floor is very low.

It may be inferred from the graphics that this planification scheme, with outdoor macro stations, would have a good performance because when the indoor microcells tend to saturation (because of the number of attended users and also of the interference level) the macro stations start absorbing this traffic. This would be a proper solution when a hotspot in the building happens and the situation outside remains constant. Then, the solution would be locating macrostations outside the building and prioritizing data services.

2.12 DIFFSERV AWARE SCHEDULING

The research presented in this section focuses on the color aware RRM algorithms to support DiffServ over CDMA radio systems. The schemes are designed to use the IP layer traffic information to increase the utilization of radio network. Based on the architecture presented in [24], the DiffServ traffic information is assumed to be provided to the RNC through the interface between CN BS and RAN BS such as translation functions. With the information, the radio resource is allocated by these proposed schemes in a way consistent with the QoS requirements of CN DiffServ traffic in order to improve the resource utilization in terms of user's QoS satisfaction. e.g. the QoS of conforming traffic is guaranteed while the non-conforming traffic's performance is penalized in the congestion state.

2.12.1 DiffServ Marking

Metering and marking functionalities of packets inside DiffServ AF classes are on per flow basis (even though these policies can also be applied for aggregated flows). Marking incoming traffic at the ingress DiffServ router is one of its most important functionalities. Traffic that conforms the service profile will be handled with low drop precedence, while non-conforming traffic will be handled with high drop precedence. In the sequel, the marking algorithm that we use for our study is called the Time Sliding Window three color marking TSWtcm [34]. In this marking scheme, the arrival rate is calculated according to the weighted average of the arrival rate over a certain time window. Thus, whenever a packet arrives, the marker calculates the estimated arrival rate. If the estimated arrival rate is less than the Committed Information Rate (CIR), the arriving packet is marked as green; otherwise, they are marked as green, yellow or red according to a calculated probability which also depends on Peak Information Rate (PIR). If we denote the measured average rate of the flow as r , the pseudo code of the TSWtcm can be described as follows,

Figure 90 shows the probability of packet marking with the arrival rate normalized to CIR (PIR=1.5CIR). The system architecture that we consider is based on the end-to-end QoS architecture for UMTS presented in D06. In this architecture, DiffServ edge function is provided by the gateway. In the gateway, IP BS manager is responsible to communication

with other DiffSer domains as a Bandwidth Broker. If policy based QoS end-to-end management is applied, Policy decision function can be regarded as a Policy Decision Point. In this case, mobile network can be seen as an autonomous stub network. Thus, through the translation function located in the gateway, we assume that each DiffServ flow's marking information can be made transparent to radio network controller for a better resource management.

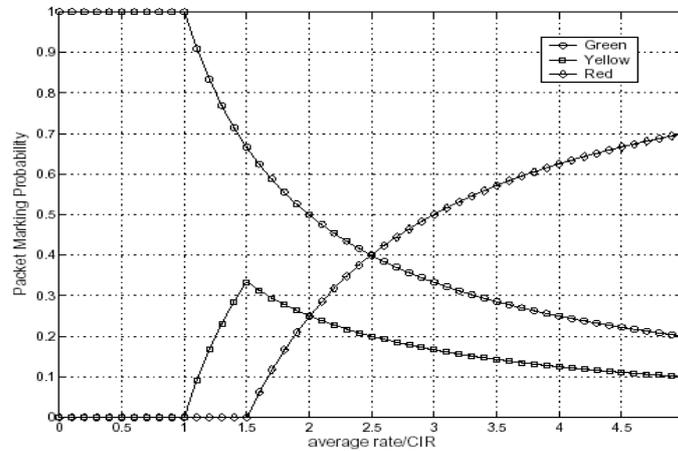


Figure 90 Probability of Marking TSWtcm

2.12.2 Motivation of Color Aware RRM

The critical impetus of the proposed approach, beyond minimization of the transmitted power, is mainly twofold. Firstly, to achieve the required per-class aggregate data rate while prioritizing and ensuring QoS of in-profile packets, and secondly to increase power gains by penalizing out-of-profile packets in sense of power consumption. The seminal aspect of the proposed scheme is that tangible power gains can be achieved by differentiating transmission of conformant and non-conformant packets while at the same time the aggregate power gains of AF classes can be utilized to enhance the performance of in-profile traffic.

Traditional power and rate adaptation techniques treat all packets equally and the differentiation is only based on lower layer criteria such as channel conditions and inter-cell interference for each user. In that sense, allocated power and rate vectors for packets in the same or different DiffServ classes are colour blind. Thus, we propose an integrated framework for power and rate adaptation that is colour aware, takes into account lower layer information and provide the required aggregate per class transmission rate. Using a different weighting factor (ξ_{green} , ξ_{yellow} , ξ_{red}) depending on the colour of the packets, transmission times and related power allocations will be differentiated in order to firstly prioritize in-profile packets and secondly to reduce required power consumption for out-of-profile packets. Instead of fixed degradation of the out-of-profile packets, the weighting factors increased linearly depending on channel conditions and inter-cell interference as shown in Figure 91(a). In the figure, there is no degradation for green packets while the difference between the yellow and red packets can be controlled by the $d\xi$ parameter. Based on this framework the actual transmission time of user i for a yellow or red packet will be augmented by ξ_i . Also, in the worst case scenario a red packet have been selected to have a degradation factor equal to $\xi_{\text{red}} = 0.5\tau$ (where τ denote the average transmission time per user in the AF class in order to achieve the required aggregate data rate). If by τ_i we denote the transmission time for mobile host i , then based on the output of the TSWtcm the average transmitted rate can be written as,

$$R_i = [p_i^g \tau_i + p_i^y (\tau_i + \xi_y) + p_i^r (\tau_i + \xi_r)]^{-1}$$

14)

where p_i^g , p_i^y and p_i^r represent the probability that a packet from user i is marked as green, yellow or red respectively. A poignant observation of the proposed linear function on degradation reveals a reminiscent with the RIO technique where instead of the queue length the channel and inter-cell interference are taken into account and it is deterministic rather than probabilistic.

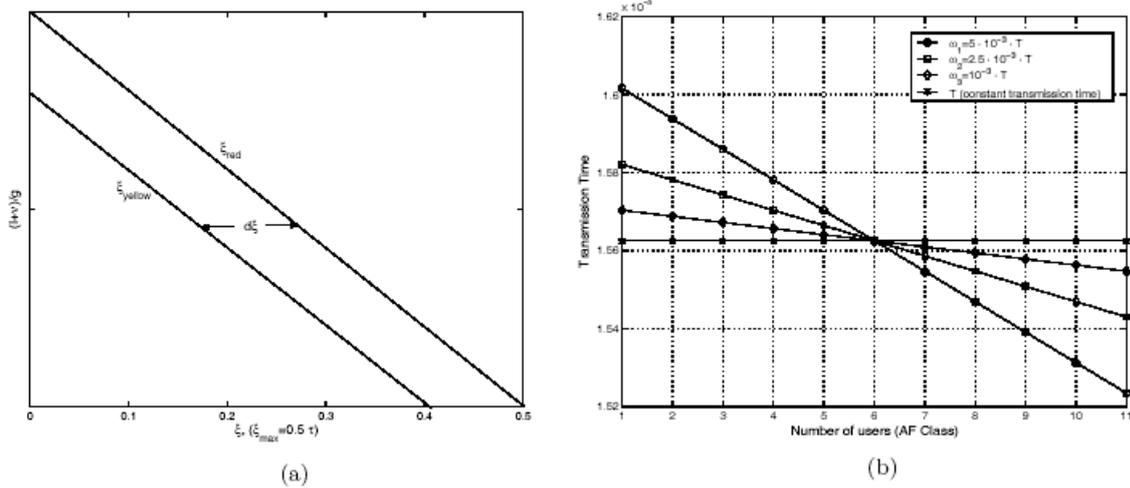


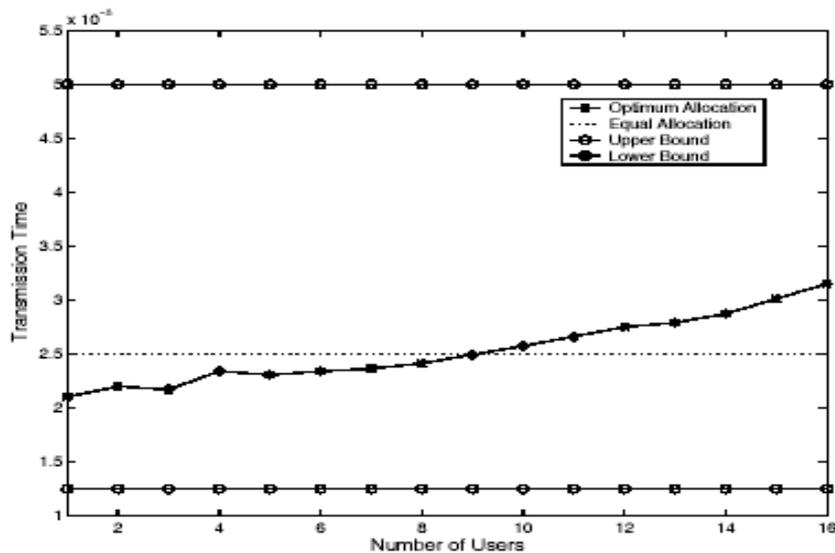
Figure 91 (a) A linear function that represent the additional delay introduced for non-conforming packets (yellow, red) (b) Constant versus adaptive allocation of transmission times for users sorted according to channel and inter-cell interference conditions (shown for different values of u).

With the weight factor presented in (14) and the procedure presented in [1] section 2.8, we can derive the constrained optimization for color-aware link adaptation as following

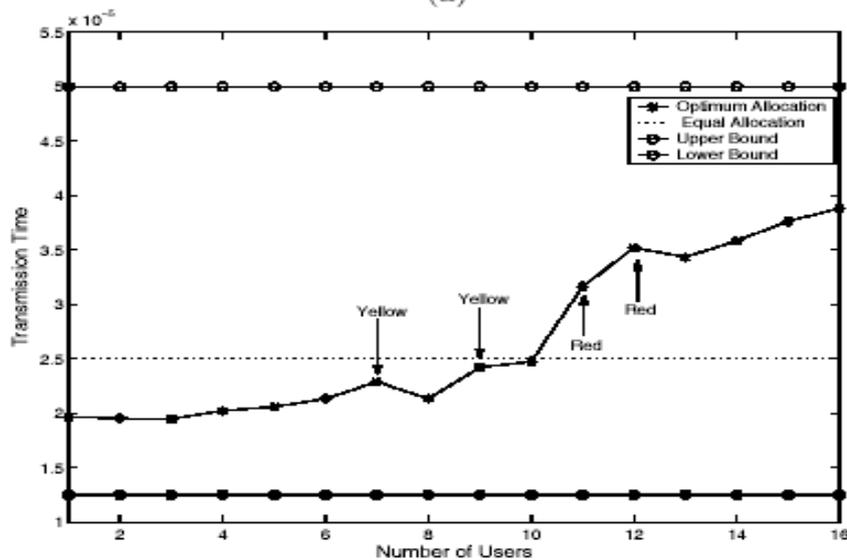
$$\begin{aligned} \min_{\tau_i} \quad & \frac{1}{1 - \sum_{i=1}^{n_j} \frac{\theta \Gamma_j}{W(\tau_i + \xi_i) + \theta \Gamma_j}} \sum_{i=1}^{n_j} \frac{\Gamma_j}{W(\tau_i + \xi_i) + \theta \Gamma_j} \cdot \frac{I_i + \nu}{g_i} \\ \text{s.t.} \quad & \sum_{i=1}^{n_j} \frac{\Gamma_j}{W(\tau_i + \xi_i) + \theta \Gamma_j} \cdot \left(\frac{I_i + \nu}{p_{n_j} g_i} + \theta \right) \leq 1 \\ & \sum_{i=1}^{n_j} \frac{1}{\tau_i + \xi_i} = R_j, \quad \left(\frac{E_b}{I_0} \right)_j = \Gamma_j, \quad \forall j, i \end{aligned} \tag{15}$$

Figure 92 shows the optimum allocation of transmission times for the N AF users which have been sorted based on the channel conditions and inter-cell interference. The aggregate transmission rate for the AF class has been set to 640Kbps, which give an average rate of 40Kbps per user. Figure 92(a) depict the case when all the transmitted packets are marked as green, while (b) Figure 92(b) shows how transmission times are affected when a specific number of transmission packets, $K = 4$ have been marked as yellow and red assuming the

same conditions. The bounds of the optimization problem control the worst case unfairness between active flows. As the distance between the bounds and the average transmission time is decreased so do the unfairness on assigned rates between flows. The additional delay introduced to the non-conforming traffic has been calculated based on channel conditions and inter-cell interference as have been shown in Figure 91(a) in section IV. The upper and lower limits of the linear function are dynamically updated influenced by the best and worst channel and inter-cell interference of the N active users, while the $d\xi$ parameter have been chosen to be equal to 0.04τ (τ is the average transmission time). We should point out that not only the transmission times of the yellow or red packets are affected but there is a global permutation on the allocated transmission times for all users. This should come by no surprise, and the simple explanation of this behavior is based on the fact that the local (or global) minimum will not only change in the K -dimensional subspace defined by the yellow and red packets but in the whole N -dimensional space because they are not linearly independent.



(a)



(b)

Figure 92 (a) Optimum allocated transmission times with only green packets.(b) Optimum allocated transmission times when transmitting yellow packets for users 7, 9 and red packets for users 11 and 12.

2.12.3 Applicability example: Color aware link adaptation protocol

In this section, a link adaptation protocol is proposed based on the theoretical study shown in the previous section. This protocol is designed to integrated the QoS information from the IP layer of the Differentiated Services Architecture (DiffServ) with lower layer criteria, in order to support packet transmission over a WCDMA-based wireless interface in the differentiated way in consistent with DiffServ core. To achieve this, a color aware scheduler, which is based on the output of a Time Sliding Window three colour marker (TSWtcm) of the ingress DiffServ node of radio access network, and a rate adaptation algorithm are proposed. Under the joint control of the scheduler and rate adaptation, the resource in the CDMA systems is dynamically allocated to differentiate the in and out profile of DiffServ packets, where the conforming traffic performance is enhanced whilst the performance of non-conforming traffic is penalized in the congestion state

2.12.3.1 Proposed scheme

The proposed link adaptation concentrates on the differentiation of the radio resource based on the colour marking scheme and improving the data transmission subject to the BS power constraint. To achieve this, two types of queues are constructed in the BS as shown in Figure 93. One is used as a buffer for each user to store the incoming traffic packets, the other is an ID queue, which contains the active users' IDs. In the ID queue, queuing policies are applied to keep an order list for users' transmission. Two queuing policies are considered for comparison. They are Modified Round Robin (MRR) and Interference Factor (IF).

To cooperate with the DiffServ marking scheme, we further divide the ID queue into three sections corresponding to Green Section, Yellow Section and Red Section. In this separation, the green section is always on top of the other two sections and the yellow is on the top of the red section. In each section, the above mentioned queuing policies such as MRR and IF are applied. With this queuing structure, the high transmission priority is guaranteed for the conforming traffic over non-conforming traffic. The resource over air interface is allocated in the consistent way with the DiffServ core.

Based on the priority in ID queue, rate adaptation is designed here to minimize the packet delay subject the power constraint as shown in Figure 93. This is to cooperate with the priority given by the scheduling scheme and discrete transmission rate subject to the constraint of BS power.

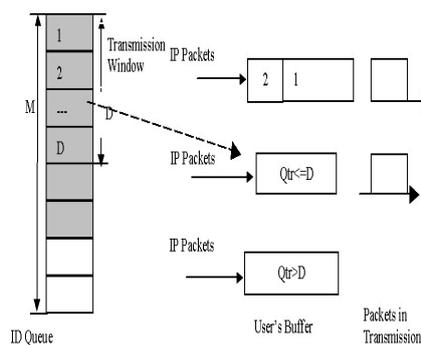


Figure 93 BS structure and Scheduling Transmission

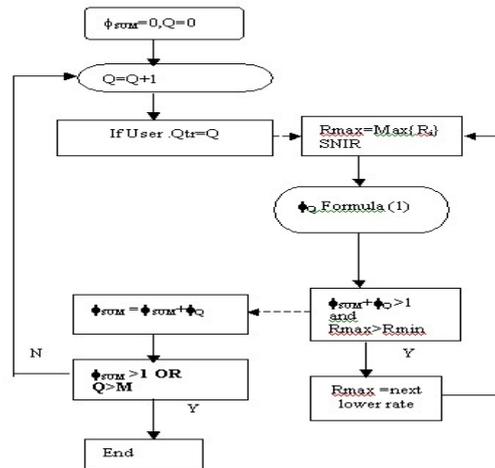


Figure 94 Rate Adaptation Algorithm

So, in the proposed link adaptation scheme, AF users' transmission order is given by the ID queuing algorithm, and the priorities are updated every frame. Then the rate adaptation algorithm is performed starting from the user on the top of the queue to choose a highest transmission rate. Through this rate adaptation, each transmitting user can transmit at the highest available transmission rate, and thus at the same time the transmission window size (D , Figure 93) is kept to a minimum in terms of the number of simultaneous transmission data users, and this reduces the mutual interference between data users and thus increases aggregate throughput. More detailed descriptions regarding the protocol are presented in [1] section 2.8. 4

Results in Figure 95 clearly show the differentiation in resource allocation in terms of transmission rate with the proposed scheme. In this figure as expected, the transmission rates decrease for all type of users as traffic load increases. And the transmission rates for all types of packet are almost the same without color aware shown by the solid line. However in the case with color aware queue, the transmission rate for each color is different to each other. Green users always have a better rate than other users because they are given high priority than other users. Also as traffic load increases, the difference becomes larger. The red users' transmission rate decreases more rapidly (from more than 145bytes drops to just over 100bytes) and in the contrast, the green users rate are only drop about 10bytes.

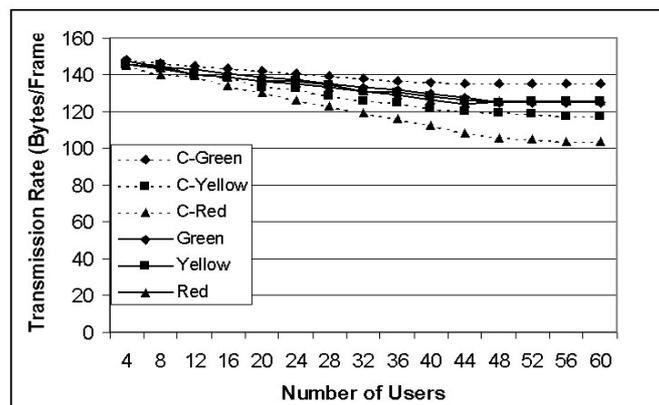


Figure 95 Transmission Rate.

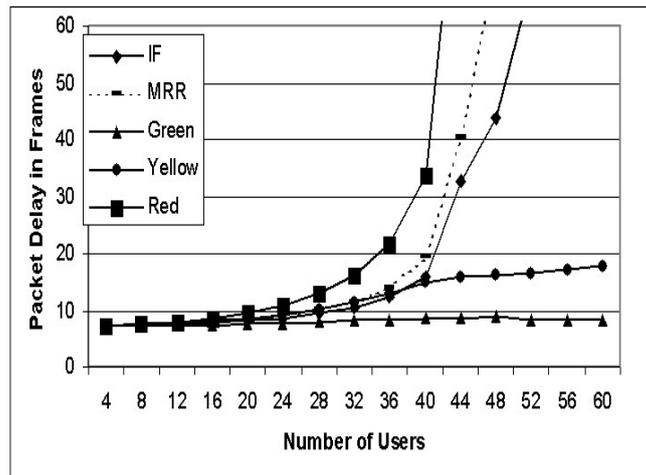


Figure 96 Delay Performance of IF and MRR with color aware.

Figure 96 shows the delay performance of MRR and IF queue both with color aware. IF offers a better performance in the average delay. The diversity of delay performance for different colored packets shows the consistence with the differentiation in transmission rate allocation presented in Figure 95. Again to enhance the delay performance of the in profile packets such as green packets, the out profile packets such as red packets are penalized in high traffic load situation.

2.12.4 DiffServ aware RRM for CDMA

The proposed link adaptation concentrates on the differentiation of the radio resource based on the proportional DiffServ QoS class information to improve the data transmission subject to the BS power constraint. The proposed link adaptation scheme consists of two levels as shown in Figure 97: inter-class power partition and intra-class scheduling/rate adaptation. In the following we give a detailed description of the two schemes.

a. Inter-class congestion control:

Inter-class congestion control is designed to allocate the power resource (indicated by ϕ less than 1) to each class to guarantee the delay requirements of each class with BS output power limit. The dynamic power-based partition scheme is able to dynamically adjust power based on each class service rate following the variations in both traffic volume of each DiffServ class and the total bandwidth over the air interface under the constraint of the BS transmission power. It is proposed to carry out this partition in two steps.

In the first step, high priority is given to EF class over AF classes to reserve the resource (ϕ_{EF}) for its transmissions to guarantee its premium service quality. The residual power unused by EF class (ϕ_{res}) will be allocated to AF classes as the following

$$\phi_{EF} = \sum_i \phi_{EF,i}$$

$$\phi_{res} = 1 - \phi_{EF} \quad (16)$$

And in the second step, partition is used to allocate the power resource to different AF classes according to proportional relative QoS. In our numerical example, the partition is to allocate the resource in terms of the transmission rate for each class in a way to meet the delay ratio between the two AF classes (AF0/AF1) based on the following

$$\frac{B_{AF1}}{R_{AF1}(\phi_{AF1})} / \frac{B_{AF0}}{R_{AF0}(\phi_{AF0})} = \kappa$$

$$\phi_{AF1} + \phi_{AF0} \leq \phi_{res} \tag{17}$$

where B_{AF0} and B_{AF1} are the number of IP packets in buffer and R is the aggregate output rate for a AF class, which can be regarded as a function of ϕ as follows

$$\phi_i = \frac{SINR_i * R_i * (I_{other} + I_{intra})}{\beta_i * S * C},$$

$$0 \leq \phi_i \leq 1$$

This partition follows is done through the Delta Modulation scheme, as follows

$$\text{If } \overline{D_{AF0}} / \overline{D_{AF1}} > \kappa, \phi_{AF0,t} = \max(\phi_{AF0,t-1} + \Delta\phi, \phi_{res,t})$$

$$\text{If } \overline{D_{AF0}} / \overline{D_{AF1}} < \kappa, \phi_{AF0,t} = \min(\phi_{AF0,t-1} - \Delta\phi, 0)$$

$$\text{else } \phi_{AF0,t} = \phi_{AF0,t-1} \tag{18}$$

In this scheme, $\overline{D_{AF0}}$ and $\overline{D_{AF1}}$ are estimated delays for AF0 and AF1 based on their previous transmission. Since the intra-class link adaptation is also to be proposed to manage the transmission, the delays are derived based on the feedback information of uplink regarding the data throughput of each class. And $\Delta\phi$ is a design parameter for this scheme, which is also examined in our numerical evaluation. With this proposed scheme, the residual resource unused by EF class is further divided into ϕ_{AF0} and ϕ_{AF1} . And these values are dynamically adjusted by the delta modulation following the changes in EF power consumption and AF users' distribution from frame to frame.

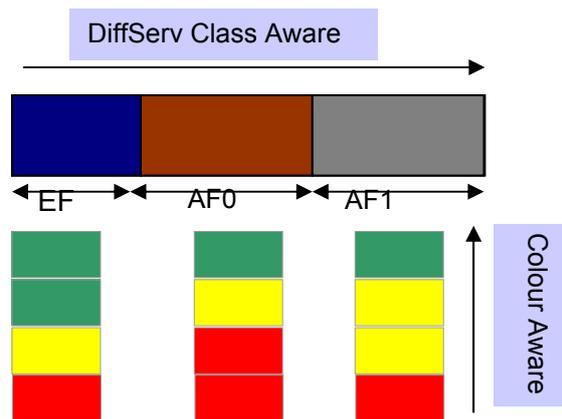


Figure 97 Two-Dimension Management in DiffServ Aware Link Adaptation

b. Intra-class Scheduling and rate adaptation for AF classes

Intra-class scheduling and rate adaptation is designed to differentiate the resource allocated to each AF class based on the color-marking schemes and also maximize the throughput under its own class resource constraint (ϕ_{AF0} or ϕ_{AF1}). The color aware solution presented in section 2.12.1 applied as the intra-class link adaptations scheme.

Figure 98 shows the admission region for AF0 and AF1 with these two inter-class link adaptation, under the same intra-class configuration. This figure clearly shows the capacity gain achieved by the proposed scheme over simple priority queue scheme in which absolutely priority is applied among EF/AF0/AF1. By calculating the admission area (the area under the curve), we find about 25% gain in the admission region with DiffServ inter-class link adaptation. The reason for this improvement is shown in Figure 99. In this Figure, the delay difference between two classes is quite large with the simple priority queuing scheme. For example, with 20 users (AF1/ AF2) users, the admission region is limited by AF1 class by reaching its delay boundary of 2s with the simple queuing scheme. While at this working load, the delay for AF0 is just 0.15s which is far below its boundary. On the contrast, the difference between classes with DiffServ aware queue is rather smaller, and a better QoS balance is achieved by the proposed scheme, thus an improvement in the admission region is achieved (it is limited at about 34 AF0/AF1 users with both of classes about reaching their delay boundaries).

We also investigate effects of the value of $\Delta\phi$ on the systems performance. We find out that real-time traffic load is the main factor related to choice of the value of $\Delta\phi$. It is shown in Figure 100. In general, a smaller value gives a better performance with whether the low or high EF traffic load. Also, it is found that, the effect of the value of $\Delta\phi$ on delay is rather smaller in the light EF traffic situation and this effect increases as the EF traffic increasing, e.g. with no EF links, the delay performance are almost the same for various values of $\Delta\phi$, however when EF traffic increase to 15 and 30, the difference between 0.05 and 0.5 is about 31 frames and 100 frames respectively. This mainly because as the EF traffic increasing, the average residual resource value for AF is squeezing, and thus a bigger value of $\Delta\phi$ make the system difficult to achieve a convergence and therefore a bigger error is expected in the adjustment which results in degrading the performance.

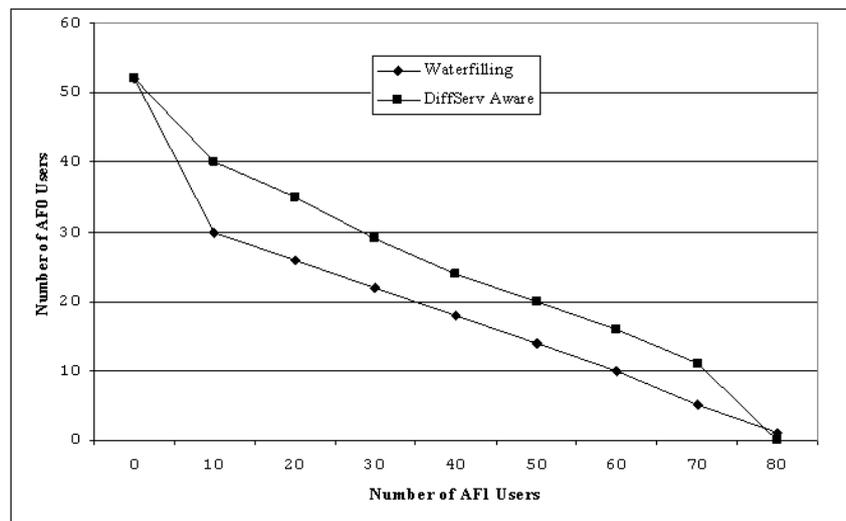


Figure 98 AF Admission Region Comparison

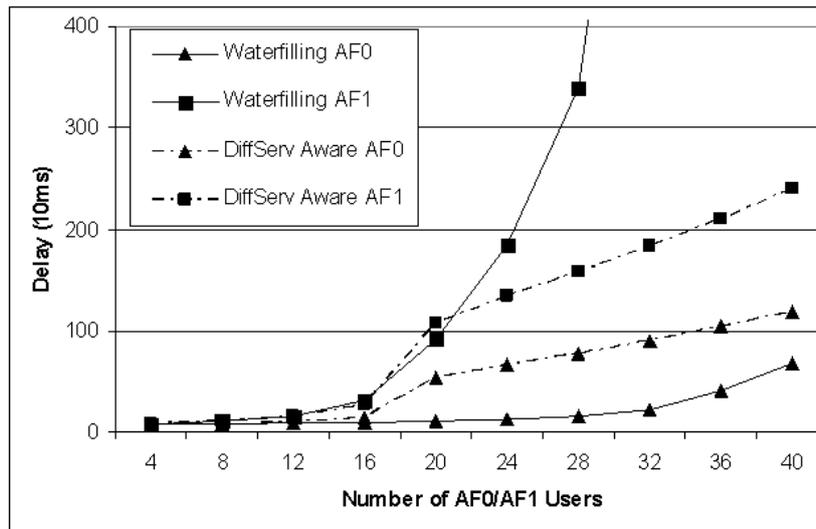


Figure 99 Delay Performance for AF0/AF1 with 10 EF users without shadowing

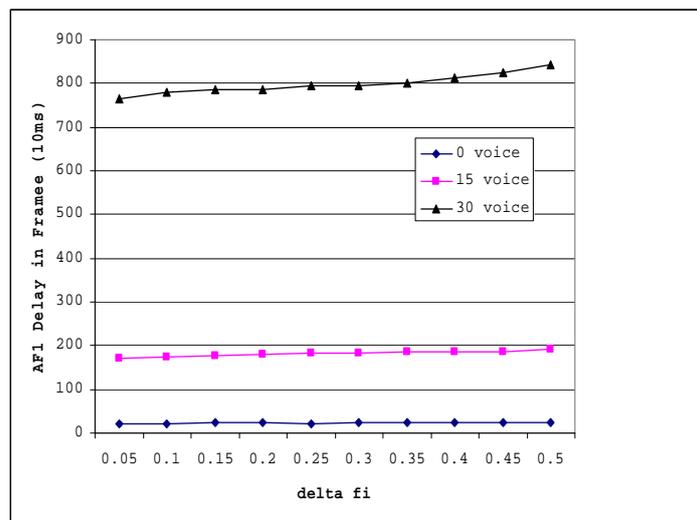


Figure 100 Effects of $\Delta\phi$ on the system performance with 4dB SNIR requirement for EF links (no shadowing)

2.12.5 Applicability example: Color aware coverage control

In this section, following the previous studies in color aware RRM, color aware rate/coverage controlled downlink transmission schemes that utilize the QoS information from the IP layer of the differentiated services (DiffServ) architecture as that shown in D06 are compared. In the rate control scheme, we follow the approach presented in the previous section, the out-of-profile packets, compared to the in-profile packets, are given lower priority in the physical rate allocation and in the coverage control they provided with a less coverage. Simulations results show that the coverage control scheme outperforms the rate control scheme. In [1] section 2.8.5, the details of the coverage control is presented.

To give a comparison to the proposed scheme, Figure 101 shows the outage as a function of the offered traffic, using rate/coverage control schemes similar to the proposed algorithm, but without color differentiation. The outage is equal for all colors in the case of rate control. In

the case of coverage control, the outage is slightly different for each color, with red, yellow, and green in the descending order. This is because a user with more red packets requires a larger power and thus is more likely to be blocked. Figure 101 indicates that the capacity is about 18 users per sector on average if the target outage is 0.05, and the coverage control provides lower outage especially when the offered traffic is more than 20.

Figure 102 shows the outage as a function of the offered traffic for the proposed color-aware rate/coverage control. The outage for red, yellow, and green packets are clearly differentiated and the green packets are provided with the lowest outage. The offered traffic that provides an outage of 0.05 for the green packets is extended to about 32 users per sector from 18 in the non-differentiation case (Figure 101). Therefore, the color-aware schemes provide better QoS support or capacity in a DiffServ architecture. The result also shows that the coverage control scheme provides lower outage than the rate control scheme.

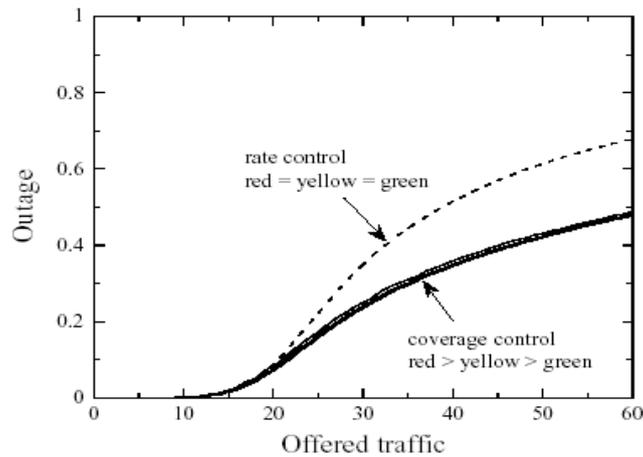


Figure 101 Outage by non-differentiated resource allocation ($k = 3$).

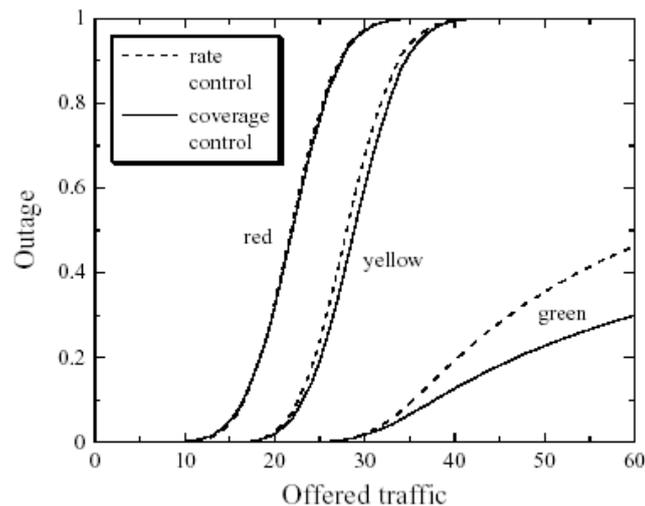


Figure 102 Differentiated outage by the color-aware rate/coverage control ($k = 3$).

2.13 HIGH SPEED DOWNLINK PACKET ACCESS

The main benefit with the high-speed downlink packet access (HSDPA) evolution of the WCDMA is that it reduces the user packet call delay and it increases the system capacity in downlink. This increased capacity can be used to either increase the number of users in each cell, and/or to provide the existing users with higher average data rates.

The HSDPA concept is based on a shorter transmission time interval, a fast retransmission with Hybrid ARQ, adaptive modulation and coding, as well as fast Node B controlled scheduling. The instantaneous data rates to a single user in HSDPA can vary between 900 kbps to above 10 Mbps. More details on the HSDPA concept can be found in [17].

It is important to note that the characteristics of HSDPA imply that there is a strong dependence between very high user data rate and radio channel quality. We must also consider the fact that HSDPA is only a downlink enhancement of WCDMA. The radio link performance can be improved by means of adding more antennas both at node B and UE in combination with advanced space-time coding and equalization in addition to the rake [35]. In scenarios with interference and severe channel conditions advanced UE receivers are required to achieve high throughput by means of high order modulation. Typically we can increase the link throughput by some 30-50% at a given cellular deployment geometry G-factor. The G-factor describes the ratio between average received power from wanted base station to the average inter cell interference power.

In [17] we simulated the HSDPA link performance in the two theoretical scenarios, using the link simulator described in [36] and [37]. We showed that the trade-off between HSDPA system throughput versus QoS is critical, and several different parameters needs to be taken into account when designing RRM strategies for HSDPA.

2.13.1 Scheduling methods for TCP traffic over HSDPA

The HSDPA scheduling is very flexible. It allows sharing of the spreading codes (here denoted as "channels"). This means that one or several channels are allocated to one user and several users may receive data under one transmission time interval (TTI). It is also possible to schedule all channels to one user under one TTI, and this is considered as the primarily means of sharing. The scheduling method used to decide which user is allocated to the channels is of great importance.

In [17] we performed simulations to test three HSDPA schedulers: Round Robin (RR), Max CQI and Proportional Fair (PF). The simulation model used when testing these schedulers is described in [38].

We showed that in a low load scenario and good radio conditions the proportional fair scheduler method gives a good trade of high average bit rate and fairness, while when increasing the packet loss probability the schedulers' performance becomes more and more similar. Already at 1% packet loss there is not much difference in average bit rate between the methods, and at 3% packet loss the performance and fairness is more or less the same. In this later case the bit rate is determined by TCP rather than by the scheduler. None of the scheduling methods can be considered as TCP friendly in the sense of helping the bit rate to recover after TCP congestion control has been activated.

We also showed in [17] that when the traffic load increase the benefit of max CQI and PF schedulers over RR becomes clear, even at packet losses. When comparing max CQI and PF, max CQI gives the best average bit rate to the users, but PF gives the highest minimum bit rate at high loads. We indicated the well known problem with max CQI, that some users might get zero bit rate (starved) or extremely low bit rate, and that one solution would be to allocate such users to dedicated channels. Moreover, we noted that PF does not maximize the minimal bit rate, so if this is the preferred QoS measure of an operator a new scheduling method ought to be developed.

2.13.2 Reference scheduling evaluations

In this section, two link adaptation algorithms are proposed and analysed. Moreover, different scheduling algorithms are studied in order to maximise the throughput. The obtained results are compared with the Round-Robin scheduler.

In order to select the appropriate modulation format and convenient transport block size, Node-B must be aware of the actual Channel Quality Indication (CQI). In [39] a relation of CQI identification and the corresponding transport block size is presented. Among the CQI values, it was selected four CQI BLER curves derived from link layer results [40], both for Indoor and Urban environments. [17] presents, for each CQI combination, the modulation, transport block size, the bit rate achieved, number of HS-PDSCH channel used, the number of codes that is used when entire sub-frame is allocated and the payload for each sub-frame.

Two link adaptation algorithms are proposed and analysed:

- a.) Maximize the user throughput. This algorithm is traduced by the following expression.

$$CQI = MAX (NbBits(CQI_i) * (1 - BLER_{Test}))$$
- b.) Allow a probability of transmission success of 90%

$$CQI = MAX (CQI_i((1 - BLER_{Test}) \geq 0.9) ,$$

Where, CQI is as defined in [39], $NbBits(CQI_i)$ is the transport block size of the CQI_i , and $BLER_{Test}$ is the predicted Block Error probability related to the CQI identification. HSDPA associates modulation and coding scheme to CQI.

2.13.2.1 Hybrid Automated Request

The HARQ type that was chosen is the partially asynchronous HARQ with Chase combining because it requires very few signaling [41].

2.13.2.2 Scheduling

Three scheduling scheme are evaluated. The schedulers aim to maximize the throughput and the results are compared with Round-Robin scheduler.

Maximum C/I

Schedules data for transmission of MS with highest C/I.

Maximum C/I over averaged C/I

Schedules data for transmission of MS with highest C/I over averaged C/I in a certain window. The window length is 10 TTI.

$$\max\left(\frac{C/I}{\text{avrg}(C/I)}\right) \quad (19)$$

The C/I is averaged over the last 10 measured C/I values.

Round Robin

MS are scheduled one after the other according to a fixed list of MS Ids.

2.13.2.3 Results

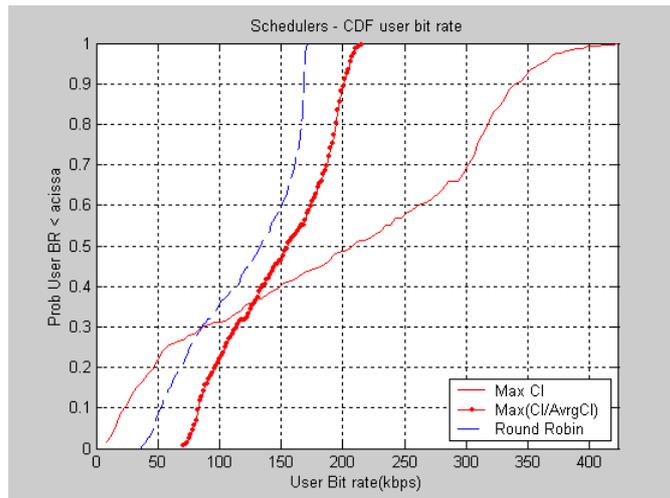


Figure 103 - CDF of user average bitrate

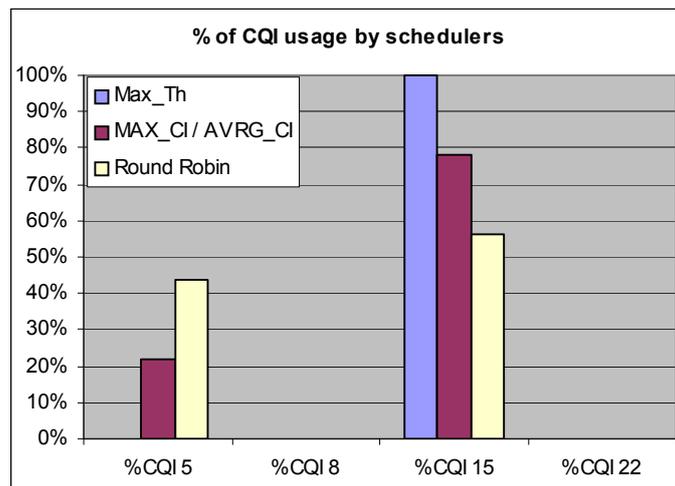


Figure 104 - Percentage of usage of each available CQI

Comparing Link Adaptation (LA) schemes

- Maximum Throughput
- Maximum CQI that allows BLER < 0.1 (Prob_success ≥ 0.9)

These simulations were performed using the maximum C/I scheduling scheme

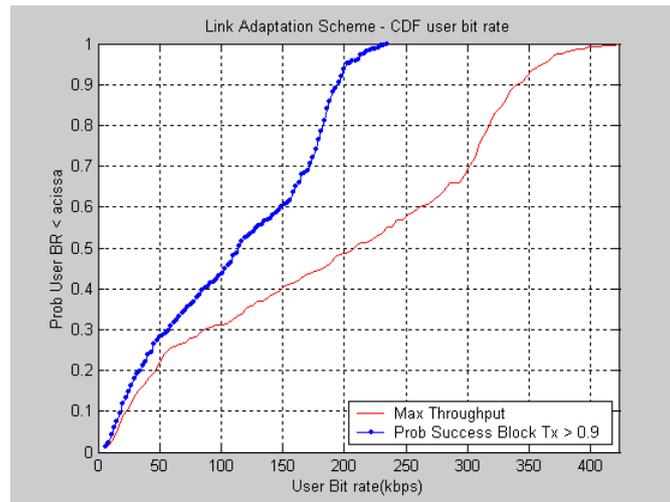


Figure 105 - CDF of user average bitrate

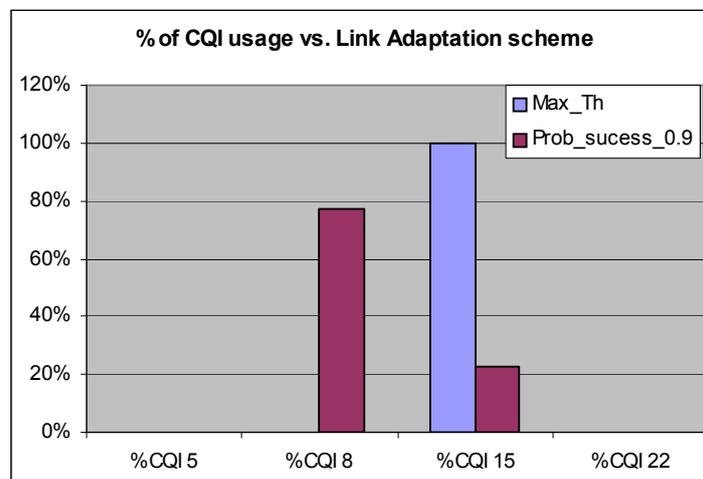


Figure 106 - CQI usage for each LA scheme

Results obtained with these schedulers shown that the Max (C/I) scheduler obtained best throughput results. With the available CQIs the system achieves the service capacity of near 4 Mbps. Max (C/I over average C/I) and Round robin achieve 3Mbps and 2.5Mbps of service throughput respectively. CDF of user average bit rate shown that Max (C/I over average(C/I)) have most equilibrated values, with bit rates varying from about 70 kbps to 215.kbps. Max (C/I) obtained average bit rate varying from about 5kbps to 430Kbps.

Concerning BLER, Max C/I scheduler obtained average value bellow 3E-1. Worst average BLER in cell was obtained by the Round Robin, with 4E-1.

Concerning the average attempt Tx. per block, best value was obtained was: max (C/I) 1.4, 1.6 and above 1.8 for the RR.

Evaluating the CQI usage, the CQI 15 was used for all transmissions for the Max(C/I) and used in most cases for Max (C/I over average(C/I)) – 80%of times, and Round Robin, near 60%. Note that CQI 22 is never used since its BLER curves does not allows good throughput. Should be also noted that the use of the CQIs that does use all codes for all sub-frame allocated to one MS is not good policy since there will exist codes that is not used. For this reason, CQI 5 (use the 15 codes over 15 available) achieve better throughput than CQI 8 (use max 14 codes over 15 available).

Comparing the two link adaptation scheme, one that will aim to provide best system throughput the other that selects higher CQI that provide probability of correct transmission of 0.9. Results shown that, in fact, best throughput result is obtained with the 'Maximum throughput' link adaptation scheme. However this scheme has associated higher BLER. With probability correct transmission of 0.9, less block error probability and less transmission attempt per block is obtained. Less transmission attempt will leads to less block delivery delay.

2.13.3 Advanced scheduling evaluations

2.13.3.1 Introduction

The advanced HSDPA RRM is an upgrade of the reference HSDPA RRM presented in section 2.13.2. The proposed advanced HSDPA RRM contains all the specific entities pertaining to generic resource allocation at the MAC layer, however additional functionality has been added and includes a support for quality of service (QoS) in packet oriented system, while trying to ensure throughput optimisation over the UMTS air interface. In particular, the RRM considered contains the following functional sub-blocks: the Scheduler, User Quality Tracking using Signal to Interference (plus Noise) Ratio (SIR) reported by each mobile, and Hybrid ARQ (HARQ) with Chase Combining.

The advanced RRM addresses specifically the design of the scheduler. In this RRM, specific schedulers in Near Real Time video (video streaming) service were implemented to support quality of service, exploring the Layer 3 while trying to maximize the cell throughput, exploring the PHY layer to provide balanced scheduling. The support of quality of service is achieved by three schedulers prioritizing packets concerning single packet delay, average delay of packets in queue of each user, and queue size in bits. FTP traffic although proposed in section 2.13.2, is not the most appropriate to be evaluated with the proposed scheduling algorithms since is a background service. The Max CI scheduler (presented and evaluated in section 2.13.2) can be the indicated scheduler for the FTP traffic. For that reason the schedulers proposed for advanced studies are evaluated with NRTV service.

Proposed schedulers were evaluated and the results are presented in terms of satisfied users and cell throughput both for indoor and vehicular environments. In particular, the algorithm is supposed to provide QoS support to delay constraint service by assigning a priority value to the packets. Packet prioritization is based on weighting specific metrics pertaining to the Predicted Reliability in transmission, the Delay that packet experiences in the queue and the attempted packet transmissions, the latter metric is related to the Automated Repeat Request (ARQ) stop and wait protocol which is included in the reference RRM in section 2.13.2.

2.13.3.2 Simulation assumptions

Some assumptions were made for the performance evaluation of the advanced HSDPA RRM. This RRM concerns the CQI and Look-Up Tables (LUT) used for reference HSDPA described in detail in [17].

2.13.3.2.1 Service and traffic load

The simulations were performed with Near Real Time Video (NRTV). The traffic load, in terms of number of users of each service, was selected taking into account the system capacity. The simulations were performed with 20, 25, 30 and 35 users. The model used to generate packet traffic is specified in [42].

2.13.3.2.2 Link adaptation

The Link adaptation algorithm for the selection of suitable CQI to be used in transmission of each packet is based on the highest CQI that corresponds to BLER values not higher than 0.1 [1]. The algorithm is traduced by the following expression:

$$CQI = \underset{i}{MAX} (\arg(BLER(CQI_i) \leq 0.1)) \quad (20)$$

2.13.3.2.3 Packet size and link adaptation

The RRM model supports packet traffic with variable packet sizes and scheduling, and transmissions are oriented to packets. Since this RRM is addressed to packet transmission, it supposes the integration of an adapter that maps packets of variable size into variable length radio blocks (which depends on the CQI). For such a system the following actions are performed by the scheduler in order to map packets into radio blocks through ARQ processes:

1. CQI is selected according the link adaptation algorithm;
2. Packet is prioritized according the prioritization scheduling algorithm;
3. Scheduler calculates number of Transport Blocks to transmit the packet. The number of blocks is calculated according to the expression (21)

$$NBlocks = \left\lceil \frac{PacketSize}{BlockSize(MCS_i)} \right\rceil \quad (21)$$

where $\lceil x \rceil$ means lowest integer higher or equal to x ;

4. Idle ARQ channel j is selected to hold packets and the number of allocated Transport Blocks;

Figure 107 illustrates an example of mapping three packets into radio blocks. ARQ holds the packet and blocks that will transport the packet.

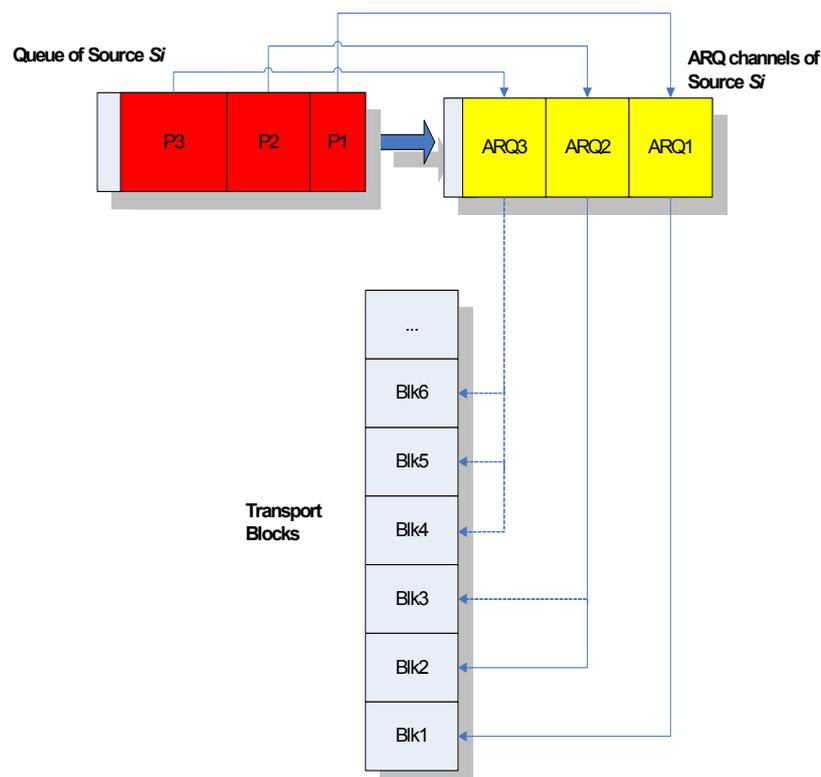


Figure 107. Mapping packets in radio blocks.

The proposed RRM is characterized by simplicity and requires minimal upper layer signalization concerning packet deliver, since ARQ protocol will be responsible for it. However it has trade-offs of some waist of useful bandwidth since it requires inclusion of padding bits, as illustrated in Figure 108.

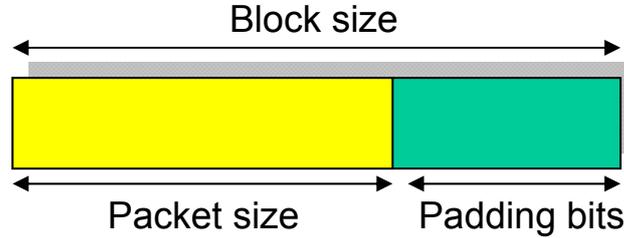


Figure 108. Block holding packet data

2.13.3.2.4 Schedulers

The scheduler algorithms were proposed for the video streaming and are based on weighting parameters that affect the application functionality and the system (cell) throughput. Complementary to the scheduler, the RRM allocation policy tries to maintain flexibility by assigning several users to a single TTI, scheduling individual packets.

Proposed scheduling algorithms combine reliability in packet transmission of each user with upper layer service requirements. Three scheduling algorithms were proposed, each one weighting the reliability in transmission, service requirements and number of attempted transmissions.

A priority value for each algorithm ($Priority_1$, $Priority_2$, and $Priority_3$) will be calculated every TTI combining W_1 , W_2 and W_3 , where W_1 , W_2 and W_3 concern respectively to the reliability on block transmission, the service delay requirement and attempted transmissions. The analytical expressions of each scheduling algorithm are presented by (22), (23) and (24).

$$Priority_1 = W_1(CQI, SIR)(W_2(packetTimeOut) + W_3(\#attempTx)) \quad (22)$$

$$Priority_2 = W_1(CQI, SIR)(W_2(queueAvrgTimeOut) + W_3(\#attempTx)) \quad (23)$$

$$Priority_3 = W_1(CQI, SIR)(W_2(queueSize) + W_3(\#attempTx)) \quad (24)$$

$Priority_2$ differs from $Priority_1$ by the fact that instead of actual individual packet delay value, the average delay of all packets in queue of one user is used. For the $Priority_3$ instead of delay, the queue size in bits is used to prioritize the packet. The details of W_1 , W_2 and W_3 are present in next subsection.

The expected reliability is obtained with the value of the CQI, which reflects the SIR that the mobile experiences. For the transmission reliability three values are associated with the weight function: transmission with low probability of correct packet reception, transmission with reasonable probability of packet reception and transmission with high probability of correct packet reception. To distinguish between the three cases, the SIR target, and threshold values are assigned accordingly depending on the service requirements. The W_1 function in relation to the SIR is given by Figure 109.

$$W_1(CQI, SIR) = \begin{cases} \sim 0 & \text{if } SIR < Trget \\ 1 & \text{if } Target < SIR < Trget + Threshold \\ 2 & \text{if } SIR > Trget + Threshold \end{cases} \quad (25)$$

Where *Target* and *Threshold* are functions of the video streaming service.

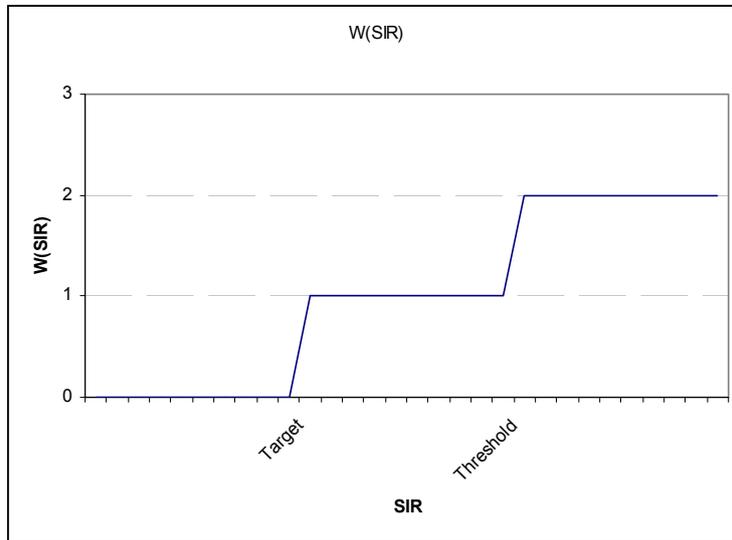


Figure 109: Reliability function

The time-out function is shown by Equation (26), and illustrated in Figure 110. A value for the constant *k* is obtained, assuming that when time-out is reached the function will have the same value of as the reliability function for *SIR > Target + Threshold*.

$$W_2(TimeOut) = k(AllowableDelay - time_out) \quad (26)$$

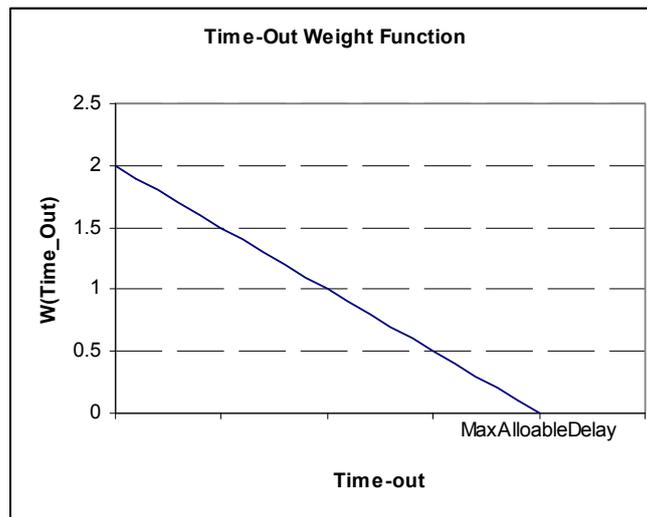


Figure 110: Scheduling delay function

The queue length function is shown by Equation (27), and illustrated in Figure 111. A value for the constant *k* is obtained, assuming that when maximum size of the queue/buffer is reached, the *W₂* function will have the same value of as the reliability function for *SIR > Target + Threshold*.

$$W_2(queueSize) = k(QueueSize) \tag{27}$$

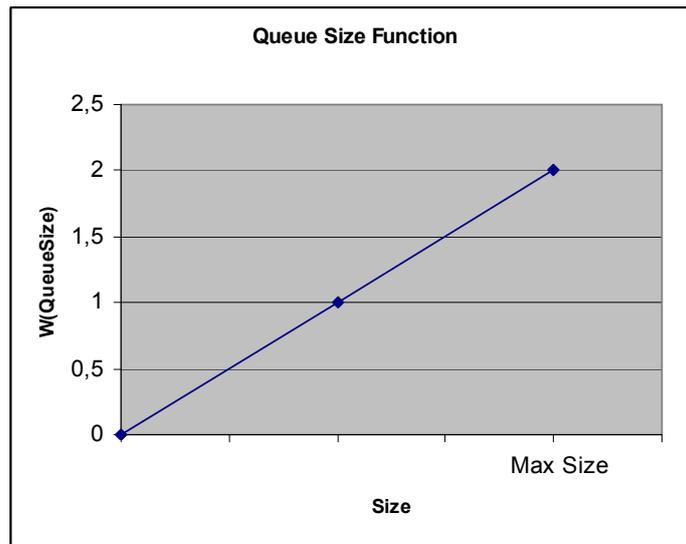


Figure 111: Queue size function

The Attempted transmission function increases with the number of transmissions that the packet has experienced. The maximum number of transmissions will help to minimize the BLER results. After the third attempt the packet is discarded. The values for the attempted transmissions are presented in (28) and graphically depicted by Figure 112 for a maximum number of transmission equals to three. Three is the number for the maximum transmission attempts. However, other values will also be explored.

$$W_3(AttemptTx) = \begin{cases} 0 & \text{if } Attempt\ Tx = 0 \\ 1 & \text{if } Attempt\ Tx = 1 \\ 2 & \text{if } Attempt\ Tx = 2 \end{cases} \tag{28}$$

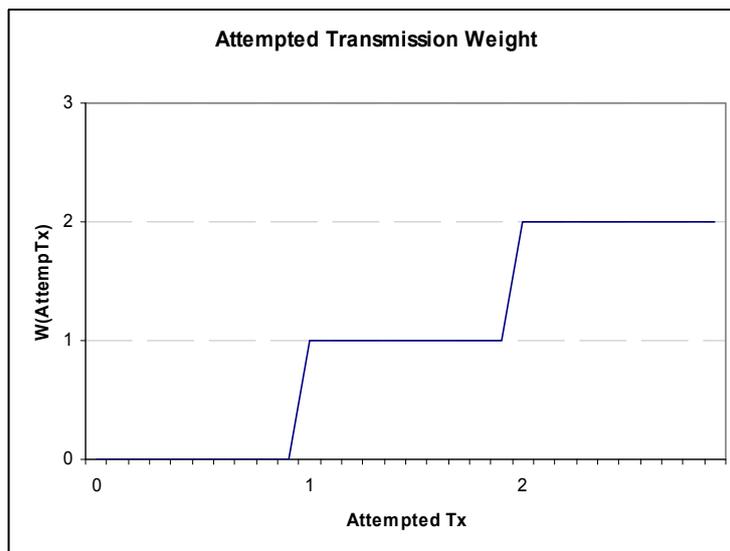


Figure 112: Attempted transmissions function

2.13.3.2.5 Scenario and simulation mode

The system will be evaluated in indoor and vehicular scenarios at 3km/h and 50 km/h respectively.

The duration of each run corresponds to 300 seconds of real time. The Mobiles are created at the beginning of each run, and remain active for the complete run duration. The traffic start generation instant is random which means that all users are not transmitting immediately as simulation started.

Most relevant simulation parameters concerning the indoor and vehicular scenarios are presented in the Table 23 and Table 24 respectively.

Table 23 Simulation parameters for the indoor scenario

Environment	Indoor
Cell type	Omni
Cell radius	10 m
Mobiles velocity	3km/h
Channel Model – Path loss and shadowing	3GPP indoor channel
Channel Model – fast fading	ETSI Indoor A
Services	NRTV and FTP. There is no mixing of different traffic types within a single simulation
Number users (NRTV)	20, 25, 30 and 35 users
TTI duration	2ms (corresponds to frame duration)

Table 24. Simulation parameters for the vehicular scenario

Environment	Vehicular
Cell type	Sectorized
Cell radius	300 m (per sector)
Mobiles velocity	3km/h
Channel Model – Path loss and shadowing	3GPP urban channel
Channel Model – fast fading	ETSI Vehicular A
Services	NRTV and FTP. There is no mixing of different traffic types within a single simulation
Number users (NRTV)	20, 25, 30 and 35 users
Frame duration	2ms (corresponds to frame duration)

2.13.3.3 Performance metrics

The performance metrics used are related both with individual user and overall system performance. For individual users the metric is the Satisfied Users and for the overall system is the System Throughput.

2.13.3.3.1 Satisfied users

For the individual user, in NRTV service, the performance metric is the satisfaction in terms of packet average delay and session bit rate.

Target values of delay and session bit rate for NRTV is presented bellow:

- Maximum average packet delay for NRTV users: 300ms. The delay of each packet concerns the time since that packet arrives to the queue of the base station until received correctly by the MAC layer of the mobile station.
- The session bit rate is equivalent to the NRTV service requirements and is 64 kbps.

2.13.3.3.2 System throughput

The system throughput is averaged over a single cell and is evaluated in terms of:

- Over the Air throughput (OTA)

It is the number of bits that have been transmitted by the given cell during the simulation duration divided by the total duration during which the cell has been transmitting.

$$R = \frac{b_{OTA}}{k \cdot T} \quad (29)$$

- Service throughput

It is the number of bits that have been transmitted correctly received in the cell during the simulation duration divided by the total simulation duration.

$$R = \frac{b_{service}}{k \cdot T} \quad (30)$$

- QoS throughput

It is the number of bits correctly received within allowed delay in the cell during the simulation duration divided by the total simulation duration.

$$R = \frac{b_{QoS}}{k \cdot T} \quad (31)$$

2.13.3.4 Simulation results

The advanced RRM performance is evaluated comparing the results obtained with proposed schedulers presented in previous section with maximum channel over interference scheduler (Max. CI), which tends to optimize system throughput scheduling users experiencing best channel characteristics. However, this is done without regarding to service requirements.

The main results focus the comparison of the priority P_1 scheduler and Max CI scheduler. The behavior of weight priority function w_2 (using packet delay, average delay of packets in queue and queue size) is analyzed comparing the schedulers P_1 , P_2 and P_3 only in the indoor scenario.

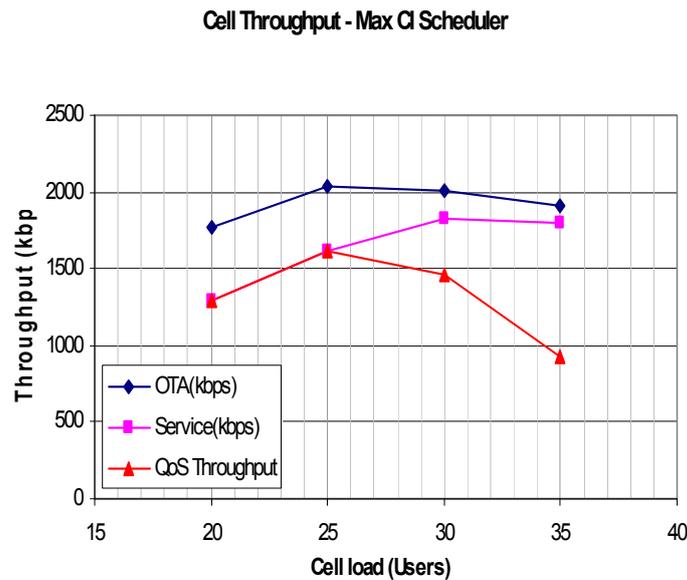


Figure 113. Average cell throughput for Max. CI scheduler

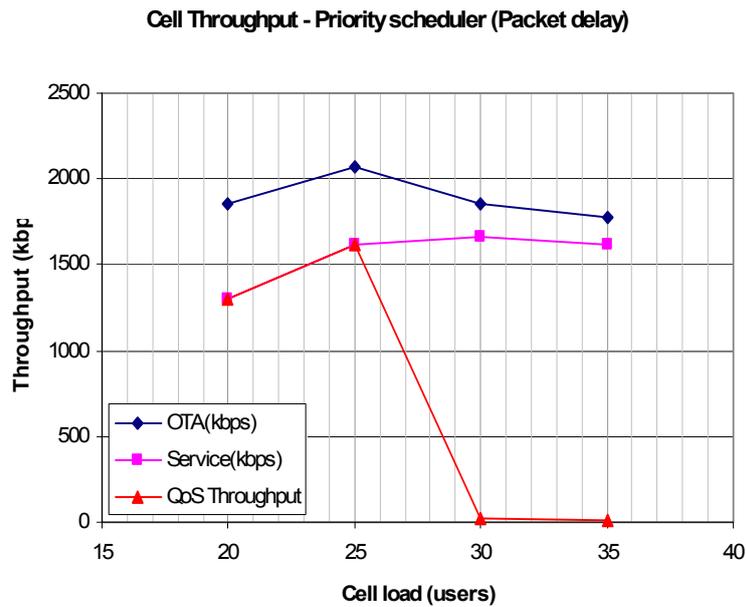


Figure 114. Average cell throughput for Priority standard scheduler (based on packet delay)

The results shown that better throughput results are obtained with Max. CI scheduler. The optimal number of users seems to be, for each scheduler, around 25 users.

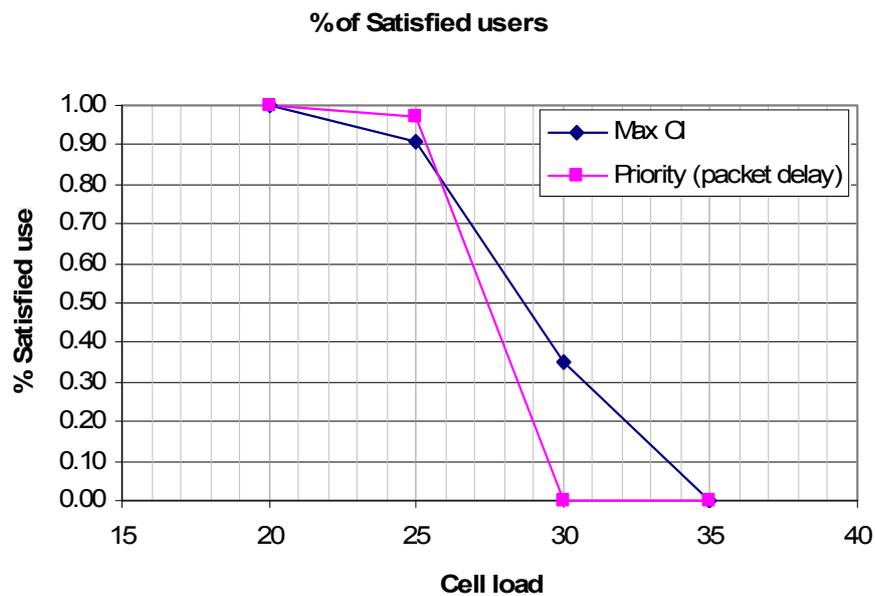


Figure 115. Ratio of satisfied users (Max CI and Priority standard schedulers)

If we appoint a target of 90% of satisfied users results shows that Priority Standard scheduler lead to better results concerning the percentage of satisfied users.

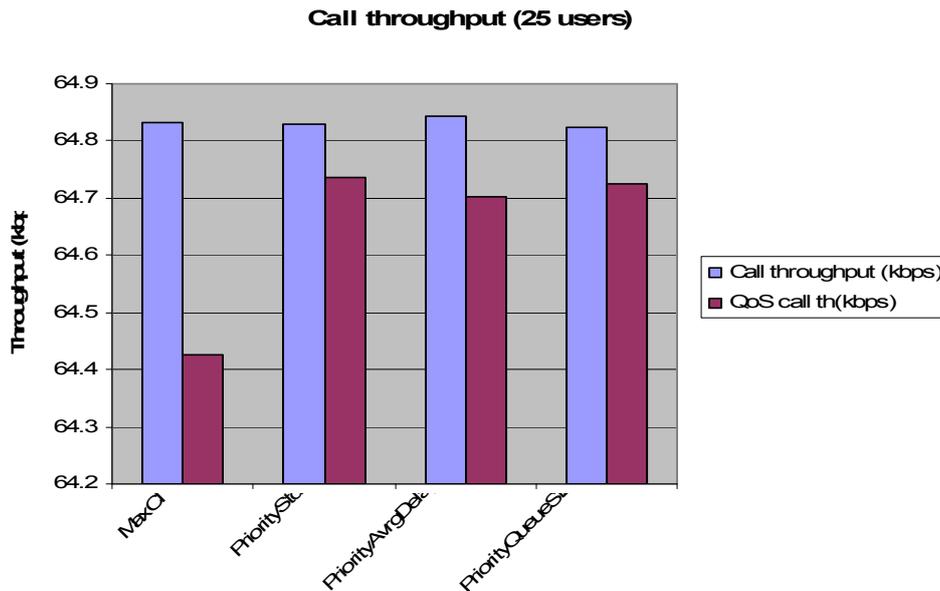


Figure 116. User average bit rate – Comparing all schedulers.

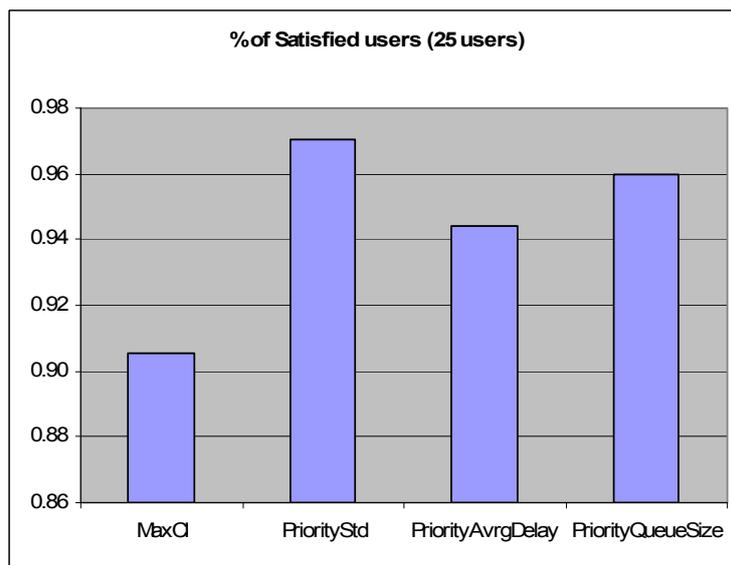
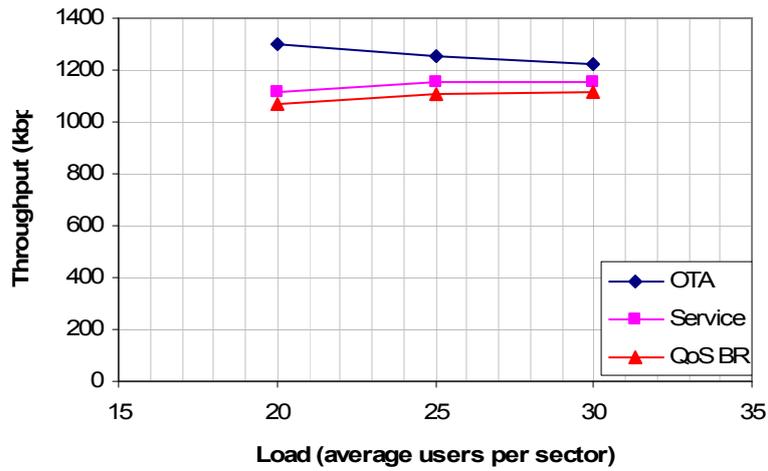


Figure 117. Percentage of satisfied users – Comparing all schedulers.

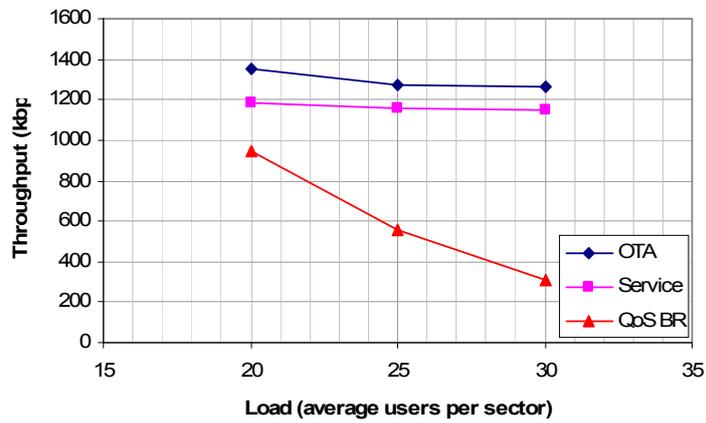
Comparing the schedulers, concerning QoS, better results both for QoS bit rate and satisfied users, were obtained with the Priority Standard scheduler, P_1 . The Max. CI scheduler is the one that lead to worst QoS results, since it does not consider services constrains.

Simulation parameters concerning vehicular scenario are presented in Table 24. Vehicular scenario comports sectorised cells with deterministic number of users in site, but averaged in each sectorised cell.

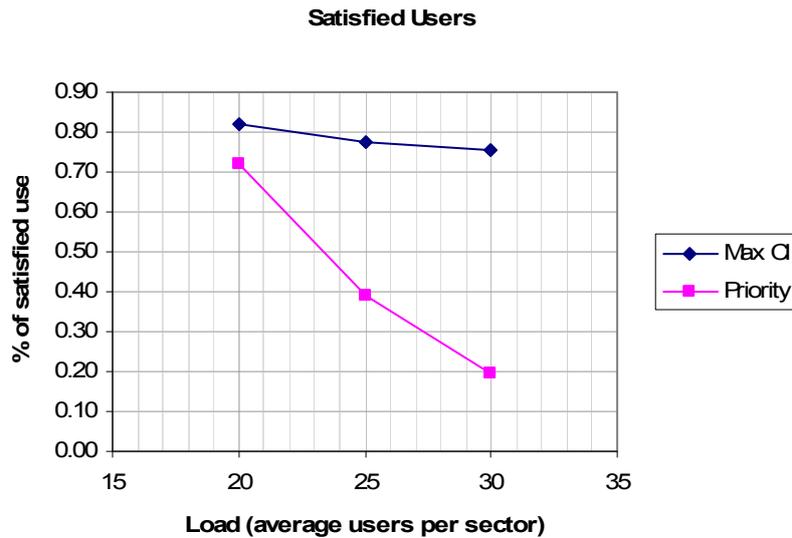
Average Sector Throughput (kbps) - Max CI Scheduler



Average Sector Throughput (kbps) - Priority Scheduler



The results shown that better QoS throughput results are obtained with Max. CI scheduler. The behavior is similar to the one in indoor scenario. In the other hand, lower QoS throughputs are obtained within vehicular scenario with both schedulers. The main reason is because the number of users in cell is averaged in vehicular scenario and for indoor this number is deterministic.



Concerning the number of satisfied users, results shown that worst performance were obtained with the priority scheduler. Results shown lower than 20 users to obtain 90% of satisfied users compared to the 25 users in the indoor scenario.

2.13.3.5 Summary

The advanced HSDPA studies concerns a RRM scheme based on scheduler that combines both channel information and service requirements. The proposed RRM was evaluated both in terms of individual user parameters (satisfied user), and cell throughput. In packet traffic, and using NRTV service, overall obtained results shown that best QoS performance were obtained in the Indoor scenario. The results shown that although Best SIR based algorithm lead to best system throughput, it is not the fairest algorithm especially if delay is one of the requirements that needs to be kept within a tolerable delay window. Overall results, which combine both satisfied user parameters and throughput, shows that the prioritization algorithm based on channel state and service requirements is a more suitable mechanism for scheduling within a system than the conventional "Best SIR algorithm". It is important to refer that Best SIR algorithm, which is the principle proposed for the HSDPA mode of the UMTS is more suitable where cell throughput is the unique requirements, for example for background applications such as ftp or email.

It is important to refer that the results were obtained without any kind of CAC. System with admission of homogeneous users (in terms of channel quality) will improve the throughput of the proposed algorithm. The CAC will decide how to attain the correct trade-off between dropped, and blocked users in the system, for a given traffic load. At is often more desirable to be blocked, than to be dropped, then the CAC must have information surrounding the CQI for the admitted user, and the current mean packet delay experienced in the buffers. Given this information, the CAC will decide whether there are sufficient resources to satisfy the service requirements of the admitted user. If the CAC is biased towards channel strength readings, if the channel quality is good, then it may improve the service throughput of the system, however at the expense of increasing the mean packet delay for the remaining users.

2.13.4 Performance Enhancement for HSDPA

2.13.4.1 Biased Adaptive Modulation/Coding to Provide VoIP QoS over HSDPA

In [17], we have reported the feasibility study on supporting VoIP over HSDPA by using biased AMC scheme. That study has clearly shown the improvement in capacity brought by biased AMC techniques. In our further research on this issue, we observed that the channel estimate is less reliable when a user is in poorer channel conditions. Since the link adaptation in HSDPA relies heavily on the channel estimation, this would degrade the general performance of HSDPA and also affects the improvement brought by the adaptive biased AMC. Therefore, we refined the adaptive biased AMC techniques by enhancing its capability on compensating the channel estimation errors. This is done by adjusting the AMC according to the appropriate bias suggested by a statistical analysis on the channel estimation errors. Since the channel estimation is also closely related to the mobility issues, we also examining the effects of mobility on the proposed biased AMC. Since VoIP traffic is quite different from traditional service carried by HSDPA, new requirement is also posed on the scheduler. We have study the performance of the proposed biased AMC with the PF-LDF scheduler rather than just PF scheduler as presented in [17] section 2.3.4.2. With taking consideration of mobility and new scheduler, the effects of RTP block size on the VoIP performance is also studied.

The protocol stack depicted in Figure 124 has been detailed presented in [17] section 2.3.4.2. Here we still use it to support this study. The voice data that arrive to the VoIP gateway via the PSTN are translated into a low rate codec suitable for UTRAN, and delivered through the IP backbone to the radio network controller (RNC). Note that the codec is not necessarily AMR. At the RNC VoIP packets are segmented into RLC packets. These packets are carried as a single MAC-d flow by the lub/lur frame protocol (FP). The FP conducts a flow control over lub/lur so that the NodeB buffer does not starve nor overflow. At NodeB the MAC-hs schedules the MAC-d PDUs to transmit on the shared physical channel (HS-PDSCH) and conducts AMC and HARQ. Since the residual error rate after the HARQ is very low, the ARQ functionality of the RLC, which will impose an extra delay, is likely to prove futile, if not negative. Thus, the unacknowledge mode (UM) shall be used on the RLC. To reduce the overhead, the packet data convergence protocol (PDCP) is used to compress the IP header.

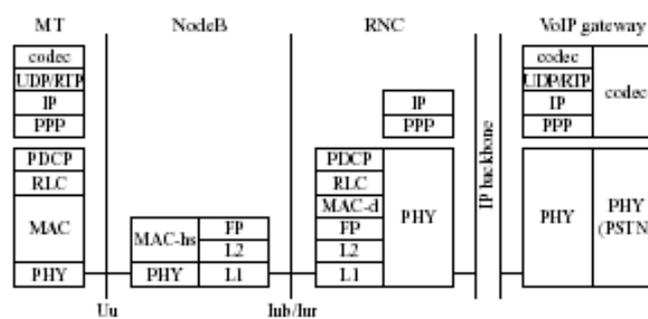


Figure 118 Protocol stack for VoIP over HSDPA.

2.13.4.1.1 HSDPA Packet Scheduler for VoIP

The fast scheduling, conducted by the MAC-hs, is a key feature to determine the overall system performance of HSDPA. The MAC-hs schedules the MAC-d PDUs as well as HARQ retransmissions to serve on the HS-PDSCH. The channel quality indicator (CQI) fed back from the mobile terminals may be utilised in the scheduling process. Various schedulers have been studied in the literature, mostly under the best effort discipline to exploit the multiuser diversity, with attempts to add a degree of fairness. A scheduler that chooses the

user with the best channel quality (known as a MaxC/I scheduler) maximizes the system throughput, however, with a major deficiency on fairness. A proportionally fair (PF) scheduler [43] exhibits a comparable system throughput with much more tolerable fairness [44], [45].

A number of modifications have been proposed to fit the PF scheduler for streaming services. The main aim of these attempts are to guarantee a lower delay jitter, so that a compact dejittering buffer enables the streaming QoS. In [46] a simple but effective method to calculate the priority metric is presented, in which the head of line delay is multiplied to the PF priority metric. Simulation studies in [47] have shown that the delay sensitive PF improves the streaming performance over HSDPA. A similar discipline is viable to schedule VoIP traffic.

$$p_i = \lambda_i - f \cdot \bar{\lambda}_i + 10 \log_{10}(\tau_i/T), \quad (32)$$

where λ_i and $\bar{\lambda}_i$ are the newest reported CQI and its time average for the i -th user respectively, τ_i is the head of linedelay for the i -th user, and T is the packet due time. The third term introduces a longest delay first (LDF) factor into the priority metric. The parameter f controls the fairness among the users; $f = 0$ behaves as MaxC/I with an added LDF factor, $f = 1$ is equivalent to PF (fair time for i.i.d. channels) combined with LDF, and $f > 1$ gives more frames to inferior users to even out throughputs. The HS-PDSCH, on which the user data is conveyed, has a frame size of 2 ms. Hence, if we assume a due time T of 100 ms, only 50 frames are provided before a packet expires without code multiplexing. This implies that that HSDPA accomodates a maximum of mere 50 flows (assuming continuous conversations without a voice activity detection). This number is further decreased by HARQ retransmissions. We must also budget for call blocking. Since a single VoIP flow requires only a fraction of the total HSDPA bandwidth, code multiplexing is essential to serve more VoIP flows. Consequently, we multiplex users on each frame by waterfilling available power and code resources. Nevertheless, since a shared control channel (HS-SCCH) is necessary per scheduled user to indicate the data presence on the HS-PDSCH, the HS-SCCH provision limits code multiplexing. The 3GPP specification [48] mandates UEs to monitor four HS-SCCHs simultaneously. Hence, if more than four HS-SCCHs are used, each user must be mapped carefully to a set of HS-SCCHs, as to minimize the inherent partition loss. Hence, we limit up to four users per frame in the sequel.

Figure 115 compares the unsatisfied user ratio (outage probability, hereafter) as a function of the offered traffic load for various schedulers, i.e., round robin (RR), proportional fairness (PF), and combined PF and LDF (PF-LDF) with fairness parameter $f = 1$ and 2. The results were obtained from dynamic system level simulations (see [37] for more details). Note that a user is assumed to be "satisfied" if more than 99 percent of the packets are received correctly within a due time T . As the traffic load is increased, the outage probability increases due to longer queues, with the RR scheduler being the worst. However, the difference is insignificant between the schedulers. This is due to the stringent delay requirement that refrains PF schedulers from effectively exploiting the multiuser diversity. Figure 119 shows that the delay-awareness slightly reduces the outage probability at high loads. However, the impact of the fairness parameter f is negligible between one and two. The results show that VoIP over HSDPA is feasible, producing an outage probability of about 0.05 at 70 Erlangs.

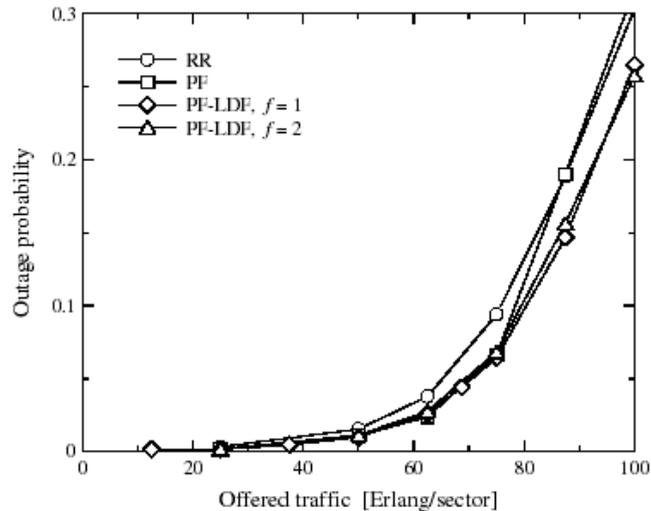


Figure 119 Ratios of unsatisfied users as a function of the offered traffic load with various schedulers

2.13.4.1.2 Impact of block size

VoIP over HSDPA is to offer elastic radio access to various applications, by supporting various speech formats. However, key parameters such as the data rate and block size (the codec block size or the RTP packet size) affect the overall system performance. A larger block size reduces the protocol overhead and the processing load on backbone routers. However, a larger block size inherits a longer encoding/packetization delay. Hence, the block size yields a tradeoff. We compared the block sizes of 20, 40, 60, 80, and 100 ms for 12.2 kb/s AMR, which generates octet aligned RTP packets of 33, 63, 94, 124, and 155 bytes respectively, including a 2 byte compressed IP header. This can be viewed as different codec block sizes, or as packetizing a different number of 20 ms AMR blocks into a single RTP packet. The packet due time is adjusted accordingly: 100, 80, 60, 40, and 20 ms respectively for a 20, 40, 60, 80, and 100 ms block size.

Figure 120 shows the outage probability for various block sizes. We immediately notice that a block size of 100 ms is unbearable, whereas 20 to 80 ms exhibits similar performance with an outage probability less than 0.05 up to about 70 Erlangs. However, if we look closely, the 60 ms block size is the most favourable, realizing a low outage probability (< 0.05) up to about 80 Erlangs. As the block size gets larger than 60 ms, the outage probability rises due to shorter deadlines. The deadline of the 100 ms block size is simply too severe. On the other hand, the slight degradation with a smaller block size is associated with the lack of radio resources, i.e., frames, codes, and power. A smaller block size generates more RTP packets per unit time. These packets arriving to NodeB in rapid succession increases the probability of each radio frame being transmitted with few data (using lower MCS with less packing and spectral efficiency). These lightly loaded frames occupied a considerable number of radio frames, while lengthening the queue and causing outages.

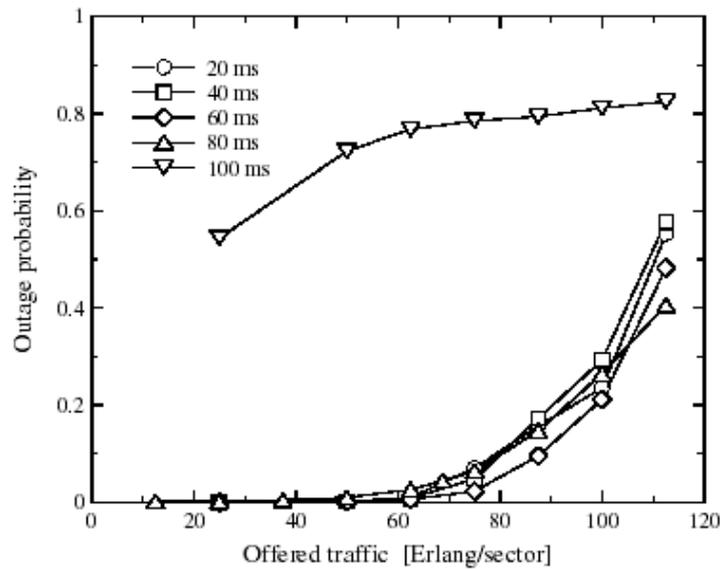


Figure 120 Impact of the AMR/RTP block size on the outage probability

Figure 121 supports the above discussion. A smaller blocksize increases the average number of code multiplexed users per frame. With a 20 ms block size, the average number of multiplexed users increases rapidly as the traffic load increases, and soon approaches the four-user limit. This implies that the radio frames often ran short, resulting in outage rise (Figure 120). The dropping number of multiplexed users at higher loads is a result of code and power exhaustion; associated dedicated physical channels (A-DPCHs) and HS-SCCHs took a large fraction of these resources. This phenomenon is unseen on a larger block size.

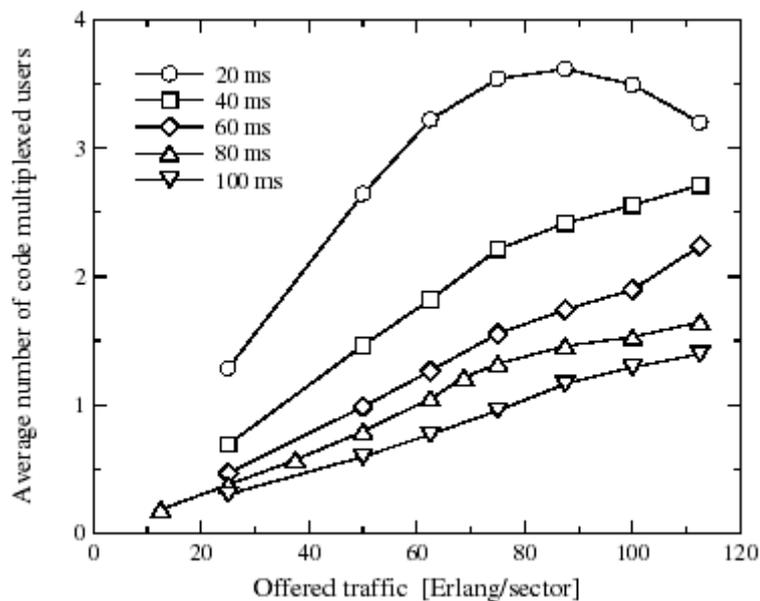


Figure 121 A smaller block size increases the number of code multiplexed users per frame

To elaborate on the discussion, Figure 122 compares the average MCS level used for various block sizes. A larger block size uses a higher MCS level, average of which decreases slightly as the traffic load increases. This is due to water-filling; users behind the queue try to transmit as much data as possible utilising residual resources. With a larger block size, a user is more likely to wait for a new packet arrival once the user is allocated a frame to transmit using a high MCS level. The increased number of starved buffers reduces

the scheduler's choices of users, and the LDF factor in the priority metric grows dominant, thereby reducing the multiuser diversity effect. In contrast, with a 20 ms block size, the average MCS level increases due to frequent packet arrivals; more users have packets in the transmitter buffer and the scheduler is able to exploit multiuser diversity. The block size imposes another tradeoff. The mobile terminal has to decode a higher (and more complex) MCS level once in a while, or has to decode a lower MCS frequently.

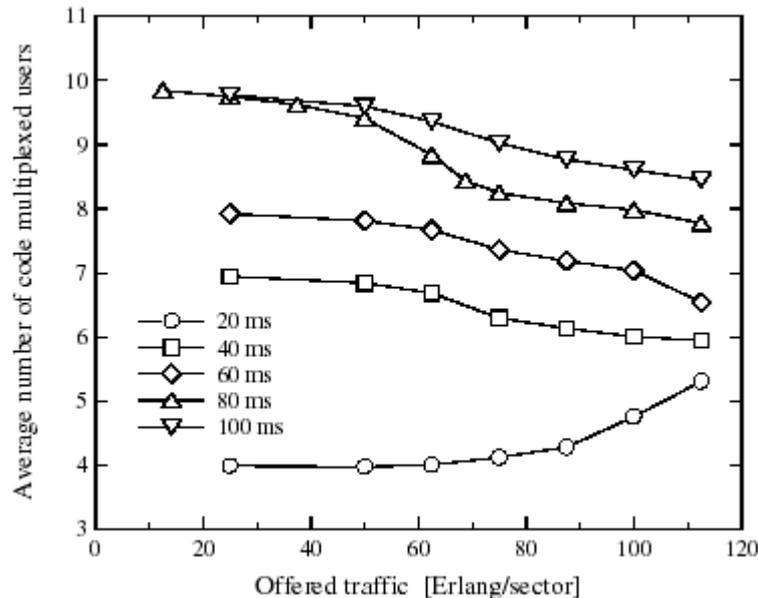


Figure 122 Average MCS levels for various block size.

We have shown that HSDPA is versatile to support various block sizes. Nevertheless, since a larger RTP packet reduces the protocol overhead and the processing load at routers, a larger block size is desirable from the IP backbone perspective. A block size of 60 to 80 ms is a potential choice.

2.13.4.1.3 Tolerance to Mobile Speed

Figure 123 shows the impact of mobility on the outage probability. As the mobile velocity increases, the outage probability effect at velocities as low as 3 km/h, and requires larger powers increases. This is due to the CQI feedback delay. At higher velocities the CQI reports arrive too late to track the fast fading channel. Consequently, the scheduler, as well as the AMC, has to rely on erroneous CQIs. At a velocity of 80 km/h, the outage probability exceeds 0.05 even at a low traffic load of 20 Erlangs. Therefore, VoIP delivering over HSDPA seems feasible only at low velocities (up to 20 km/h). However, as we shall show in the following section, the outage probability at high velocities can be reduced remarkably by adjusting the AMC.

To mitigate the fast fading, HSDPA utilises fast scheduling and AMC. In contrast the DCH employs fast transmission power control and symbol interleaving over a TTI (typically 20 ms) to counter act the fast fading. This makes the conventional (3GPP Release 99) DCH speech more robust to vehicular speeds. However, the DCH loses the interleaving effect at velocities as low as 3 km/h, and requires larger powers to satisfy the error rate. The lower outage probability at a lower velocity as seen in Figure 123 is hence peculiar to HSDPA.

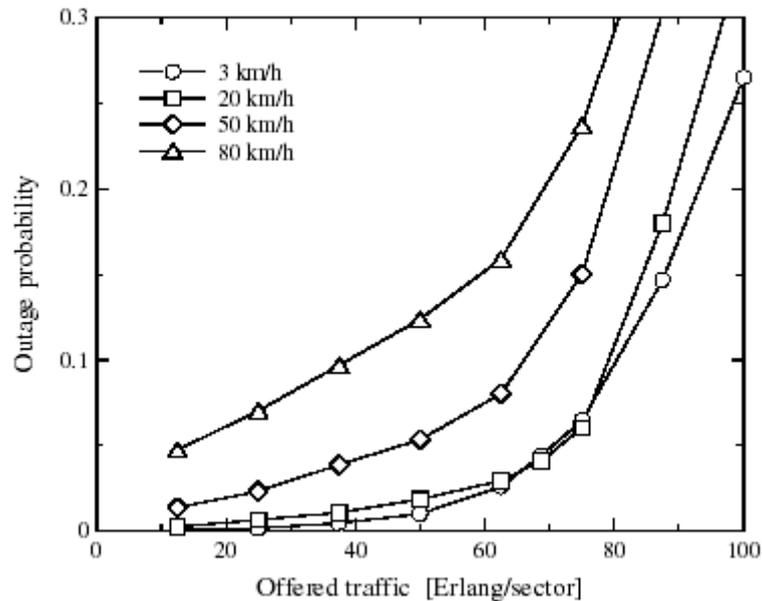


Figure 123 Impact of mobility on outage probability (80ms block size)

2.13.4.1.4 Biased AMC scheme

The QoS degradation at high speeds limits deployment scenarios of VoIP over HSDPA. According to [48] the AMC is to be performed such that the frame error rate (FER) is expected to be around 0.1. However, relying on erroneous CQIs, the AMC operates at a much higher FER especially at high vehicular speeds, thereby causing more HARQ retransmissions and subsequent delay. To mitigate this problem, in [17], we developed a biased AMC scheme, in which the channel estimation error is compensated at a cost of extra power. That is, we transmit each MCS with a larger power to reduce HARQ retransmissions, provided the power is available at NodeB. If the power is unavailable, we simply lower the MCS by an equivalent amount. Figure 124 illustrates the MCS selection under this rule. The MCS levels of HSDPA are designed such that the received symbol energy to interference power density ratio (E_s/I_0) to achieve a BLER of 0.1 is about 1 dB larger than the next lower MCS level. In Figure 124 BLER curves for five adjacent MCS levels are shown. Assuming the MCS that corresponds to the reported CQI is level three, the point at "O" represents the original operative point. If we bias the power by 1 dB, the MCS level three is still used with, however, 1 dB higher power, provided the power is available. The working point is thus moved to "A," yielding a lower BLER. If the total power runs short, the MCS is degraded to level two, thereby operating at "B." A bias of 2 dB shifts the point to "C" if the power is available, and to either "D" or "E" if not, depending on the available power. Thus, this biased adaptive Modulation/Codingscheme reduces the delay of VoIP packet through reducing the number of HARQ retransmission, but consumes an extra fraction of the limited total power. Table 25 shows the capacity gain obtained by the proposed scheme. Clearly with the trade-off between the power consumption and reduction of HARQ retransmission, a best offset value to maximum the capacity is 3dB.

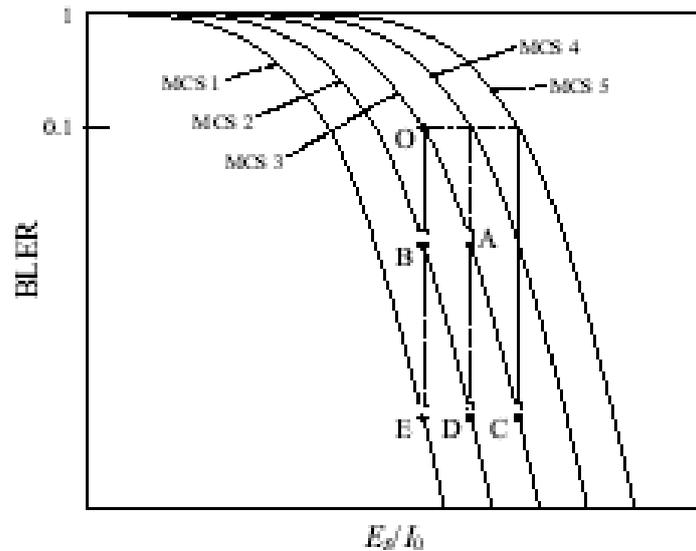


Figure 124 Biased MCS selection.

Table 25 Capacity gain by the biased AMC

Power offset [dB]	Relative capacity
0 dB	1.00
1 dB	1.15
2 dB	1.29
3 dB	1.32
4 dB	1.21

2.13.4.1.5 Derivation of required AMC bias

In HSDPA the CQI is obtained by measuring the SIR of the common pilot channel (CPICH). However, the SIR estimate is less accurate when the true SIR is low. This implies that the CQI reports from mobile terminals near the cell boundary are less reliable. Moreover, a higher mobile velocity reduces the CQI reliability due to feedback delay. Hence, to find the appropriate power offset, we examine the statistics of the SIR estimation error, considering the channel correlation at the time of CQI estimation and application on AMC, and the user location.

Because of the logarithmic nature of the SIR, we consider SIR in decibels. We denote the true SIR and its estimate by γ_d and $\hat{\gamma}_d$, respectively. The SIR estimation error, namely ε , is defined as $\varepsilon \equiv \gamma_d - \hat{\gamma}_d$. We note that a time difference, corresponding to the CQI feedback delay, is existent between γ_d and $\hat{\gamma}_d$. The user location can be represented by the short term average CPICH SIR, namely $\bar{\gamma}_c$, which can be derived by averaging the recent CQI samples. Consequently, we shall consider the conditional pdf of the error, which we denote by $pr(\varepsilon | \bar{\gamma}_c)$, in the MCS selection.

If the AMC is to operate at a certain target FER, namely ξ_{tgt} , the AMC must select the highest MCS level that satisfies

$$\int_{-\infty}^{\infty} \xi_j(\hat{\gamma}_d + \epsilon) \cdot pr(\epsilon | \bar{\gamma}_c) d\epsilon \leq \xi_{tgt}, \quad (33)$$

where $\xi_j(\cdot)$ is the FER of the j -th MCS level. However, (33) is cumbersome to evaluate every frame on each MCS level. We can simplify the problem by approximating $\xi_j(\cdot)$ by a step function

$$\xi_j(\gamma) = \begin{cases} 1 & : \gamma < \gamma_{tgt,j} \\ 0 & : \gamma \geq \gamma_{tgt,j} \end{cases}, \quad (34)$$

where $\gamma_{tgt,i}$ denotes the target SIR that satisfies $\xi(\gamma_{tgt,i}) \equiv \xi_{tgt}$. Then (34) reduces to

$$\int_{-\infty}^{\gamma_{tgt,j} - \hat{\gamma}_d} pr(\epsilon | \bar{\gamma}_c) d\epsilon \leq \xi_{tgt}. \quad (35)$$

Conveniently, the conditional pdf of ϵ can be well approximated by a lognormal distribution (Figure 125). Hence, if we target, for example, $\xi_{tgt} = 0.1$, the task is simply to find the maximum MCS level that satisfies

$$\hat{\gamma}_d + \left\{ E[\epsilon | \bar{\gamma}_c] - 1.29 \sqrt{Var[\epsilon | \bar{\gamma}_c]} \right\} \geq \gamma_{tgt,j}, \quad (36)$$

where $E[\cdot]$ and $Var[\cdot]$ represent the mean and variance respectively.

The term in brackets indicates the amount of power offset required to compensate the SIR estimation error. Therefore, the required power offset, namely ψ , is given as a function of $\bar{\gamma}_c$, that is,

$$\psi(\bar{\gamma}_c) = -E[\epsilon | \bar{\gamma}_c] + 1.29 \sqrt{Var[\epsilon | \bar{\gamma}_c]}. \quad (37)$$

Figure 126 shows an example of the required power offset to achieve an FER of 0.1 for various vehicular speeds. We used Monte-Carlo simulations to obtain $pr(\epsilon | \bar{\gamma}_c)$, and the result was applied to (33) to derive the exact required power offset (denoted as “simulation” in Figure 126) on average. The required offset was also calculated by (37) and then fitted to a form $Ae^{-B\bar{\gamma}_c} + C$ using the MMSE method (denoted as “ $Ae^{-B\bar{\gamma}_c} + C$ ” in Figure 126).

Figure 126 shows that the required offset increases rapidly as the average CPICH SIR ($\bar{\gamma}_c$) decreases. Hence, a user near the cell boundary requires a larger power offset. Figure 126 also shows that the required offset increases as the velocity offset is about 4 dB larger at 50 km/h compared to 3 km/h, at a user location represented by $\bar{\gamma}_c = 6$ dB. Note that these compensation curves cannot be given precisely in a convenient closed form (although we have shown that an exponential fitting, having the form $Ae^{-B\bar{\gamma}_c} + C$ works out quite well). Moreover, the required offset generally depends on the multipath profile and mobile terminal velocity. Unfortunately, this implies that a universal formula that suits any scenario cannot be given in a simple form. Hence, an operator has to presume a certain environment for the cell being deployed, as often done so in cell planning [49]. In any case, Figure 126 suggests controlling the AMC bias depending on the reliability of the CQI report.

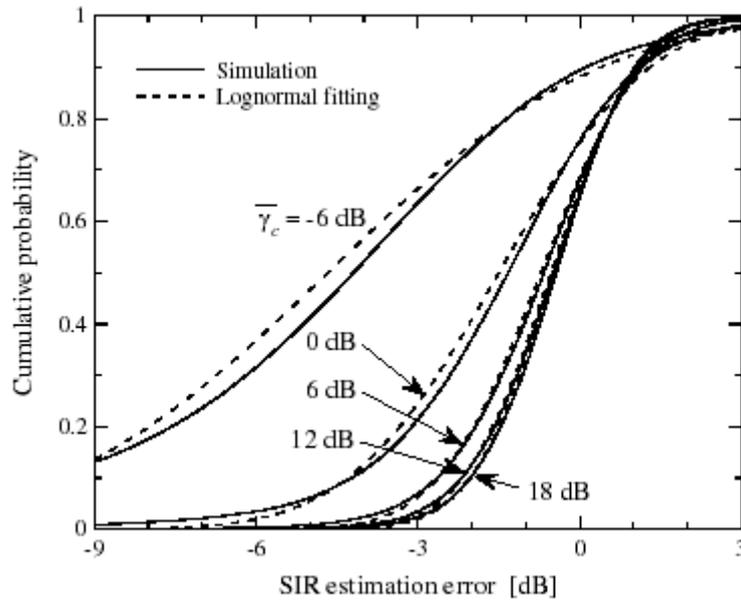


Figure 125 Lognormal fitting of SIR estimation error (3km/h).

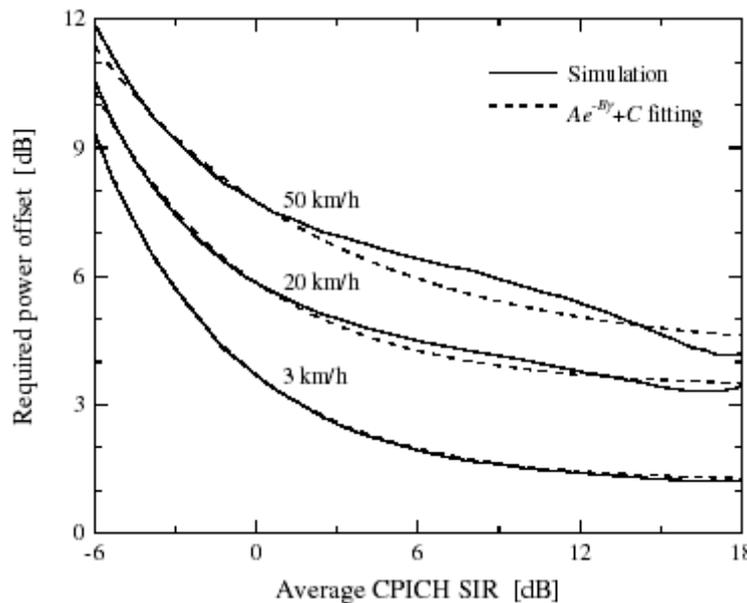


Figure 126 Power offset required to compensate SIR estimation error.

2.13.4.1.6 Performance of biased AMC

Assuming an 80 ms block size with due time $T = 40$ ms and a PF-LDF scheduler ($f = 1$), we examined the performance of the biased AMC. Figure 127 shows the outage probability for the biased AMC with offset being the fitted version, i.e., $\psi(\gamma) = Ae^{-B\gamma} + C$. To give a comparison, we have also shown the case without any bias (denoted as " $\psi(\gamma) = 0$ ") and the case with a fixed bias (denoted as " $\psi(\gamma) = C$ "). The outage probability is reduced significantly by the biased AMC at low traffic loads when the velocity is 80 km/h. The biased AMC also reduces the outage probability at low traffic loads at 3 km/h. At 80 km/h the traffic load that produces outages at a five percent rate (or system capacity) is increased from mere 12 Erlangs to 44 Erlangs by biasing the AMC. Although the exponential version ($\psi(\gamma) = Ae^{-B\gamma} + C$) performs slightly better than the fixed bias at very low traffic loads, the

outage probability rises steeply at high traffic loads. This is because the exponential version consumes a larger power and the system becomes power limited at high loads. Adapting the offset by location has a marginal significance when the system operates at controlled traffic loads (e.g., by blocking new calls to guarantee outage < 0.05). A fixed bias is sufficient to suppress outages. Our results suggest that compensating for the CQI feedback error is thus the essential part.

The proposed AMC achieves lower error rates by transmitting each MCS with a higher power. As shown in Figure 139, this imposes an extra power consumption on aggregate. However, an offset of 1 dB does not increase the average total power by 1 dB, but only by its fraction. The total power merely increases by 1 dB at 40 Erlangs at 80 km/h with a fixed offset. This is because the reduced number of HARQ retransmissions relieved the aggregate power increase. Therefore, the proposed scheme effectively improves the VoIP QoS at a cost of slightly increased power.

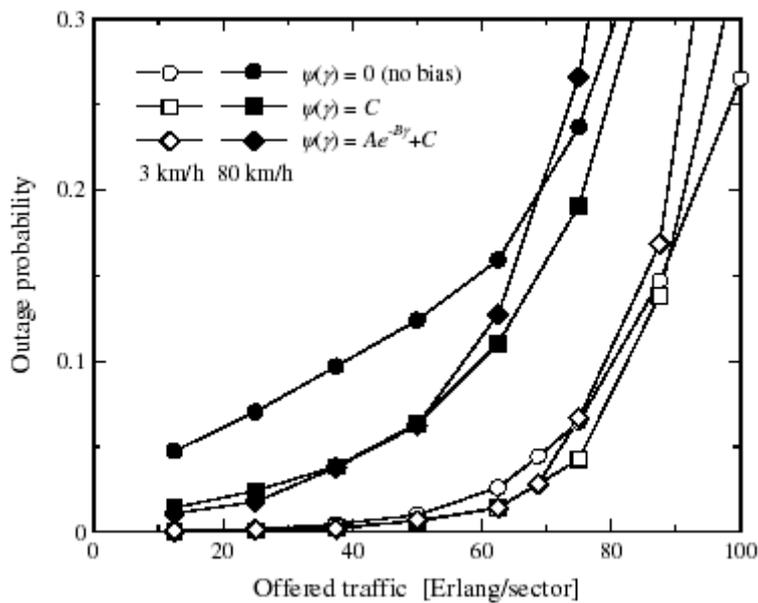


Figure 127 Biased AMC reduce the outage probability

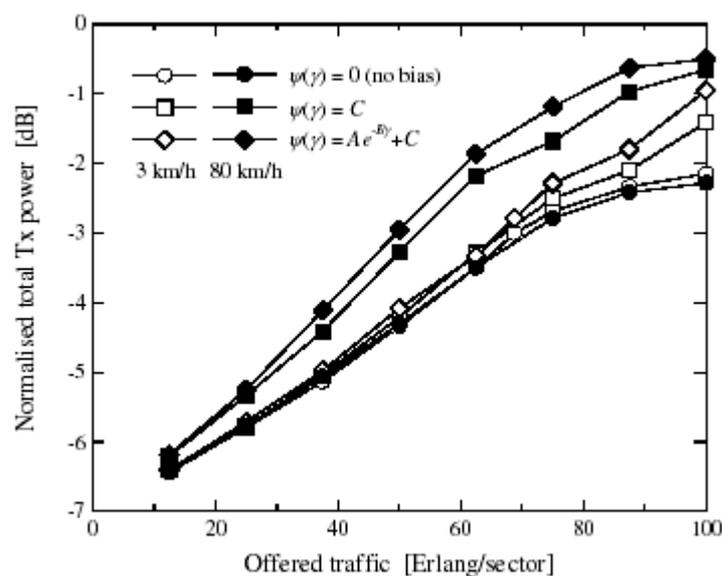


Figure 128 Biased AMC consume extra power

2.13.4.1.7 Conclusions

We investigated the possibilities of accommodating VoIP traffic in HSDPA. Assuming 12.2 kb/s AMR speech in a macro cell environment, VoIP deliveries over HSDPA was first studied with regard to the NodeB scheduler, AMR/RTP packet size, and mobile terminal velocity. Dynamic simulation results showed that a system capacity of 70 Erlangs can be achieved at velocities up to 20 km/h, by combining a longest delay first (LDF) discipline with a proportionally fair (PF) code multiplexing scheduler, and by using a block size of 20 to 80 ms. However, at 80 km/h the capacity is reduced to 12 Erlangs due to the inability to track fast fading with CQI feedback delay; the AMR and scheduler has to rely on outdated CQI reports.

Observing that the channel estimate is less accurate when the channel SIR is low, we further introduced a biased AMC scheme. A larger transmission power is imposed on each MCS to lower error rates, such that the number of HARQ retransmissions and the consequent delay are reduced. To derive an appropriate power offset, we analysed the statistics of the channel estimation error including the CQI feedback delay. Our simulation results showed that adapting the offset by the average SIR (representing location) produces a marginal significance, and a fixed bias is sufficient to improve the VoIP QoS. This implies that the CQI feedback delay is the main evil that must be mitigated. Our results showed that the biased AMC increases the system capacity by over a three-fold at 80 km/h. The current third generation systems carry speech traffic on dedicated channels (DCHs), that offer robust radio links with transmission power control (TPC) to resolve fast fading and soft handoff to provide continuity of service. However, the circuit switch orientation of the DCH limits flexibility to control QoS at IP and higher layers. Our results showed that HSDPA is able to convey VoIP traffic efficiently, especially at low mobility environments (20 km/h). The proposed biased AMC extends the feasible region to higher velocities. Since HSDPA is more flexible and allows affirmative interoperability with the Internet, VoIP over HSDPA has a potential future.

2.13.4.2 DiffServ-aware priority queuing improves IP QoS support on HSDPA

In this section, a priority queuing scheme for High-Speed Downlink Packet Access (HSDPA) that utilizes the QoS information from the IP layer of the differentiated services (DiffServ) architecture is presented. Here, we consider a most general case for DiffServ service structure based on RFC[50][51]. More details can be found in D08. In DiffServ the packets are delivered through traffic conditioning (e.g., classification, metering, policing, and shaping) at the edge of the scope domain, and simple differentiated forwarding mechanisms at routers [52]. To support various QoS, DiffServ offers different per-hop forwarding behaviours (PHBs), i.e., expedited forwarding (EF) [50] and assured forwarding (AF) [51] PHBs. The EF PHB is intended to support delay-sensitive premium services, such as VoIP, end-to-end across DiffServ domains. The EF PHB can be supported in practice by queuing the EF traffic separately from the AF traffic, and by giving the absolute priority over the AF in allocating resources when forwarding. The AF class, on the contrary, supports IP QoS for more elastic services by a simple mark and drop mechanism. The incoming packets are differentiated into four classes with three level drop preferences per class. By controlling the drop preference of the packets at congested times, the AF PHB serves better QoS than best effort.

According to the RFC proposal of the AF PHB [51], traffic aggregates are discriminated by the DiffServ code point (DSCP), and a random early detect (RED) queue is utilised to perform differentiated dropping of packets during congestion. In the simplest case the ingress DiffServ node performs only behaviour aggregate (BA) classification, where the classification of the packets is based only on the DSCP value. In general the classifier will be able to utilise more detailed header-based information from the incoming packets and perform multi-field (MF) classification. An example of MF classification is a 6-tuple classifier, where classification is based on six different fields from the IP/TCP (UDP) header, i.e.,

destination address, source address, IP protocol, source port, destination port, and DSCP. Each packet is marked with a dropping precedence according to the service profile; traffic that conforms to the corresponding service profile (in-profile packets) are handled with low drop precedence, whereas non-conforming traffic (out-of-profile packets) are handled with high drop precedence. The marking can be performed using algorithms such as token bucket and average rate estimation. For example the time sliding window three colour marker (TSWtcm) [53] estimates the arrival rate of packets by averaging the rate over a certain time window, and marks the packets into three colours (green, yellow, or red, in descending priority order).

Figure 129 shows the protocol stack of a session over HSDPA, taking HTTP for example. The end-to-end data transmission is controlled by the transport layer protocol, such as the TCP commonly used in the Internet. The user data are delivered through the IP backbone to the radio network controller (RNC). At the RNC IP packets are segmented into RLC packets. These packets are carried as a single MAC-d flow [54] by the lub/lur frame protocol (FP), which conducts flow control over the lub/lur. At NodeB the MAC-hs sublayer [54] schedules the MAC-d PDUs to serve on the HS-PDSCH and conducts AMC and HARQ. The RLC [55] has three modes of operation, i.e., the acknowledged mode (AM), unacknowledged mode (UM), and transparent mode (TM), depending on the functionality offered (e.g., ARQ and ciphering). Although the residual error rate after the HARQ is expected to be very low, the ARQ functionality of the RLC may be useful to preserve the data integrity, depending on the required QoS. If the UM or TM is used on the RLC, recovery of erroneous data must rely on higher layer protocols such as the TCP. The packet data convergence protocol (PDCP) [56] is used to compress the IP header to increase the radio spectral efficiency.

An important feature of the HSDPA that has not been emphasized in the literature is the priority queuing mechanism. The scheduling is handled in the MAC-hs sublayer of the HSDPA entity and the MAC-hs protocol header has 3 bits dedicated to indicate the queue ID [57]. Thus, each user may have up to 8 queues simultaneously. Each queue may be assigned a priority level depending on various conditions that may be exploited, such as the application, service contract, or the propagation condition. As shown in Figure 129, the MAC-d flow is transferred over the lub/lur by the frame protocol (FP). The FP has 4 bits to indicate the priority level of the FP payload. Hence, 16 different priority levels may be configured in the system. Each queue is mapped to a priority level, and the relation is signaled by the lub protocol called NodeB application part (NBAP) during the call setup [58]. An example of priority level assignments to the queues of three users is shown in Table 26. As is evident from, the priority queuing mechanism of HSDPA is highly flexible.

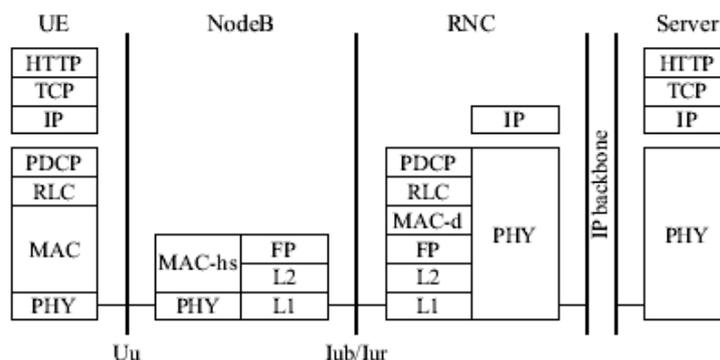


Figure 129 Protocol Stack

The allowance of 8 queues per user is especially useful when multiple services coincide for a user, e.g., downloading an email while making a speech call. The use of multiple queues is, however, not limited to such multiple service case. A single service can utilise multiple queues to differentiate the transfer of various contents. For example, the base layer

information such as texts in a web page, and the enhancement layer information such as background images, can be differentiated in a transmission utilising multiple queues.

Table 26 Example of priority level assignment to queue IDs.

Priority level	UE 1	UE 2	UE 3
15	Queue 3	-	-
14	-	Queue 1	-
13	-	Queue 7	-
12	Queue 1	Queue 4	Queue 0
11	Queue 5	Queue 6	-
10	-	Queue 3	-
9	Queue 6	-	-
8	Queue 7	-	Queue 3
7	Queue 2	-	Queue 5
6	-	-	Queue 6
5	-	Queue 2	Queue 7
4	-	Queue 0	-
3	Queue 4	-	Queue 1
2	Queue 0	-	-
1	-	Queue 5	Queue 2
0	-	-	Queue 4

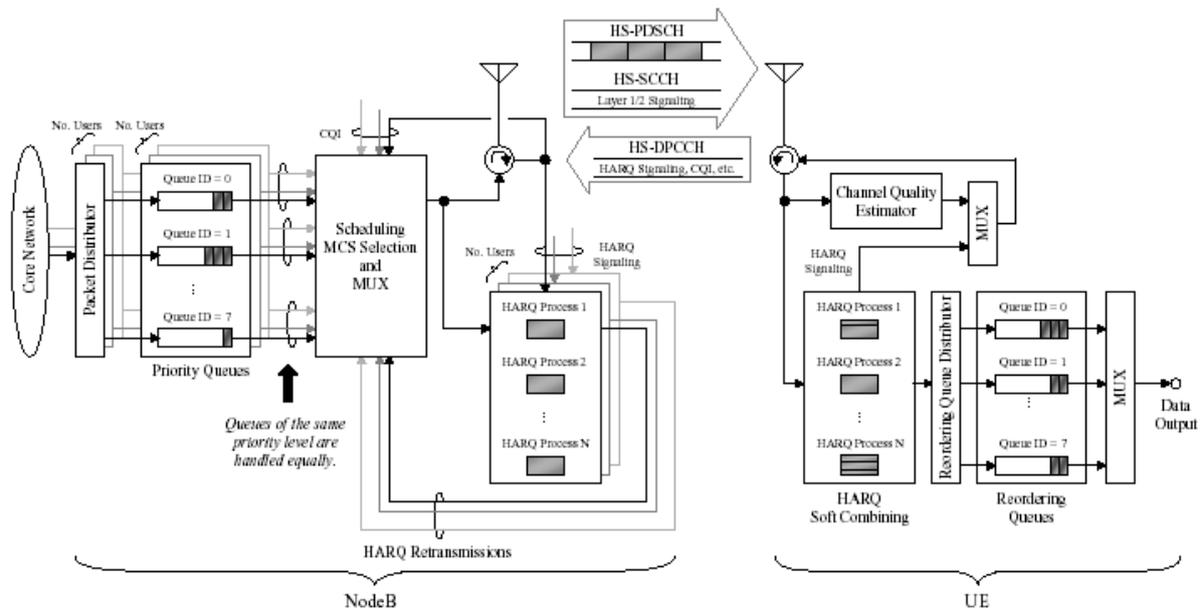


Figure 130 General Architecture of HSDPA

Figure 130 depicts the general architecture of HSDPA with priority queuing. The packets arriving in the NodeB from the core network are distributed into a maximum of 8 queues per user, according to the priority level. The queues of the same priority level are handled equally in the scheduler. The scheduler selects a set of users to transmit data on the next HS-PDSCH frame, and notifies to those users of the data presence by signalling on the HS-SCCH. A new data block transmitted is assigned an HARQ process, and stored in the NodeB memory until the data block is correctly received by the mobile terminal, or the maximum number of retransmissions is reached. If the data block is still erroneous after the maximum number of retransmissions, the data recovery is left to higher layer protocols such as the radio link control protocol (RLC) and TCP. The maximum number of parallel HARQ processes (denoted by N) depends on the mobile terminal capability and the NodeB configuration. Once an HARQ process becomes idle after a successful transmission, a new data block may be assigned to the HARQ process. Due to the N-SAW the sequential order of packets is not ensured at the receiver. Therefore, a reordering mechanism is needed per

flow in the receiver as shown in Figure 130. The HARQ acknowledgements and CQI are reported via the uplink dedicated physical control channel (HS-DPCCH).

The priority queuing mechanism is a fountain for novel ideas. In the simplest example, the priority of packets may be differentiated by application. A real-time service may be assigned a higher priority level than a web browsing service, and background services such as email may be assigned a lower priority level. The priority level may also be controlled by the contracted service profile, thus enabling a business model adaptive to the requirements of each user. In addition a retransmitted data by higher layer protocols such as RLC and TCP may be given a higher priority. This should improve the packet delay performance and the overall QoS. Another approach would be to use the propagation condition, exploiting the average CQI value for example. The scheduling algorithm and the priority handling therefore provide a platform for highly customized network behaviour, without requiring modifications on the mobile terminal.

The prospects of HSDPA and DiffServ support in UMTS yield two eminent issues in supporting end-to-end QoS involving UMTS, i.e., to map the DiffServ PHBs onto priority queues in HSDPA, and to schedule the different priority queues for radio transmission. The mapping is to assign a priority level in HSDPA to the 13 different QoS levels in DiffServ, i.e., one EF and four AF classes, each with three drop precedence levels (colours). The EF and AF classes, and so are the different colours of the same AF class, are distributed into different priority queues. The NBAP header allows the HSDPA to realise up to 16 different QoS levels. Hence, each of the 13 levels in DiffServ can be mapped onto a different priority level in HSDPA. Note that in addition to the DSCP classification, an MF classification can also be performed in the mapping process.

The optimum scheduler shall maximise the system capacity while satisfying the QoS constraints of each priority level. An operator must define the 16 levels of QoS, and guarantee each QoS level through appropriate scheduling when heterogeneous traffic coexist. A noteworthy remark is that the priority queues in HSDPA are given per flow, as opposed to the behaviour aggregate queues in DiffServ. Nevertheless, the scheduler algorithm is not specified by the 3GPP, and is up to implementation. Hence, the actual scheduler design can incorporate flow-based rules as well as aggregate-based rules, exploiting various conditions, including the CQI, HARQ status, required QoS, traffic constitution, and user mobility.

The user packets are scheduled for transmission on a set of shared code channels depending on the radio channel condition and the drop precedence (packet colour) of the IP packets marked by the ingress DiffServ node. The out-of-profile packets that are beyond the contracted service level agreement are marked with high drop precedence and thus are given less priority (served only on residual resources, i.e., code and power). By truncating the throughput for the out-of-profile packets, the in-profile traffic experiences a higher throughput, consequently improving the QoS support.

The color aware priority queue here is to control the priority of the queues with the packet color information in the AF concept of DiffServ architecture through Time Sliding Window three color marking (TSWtcm) algorithm. The green, yellow, and red packets are distributed to different queues with different priorities. The scheduler serves primary green packets, then yellows, and finally reds, on residual resources. A yellow (red) packet gets a chance to transmit only if enough codes and power are available after allocating resources to green (and yellow) packets. More technical details regarding the queuing schemes are presented in [17] Section 2.3.4.1. Our simulation results in Figure 131 show through the differentiation over the air-interface, the in-profile packet enjoy a much higher throughput than the out-profile packets (yellow and red). Also in Figure 132 shows the balance between the average in-profile user throughput and the system throughput. The system throughput increased as the

offered traffic load was increased with and without the proposed queueing. This is because at higher loads, the probability of using higher MCS levels increases since the scheduler is able to choose from many users. However, the in-profile user throughput decreased at high loads, because each user had less chance to transmit. Therefore, the traffic load greatly impacts on the system behaviour in both queueing schemes. The priority queueing shifted the system behaviour towards higher in-profile user throughput, thus improving QoS support. Although the system throughput was smaller in the priority queueing case, this is only because that out-of-profile traffic is truncated.

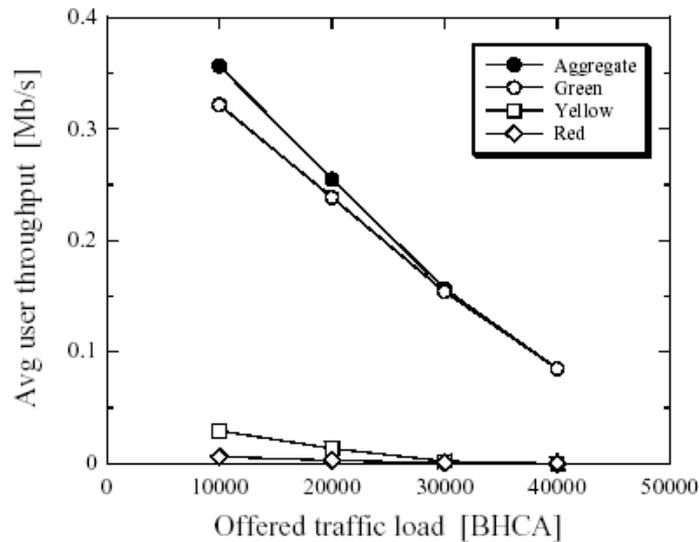


Figure 131 Average User Throughput with Priority Queuing

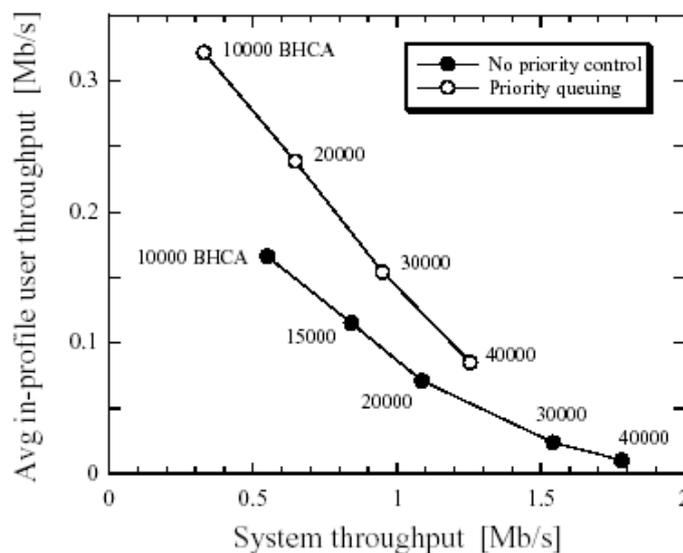


Figure 132 Priority Queuing Shifts System Behaviour towards a Higher In-profile User Throughput

2.13.4.3 Supporting Heterogeneous Traffic in HSDPA

In [1] and [17], we have reported DiffServ-aware radio resource management schemes to support end-to-end QoS. In [1], we have outlined the theory frame-work of the color aware scheduling strategies in order to minimize the downlink transmission power in DS/CDMA under QoS constraints of the DiffServ structure. And a link adaptation problem is developed based on the theory for DS-CDMA systems. A color-aware coverage control scheme was further introduced for DS/CDMA in [1] and simulation results showed that the coverage

control scheme gives a better throughput performance of the traffic than rate control strategy with conforming to the service level agreement. Then the color-aware control scheme was further extended to the priority queuing in HSDPA in [17], and system simulations on a single service scenario showed that the priority queuing effectively improves the throughput of the conforming traffic. Nevertheless, these papers have not considered heterogeneous services. In this section, we report a new DiffServ-aware link protocol for DS/CDMA to support multiple both EF and AF service classes over DS/CDMA air-interface in [17]. Here, we extend our DiffServ-aware link adaptation to support heterogeneous services over HSDPA systems.

In this section, Heterogeneous traffic support in the High-speed Downlink Packet Access (HSDPA) of the Universal Mobile Telecommunication System (UMTS) is studied. Two eminent issues are addressed for supporting mixed traffic: mapping of the different QoS classes onto the priority queues of HSDPA, and scheduling the different priority queues. As a case study, a viable scheduler is developed for a three-service scenario, i.e., coexisting voice over IP (VoIP), variable bit rate (VBR) video streaming, and web-browsing services. The scheduler gives the absolute priority to the VoIP, with biased adaptive modulation/coding to suppress the packet delay. The streaming and web traffic are served on the residual resources in a single queue, under the combined modified weighted delay first (M-LWDF) and proportionally fair (PF) rules, with a priority balancing parameter. Dynamic system simulations show that the developed scheduler autonomously balances the outage probabilities of the three services, and is robust in supporting various traffic constitutions.

Providing quality of service (QoS) support on the Internet is increasingly desired as the diverging variety of services, including realtime video streaming and voice over IP (VoIP), are provisioned over the rapidly expanding Internet. Yet another intension is to realise seamless mobility of such services, thus raising the prospect of IP-based transport on radio access networks (RANs) [59]. To realise seamless end-to-end QoS involving UMTS, the edge functions of the differentiated services (DiffServ) concept [60] have been supported in the 3GPP since Release 5, with the logical element called IP BS manager [61]. The IP BS manager is responsible for the QoS handling in the UMTS domain, and translates QoS classes in diplomacy with external domains as a bandwidth broker. As IP-networking in UMTS pervade, in [24] and D16, EVEREST project has proposed our end-to-end QoS architecture based on release 5. In this architecture, the radio link becomes the last hop along the QoS end-to-end path. Due to the wireless medium, which is characterised by scarce spectrum, unstable propagation, and stochastic user mobility, channel-aware management of the radio resource is essential in the wireless domain. Hence to support end-to-end QoS efficiently in UMTS, radio resource management, that integrates available information from the radio link and IP layers, is indispensable.

HSDPA [62][63] is capable of supporting various QoS levels. When heterogeneous traffic coexist, an HSDPA packet scheduler that discriminates QoS classes is speculated to provide better QoS support than those that consider only the channel condition. The HSDPA specification supports up to 16 priority levels and 8 queues per user [62][63]. With the prospect of DiffServ support in UMTS, two issues are eminent in supporting end-to-end QoS involving HSDPA, i.e., to map the QoS classes of DiffServ onto the priority queues in HSDPA, and to schedule the different priority queues for radio transmission. In the following, we discuss these issues in detail, and performs case studies considering three services, i.e., voice over IP (VoIP), variable bit rate (VBR) video streaming, and web-browsing. A viable solution is eventually developed through simulation studies, in which priority queuing and various QoS and radio channel criteria are incorporated, departing from the simple single queue PF scheduler.

To elaborate on the heterogeneous traffic support in HSDPA through a concrete, realistic scenario, a dynamic system level simulator was developed based on the one presented in [37]. The packet transmission process with AMC and N-SAW HARQ was simulated frame-

by-frame assuming a sectorised multicell environment and Poisson arrivals of packet calls. The signal-to-interference (and noise) power ratio (SIR) after the rake and HARQ combining was calculated per frame, and the SIR was used to look up a frame error rate (FER) table prepared for each MCS level by link level simulations. Frame errors were generated randomly using this FER. Channel estimation error was taken into account in the SIR and CQI calculations, through statistical modelling [64].

Table 27 System configuration

Cell layout	7 cells, 3 sectors, wrap around
Site separation	1 km
Carrier frequency	2 GHz
Chip rate	3.84 Mc/s
Path loss	COST231-Hata
Shadowing	spatially correlated lognormal (std. deviation = 8 dB)
Multipath fading	3-path Rayleigh (max Doppler = 5.56 Hz)
Handoff hysteresis	3 dB
Receiver noise figure	9 dB
Total Tx power	20 W
CPICH Tx power	2 W
HS-PDSCH Tx power	16 W (max)
HS-PDSCH spreading factor	16
Number of HS-PDSCH codes	max 15
HS-SCCH spreading factor	128
Number of HS-SCCH codes	4
AMCS	25 MCS levels (68.5 kb/s to 7.2 Mb/s)
HARQ	6-SAW, Chase combining
CQI feedback delay	2 HS-PDSCH frames

Table 27 summarises the main system parameters. A regular hexagonal cell layout was assumed with the wrap around technique [65] applied to avoid the boundary effect. Users were generated by a Poisson process, assuming uniform distribution over the service area. The radio propagation was simulated as a concatenation of the Hata loss [66], [67], spatially correlated lognormal shadowing [68], and multipath Rayleigh fading [43] with maximum Doppler frequency of 5.56 Hz. The AMC was performed using 25 MCS levels from 68.5 kb/s to 7.2 Mb/s (UE category 8 in the 3GPP specification [69]), with a CQI feedback delay of two HS-PDSCH frames. Moreover, the Chase combining HARQ [49] was simulated with 6-channel SAW per user. The HS-PDSCH was allocated the residual code and power resources after allocating necessary resources to the other channels.

Three services are considered: voice over IP (VoIP), variable bit rate (VBR) video streaming (MPEG, hereafter), and web browsing (HTTP, hereafter). Main parameters of the traffic models are summarised in Table 28. Details of the traffic models for VoIP, MPEG, and HTTP are found in [24], respectively. A call was labelled as an “outage” if the perceived QoS did not comply with the requirements stated in Table 28. To avoid unstable states (where the number of concurrent users diverges over time, due to the aggregate incoming rate exceeding the system bandwidth), a call was dropped and marked also as outage if one of the following has occurred.

1. The dropping criterion shown in Table 28 has occurred, for the respective service.
2. A frame remains erroneous after 20 HARQ retransmissions.
3. The CPICH SIR drops below -10 dB for consecutive 100 ms.

No call admission control was applied for the purpose to examine the achievable performance. A single Poisson process was splitted into the three services, by the nominal ratio 0.07, 0.15, and 0.78 for VoIP, MPEG, and HTTP, respectively. Although the arrival rate of VoIP is less than a tenth of the HTTP, the longer holding time of the VoIP causes more

VoIP users to remain in the system. The splitting ratio was calculated by taking the ratio of capacities for each service (that were derived internally), such that the radio resource is expected to be divided roughly equal among the three services on average.

Table 28 Traffic Model Parameters

VoIP	description	80 ms block size AMR speech
	DiffServ PHB	EF
	source rate	peak 12.2 kb/s, 40% activity
	holding time	exponential (mean 90 s)
	required QoS	40 ms delay, 1% packet loss
drop criterion	packet loss > 10% after 10 s	
MPEG	description	VBR MPEG-4 video streaming
	DiffServ PHB	AF-1
	source rate	variable (mean 53 kb/s, peak 940 kb/s)
	clip length	exponential (mean 15 s)
	required QoS	2 s delay, 5% packet loss
drop criterion	packet loss > 20% after 5 s	
HTTP	description	interactive/background service
	DiffServ PHB	AF-2
	source rate	mean 2 Mb/s
	data size	truncated Pareto (mean 25 kbyte, max 2 Mbyte)
	required QoS	32 kb/s throughput, 0% packet loss
drop criterion	throughput < 3.2 kb/s after 10 s	

Using the dynamic simulator, the outage performance of various schedulers are studied in the sequel, with the aim to develop a viable scheduler for supporting heterogeneous traffic. Note that the system capacity is defined as the maximum call arrival rate to guarantee the outage probability of five percent to all the services. Hence, the outage probability of the bottleneck service limits the capacity.

2.13.4.3.1 Single queue schedulers

The simplest scheduler would be to serve all the traffic in a single queue using per flow-oriented disciplines such as first-in first-out (FIFO) and round robin (RR) [49]. The MaxC/I, PF, and M-LWDF rules provide more sophisticated solution, utilising the channel condition (see [58], [51], for example). As shown in [58], the PF scheduler excels in supporting interactive and background services, whereas [51] and [46] showed that the M-LWDF rule is viable for streaming and conversational services, respectively. Hence, the following schedulers are considered to give a benchmark:

- ◆ Single queue PF: All the traffic is directed to and served from a single queue irrespective of the service class. The PF rule (see Annex B for more details) is applied to schedule the traffic with up to four user code multiplexing (waterfilling). Therefore, the scheduler considers only the CQI.
- ◆ Single queue M-LWDF (see Annex B for more details): All the traffic is directed to and served from a single queue irrespective of the service class. The M-LWDF metric is applied to schedule the VoIP and MPEG traffic, whereas the PF metric is applied to the HTTP. Hence, the priority metric for the j -th user is calculated as

$$p_j = \begin{cases} -\log \rho_j \cdot \frac{R_j(t)}{\bar{R}_j(t)} \cdot \frac{\tau_j(t)}{D_j}, & \text{VoIP and MPEG} \\ \frac{R_j(t)}{\bar{R}_j(t)}, & \text{HTTP} \end{cases}$$

where $R_j(t)$, $\bar{R}_j(t)$, ρ_j , $\tau_j(t)$, and D_j are the current supportable rate, average rate, allowable packet loss rate, head-of-line delay, and the packet due time, respectively for j -th user. Hence, QoS differentiation is established within the single queue. Waterfilling is

performed utilizing the available resources through code multiplexing, allowing up to four simultaneous users.

Figure 133 shows the outage probabilities for VoIP, MPEG, and HTTP served by single queue schedulers, both PF and M-LWDF. The outage probability for VoIP rises steeply, exceeding 5% at less than 1 call/s, whereas the outage probability for MPEG is maintained below 5% until about 3.5 calls/s. The HTTP outage is nearly zero even at 4.1 calls/s. The VoIP users suffered poorer QoS due to the scheduler not being able to reconcile the stringent delay. Although the M-LWDF scheduler marginally reduces VoIP outage compared to the PF, the outage probabilities are still widely dispersive for the three services. The VoIP outage limits the system capacity to less than 1 call/s. A well designed scheduler shall exhibit similar outage probabilities for all the services, to avoid any particular service (VoIP in this case) being a bottleneck. The single queue schemes with PF and M-LWDF are insufficient to support heterogeneous traffic efficiently.

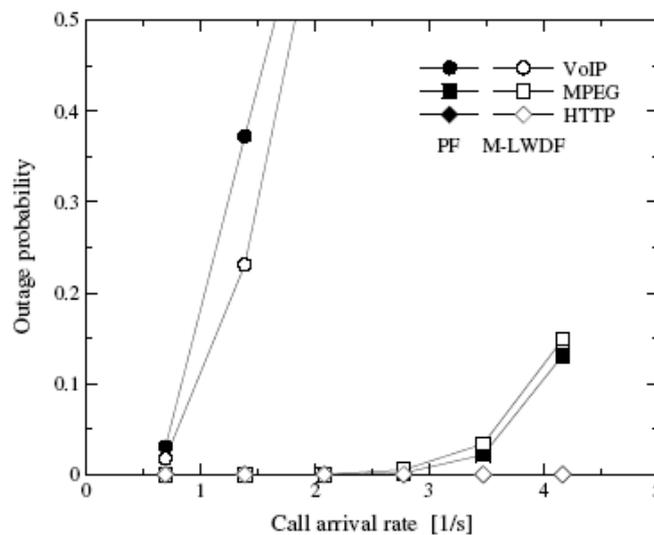


Figure 133 Outage performance for Single Queue

2.13.4.3.2 Priority queuing

The previous result (Figure 133) implies that a priority queuing mechanism is necessary to reduce the outage discrepancy, especially to improve the VoIP outage. Thus, two schedulers are considered:

- ◆ Priority queuing PF: The three services are queued separately and scheduled with strict priority. The VoIP class (EF) is mapped to the highest priority level, the MPEG class (AF-0) to the second, and the HTTP class (AF-1) to the lowest priority level. The PF rule is applied within each class.
- ◆ Priority queuing M-LWDF: The rule similar to the above is applied, however, with the M-LWDF rule applied to the VoIP and MPEG, and the PF rule to the HTTP.

Figure 138 shows the outage probabilities for VoIP, MPEG, and HTTP served by the priority queuing schedulers. The outage probability for VoIP is significantly reduced from the single queue schemes (Figure 133), because the VoIP traffic were able to enjoy the absolute priority over the MPEG and HTTP. The outage probability is maintained below 5% up to about 2.8 calls/s for all the services. Unlike the single queue case, the HTTP outage increases sharply beyond 2.8 calls/s, since the HTTP traffic had to queue behind all the VoIP and MPEG traffic. The MPEG traffic experiences lower outage probability than the HTTP, given the second priority level next to the VoIP. The M-LWDF rule marginally reduces the outage probabilities for the VoIP and MPEG traffic, while degrading the HTTP outage. A comparison with Figure 133 implies that the priority queuing is more effective in supporting

the heterogeneous traffic under study. The system capacity is increased to 2.8 calls/s, with the VoIP and HTTP being the limiting services.

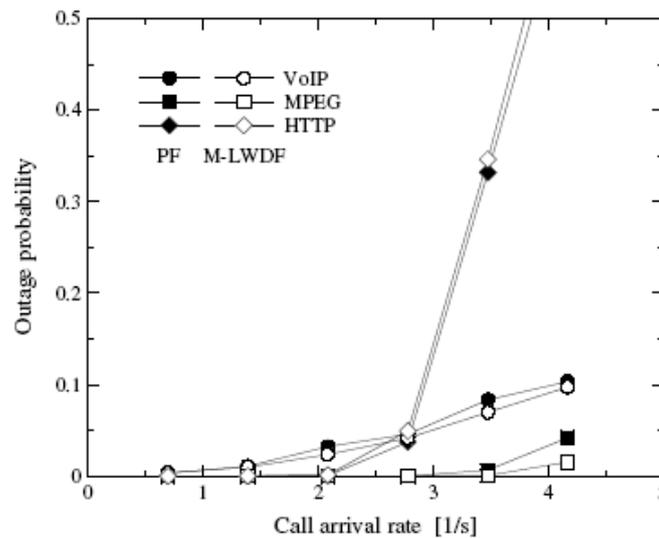


Figure 134 Outage performance with Priority Queues

2.13.4.3.3 Hybrid priority queuing with biased AMC for EF PHB

Although Figure 138 has shown that priority queuing is effective to support heterogeneous traffic, careful observations of the previous results suggest two possibilities to improve the overall outage performance, i.e., to refine the AMC for VoIP, and to balance the priority of the MPEG and HTTP.

Although the VoIP traffic were given the absolute priority over the MPEG and HTTP, the VoIP (along with HTTP) is the limiting service in terms of capacity. This suggests that further improvement is necessary for VoIP to increase the capacity. A solution is to refine the AMC for VoIP (or EF PHB, in general), by biasing the AMC as reported in the previous section (section 2 biased AMC). Each MCS is transmitted with a larger power to reduce HARQ retransmissions, provided the power is available at NodeB. If the power is unavailable, the MCS level is simply lowered by an equivalent amount. By imposing larger transmission power on each MCS level, the number of HARQ retransmissions would be reduced for VoIP packets. This shall reduce the consequent delay and prevent VoIP outages. By reducing the HARQ retransmissions for the EF class, the AF classes (in this case, the MPEG and HTTP) may get more opportunities to transmit.

Another possible improvement is to balance the outage probabilities of the MPEG and HTTP traffic. Figure 133 and Figure 134 have shown that the HTTP traffic experience higher outage probability than the MPEG with priority queuing (Figure 134), inverse of the result for single queue schedulers (Figure 133). This suggests that giving the absolute priority to the MPEG over the HTTP exaggerates the MPEG QoS, whereas straight comparison of the M-LWDF and PF metrics for the MPEG and HTTP respectively in a single queue, acts in favour of the HTTP. As such, a priority balancing parameter a is introduced to equalise the outage probabilities of the two services, sharing the same queue. The priority metric is hence calculated by

$$p_j = \begin{cases} -\log \rho_j \cdot \frac{R_j(t)}{\bar{R}_j(t)} \cdot \frac{\tau_j(t)}{D_j} \cdot a, & \text{MPEG} \\ \frac{R_j(t)}{\bar{R}_j(t)}, & \text{HTTP} \end{cases}$$

Figure 135 shows the outage results for the case when $a = 1$ and $a = 2$. For VoIP the AMC bias was 1.21 dB in [17] [Section 2.3.4.2]. At $a = 1$ the MPEG traffic experience higher outage probability than the HTTP, although this inverts when $a = 2$. This suggests that the optimum balance lies between $a = 1$ and 2. The outage probability for the VoIP decreased considerably, benefiting from the biased AMC. No service is distinctly good or poor; the outage probabilities for the three services exceed 5% at comparable arrival rates. The capacity (the maximum call arrival rate to guarantee 5% outage to all the services) is about 3.2 calls/s with $a = 1$ and 3.4 calls/s with $a = 2$. This is a considerable increase compared to 2.8 calls/s with strict priority queuing (Figure 134).

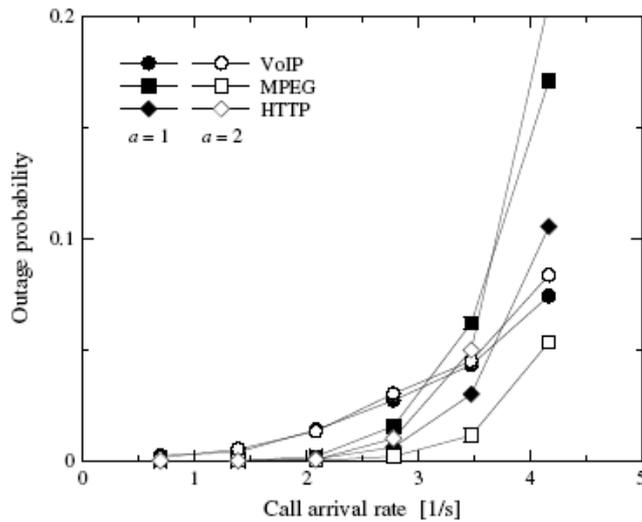


Figure 135 Outage Probability with Hybrid Queue

Figure 136 shows the maximum call arrival rate to guarantee 5% outage probability (capacity) for each service, across values of a between 1 and 2 (shown as the horizontal axis). The MPEG capacity increases whereas the HTTP capacity decreases, as the balancing parameter a is increased. This is as expected since a larger a gives higher priority to the MPEG over the HTTP. An interesting result is that the VoIP capacity is also affected by the parameter a . This can be explained since the parameter a affects the radio transmission of the MPEG and HTTP traffic, thereby affecting the interference experienced by the VoIP users. When a is around 1.26, the capacities of the three services become roughly the same at about 3.5 calls/s. The result shows that a value of a between 1.2 and 1.8 is adequate to support the heterogeneous traffic mix, enabling the system capacity (the lower envelope of the three curves in Figure 136) to reach about 3.5 calls/s.

Table 29 Traffic Mix

	Call arrival rate			Expected resource usage		
	VoIP	MPEG	HTTP	VoIP	MPEG	HTTP
Case 1	0.231	0.124	0.645	2/3	1/6	1/6
Case 2	0.048	0.413	0.539	1/6	2/3	1/6
Case 3	0.021	0.045	0.934	1/6	1/6	2/3

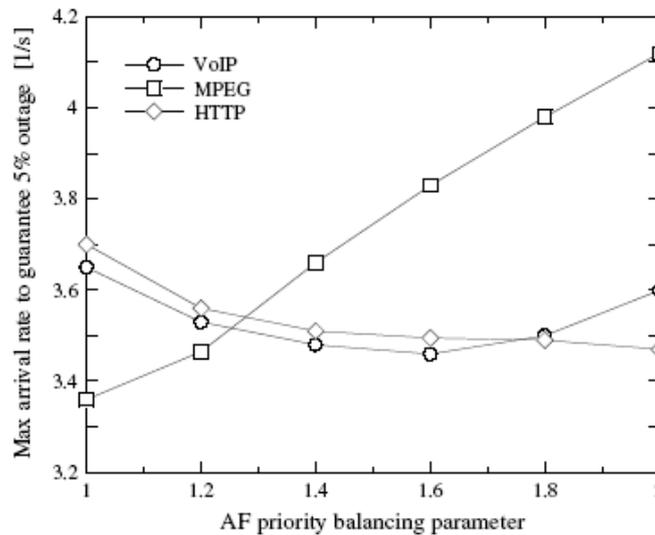


Figure 136 Impact of AF priority balancing factor on capacity.

To investigate the robustness of the developed scheduler to different traffic mix, three variations of the traffic mix are studied as shown in Table 29. As the expected resource usage (calculated from the single service capacities derived internally) indicates, the dominant service is the VoIP, MPEG, and HTTP for the cases 1, 2, and 3, respectively. The priority balancing parameter α was set to 1.26. Figure 137 shows the simulation results. Despite the variations in the traffic mix, the proposed scheduler exhibits comparable outage probabilities for the three services in each case; the outage probabilities cross 5% at roughly the same arrival rate. Therefore, the proposed scheduler is capable of supporting heterogeneous traffic, with elasticity to changes in the traffic mix. This is significantly valuable in practice, since the traffic mix changes dynamically over time. A solution that requires adjustment of parameters for different traffic mix cannot offer this scalability and is undesirable.

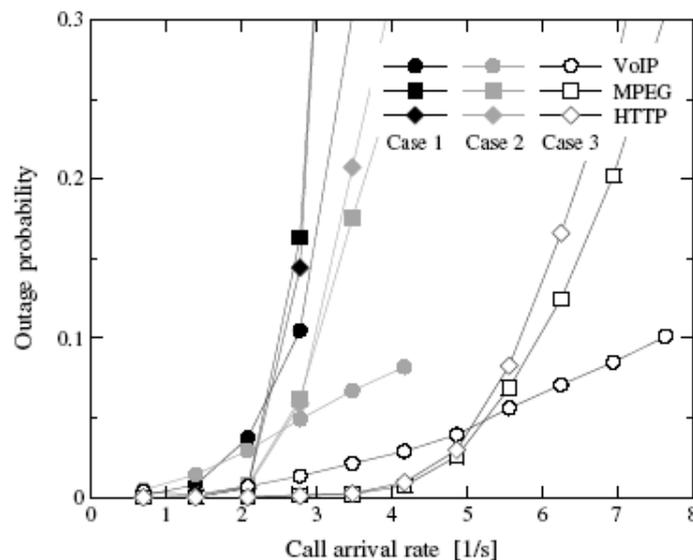


Figure 137 Hybrid priority queuing robustness to traffic mix variations.

2.13.4.3.4 Conclusions

Key issues in supporting heterogeneous traffic in HSDPA were discussed, and were elaborated through system simulations. The HSDPA specification supports up to 16 priority

levels and up to 8 simultaneous queues per user, each queue with one of the 16 priority levels. With the prospect for DiffServ support in UMTS, supporting heterogeneous traffic segregates into two key issues, i.e., to map the DiffServ traffic classes onto the priority queues in HSDPA, and to schedule the different priority queues. An operator must configure the appropriate set of QoS levels in HSDPA to support the provisioned services, and design a scheduler that supports the QoS levels efficiently under heterogeneous traffic environment. The scheduler design is up to implementation, and as such, various disciplines can be incorporated, including flow-based rules and aggregate-based rules.

A simulation study was performed to develop a viable scheduler, in which three services were considered, i.e., voice over IP (VoIP), variable bit rate video streaming (MPEG), and web-browsing (HTTP). The arrival rate of the services were nominally set such that the expected resource share is roughly even among the services (default scenario). A hybrid priority queuing scheduler was developed, into which the following rules were incorporated:

- ◆ The VoIP traffic are given the absolute priority over the MPEG and HTTP.
- ◆ The AMC of the VoIP traffic are biased (larger power is imposed on each MCS) to reduce HARQ retransmissions and the consequent delay.
- ◆ The MPEG and HTTP traffic are served by a single queue, using the M-LWDF and PF metrics for the MPEG and HTTP, respectively.
- ◆ A priority balancing parameter was introduced to equalize the outage probabilities of the MPEG and HTTP services.

The proposed scheduler was shown to support the heterogeneous traffic efficiently, without causing any particular service being a bottleneck in providing capacity. The simulation results further showed that the proposed scheduler is robust to changes in the traffic constitution, and supports call arrival rates of up to 3.5 calls/s in guaranteeing 5% outage to all the services in the default traffic scenario. This is substantially larger than some conventional schedulers, as strict priority queuing schemes support only up to 2.8 calls/s, whereas single queue schemes fail to secure 1 call/s, according to the simulation results.

The optimum scheduler design depends on the traffic QoS, and as such, the priority order and the balancing parameter, for example, must be adjusted accordingly for the provisioned QoS set. However, as the simulations revealed, by careful design, autonomous schedulers can be made robust to traffic variations. The key to success lies in the scalability of the scheduler, as was enabled by the priority balancing parameter in the proposed scheduler.

2.13.4.4 Code Multiplexing of Multiple Access Users in HSDPA

In this section, we provide a solution to the resource management problem with the code multiplexing in HSDPA. In conventional HSDPA, user packets are scheduled for transmission on a set of shared code channels in a time division manner. So the conventional scheduler serves only one user per frame dedicating all available code and power resources. According to our simulations, the resource usage on average reaches a mere 50% of the total transmission power and 40% of the available HS-PDSCH codes, when the scheduler is a single user PF. This is caused by the fact that even if PF considers the CQI, the best user does not necessarily qualify for such a high MCS level that fully utilizes the available power and code. Moreover, even if the CQI is remarkably high, a large fraction of the offered block size is wasted unless the transmitter buffer is loaded with awaiting packets. Although less transmission power means less interference, an appropriate cell planning [72] should consider the worst interference scenario where all base stations transmit at their maximum powers. Therefore, an attempt to fully utilize the available resources sounds more promising. Consequently, by letting other users transmit using the remaining resources – that is, by code multiplexing – we shall provide larger user/system throughputs as shown in Figure 47.

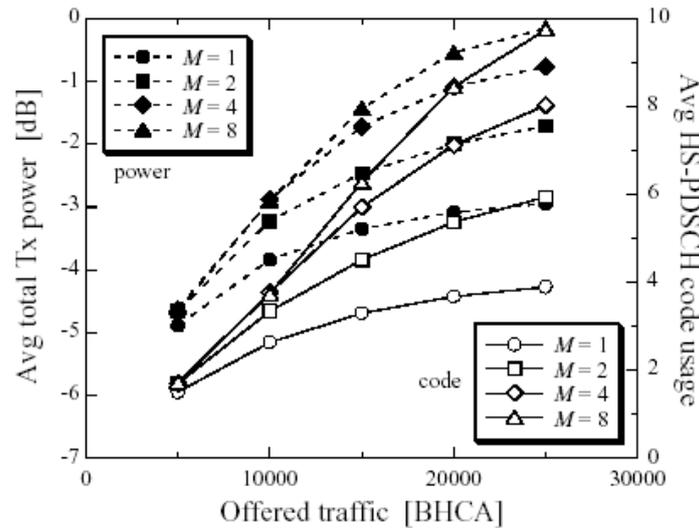


Figure 138 Radio resource usage comparison (for PF without IC).

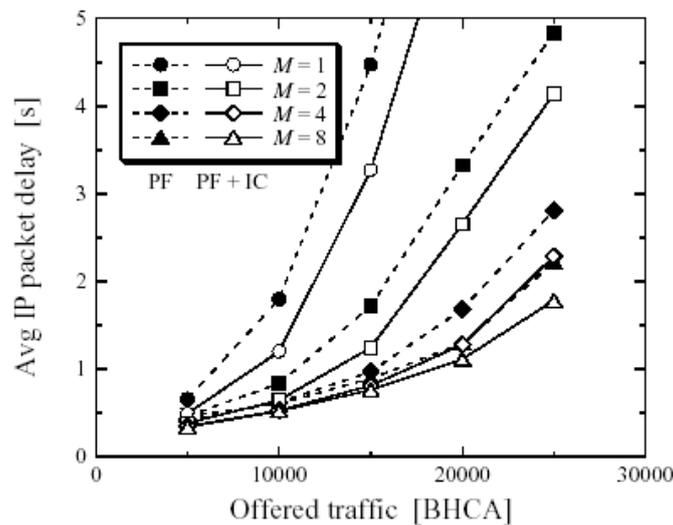


Figure 139 Average IP packet delay comparison.

The code multiplexing raises a new resource sharing problem in the scheduling process. The scheduler shall now select an optimum set of users as to maximize the throughput, however, with regard to fairness and delay constraints. Thus, one may fall into a depth of multi-variate optimization problem, which shall consider the CQI, HARQ status, fairness, packet delay, throughput, priority, and perhaps the amount of data in the transmitter buffer. If the aim is to simply maximize the system throughput, a simple solution is to perform water-filling following the MaxC/I discipline until the resources are packed. Derivatively, this shall relax the unfairness to a certain extent. Detailed performance in terms of efficiency and fairness and discussions are provided in [17] section 2.3.4.3. To provide a better degree of fairness, we consider the PF discipline and perform water-filling. We denote the maximum number of users that can be allocated the same frame for a code multiplexed transmission by M . Typically M is limited by the number of HS-SCCH codes that can be allocated. With the proposed scheduler, the users are sorted in a descending order of their relative CQI values (defined as the CQI deviation from the average CQI), as in an ordinary PF scheduler. The MCS for the first user is provisionally set according to the CQI value and the available amount of power and code resources. This MCS level is reduced if a lower level is enough to empty the transmitter buffer of that user. The first user takes the required resources for the

chosen MCS. If the resources are still available and M is not reached, this allocation process is repeated for the next user in the list.

An intriguing issue when we consider code multiplexing, is the use of interference cancellers (IC). Since UMTS employs orthogonal variable spreading factor (OVSF) codes [73], the intra-cell signals that arrive to a mobile terminal on the same path do not interfere with each other (unless the signals are spread using different scrambling codes). However, in a multipath environment, the signals via different paths interfere. The multipath interference canceller (MPIC) is a powerful tool to improve the SIR, and hence the BER/BLER performance, in a multipath environment [74]. Under multicode transmissions, a serial or parallel MPIC is able to cancel the inter-code interference among the multi-codes used by the same user, with some added complexity. Detailed modeling for the SIR calculation with and without MPIC is provided by [17] section 2.3.4.3. However, interference from other users' signals cannot be cancelled unless the mobile terminal is able to know the signal constellations. This suggests that the amount of interference that can be cancelled is maximized if a single user occupies all available codes in HSDPA. Consequently, the code multiplexing strategy seems not as effective if the mobile terminals support MPIC.

Table 30 summarizes the capacity, defined as the maximum offered traffic load that provides a larger rate than 32 kb/s to 95 % of users, for various scheduler disciplines and effects of interference cancellations. Note that the capacity values are normalized pivoting on the PF $M = 1$ case without IC. By allowing $M = 8$ on PF, the capacity is more than doubled. Interference cancellation further enlarges capacity. With $M = 8$ PF and IC, the capacity is significantly increased to 2.444. Table 30 also shows the relative capacity for RR and MaxC/I cases, for reference. Although RR and MaxC/I provide inferior capacities with the 32 kb/s QoS constraint, the code multiplexing effectively enlarges capacity with other scheduler disciplines such as MaxC/I. For more details on the other performance such as code usage, power sharing and IP packet delay are provided in [17] section 2.3.4.3.

Table 30 Capacity comparison of various schedulers.

M	RR	MaxC/I	PF	PF+IC
1	0.536	0.644	1.000	1.236
2	–	1.184	1.624	1.832
4	–	1.768	2.088	2.300
8	–	2.048	2.240	2.444

2.14 RAN SHARING

Sharing spectrum can be very attractive. For example, in rural areas UMTS coverage can be offered with much lower investment costs, but also in urban areas and hot spot areas capacity gains can be achieved. A UMTS FDD capacity gain of 28-49% speech and video Erlangs is claimed in [25], when two operators have one dedicated carrier each and two shared carriers. This capacity gain is due to the increased trunking efficiency as channels are pooled together between the operators.

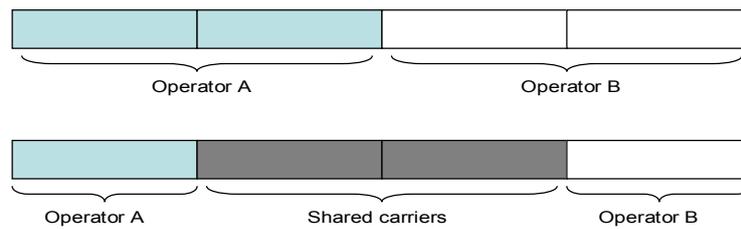


Figure 140. The capacity gain of 28-49% claimed in [25] is when operators share two carriers and use one dedicated carrier each, compared to when the operators have two dedicated carriers each.

Fairness of resource allocation can become an issue when operators share spectrum. For example, a service request from a customer of operator B might be rejected because operator A uses all radio resources.

2.14.1 Simulation study

In [17] we studied admission control algorithms affect on capacity and QoS fairness between operators in a shared UMTS network. We considered two operators that share an UMTS spectrum frequency in a hot spot area. The service mix was 25% speech (CS) and 75% HTTP (PS) traffic of the offered load. Three admission control methods, which allow some operator resource usage control, were tested. Two of the methods (algorithm 1a and 1b) use the scheme to divide the power usage between the operators. Algorithm 1a performs blocking of both CS and PS service requests at half power, whereas algorithm 1b only applies blocking of PS service requests. The third method (algorithm 2) was a new method proposed. It uses the bit rate elasticity of TCP flows in an attempt to achieve a fair QoS between the operators. A reference admission control method was used as well. It does not address any network sharing aspects, but only performs ordinary admission control for the purpose to achieve a good QoS. See [17] for more details on the algorithms and simulation setup.

It was shown that due to high CS blocking, algorithm 1a is not an attractive method. It gives a poor capacity. The QoS fairness for PS service is also poor.

For algorithm 1b, the PS blocking is high. Thus, also this method gives a poor capacity. It gives about the same bit rate as the reference method. The QoS fairness for PS service is poor even for this method.

Algorithm 2 achieves the best fairness, but not any impressively higher fairness than the reference method. It also gives the best capacity, and the highest bit rate for PS services.

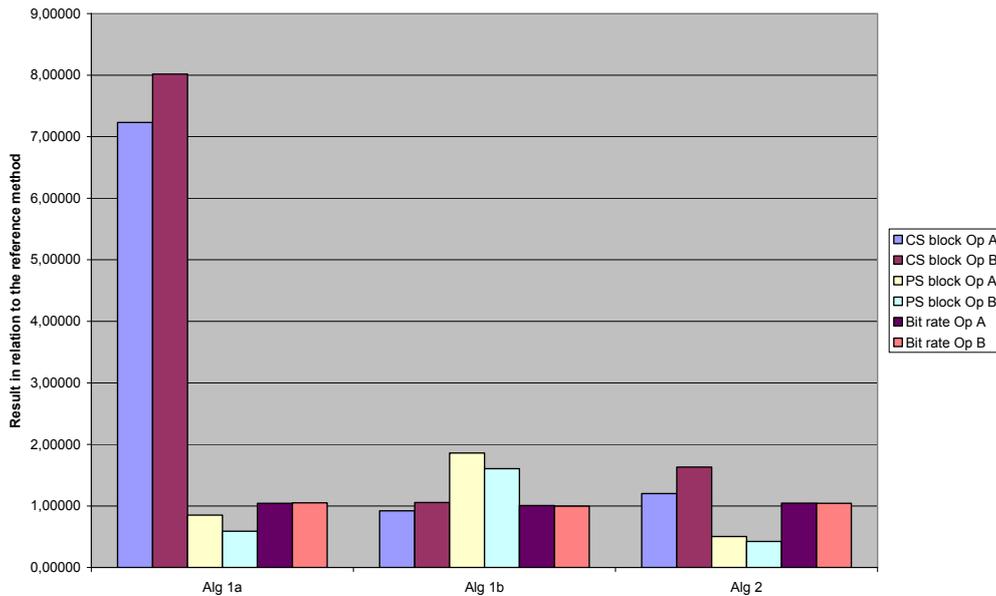


Figure 141. CS blocking, PS blocking and bit rate when two operators have equal offered load.

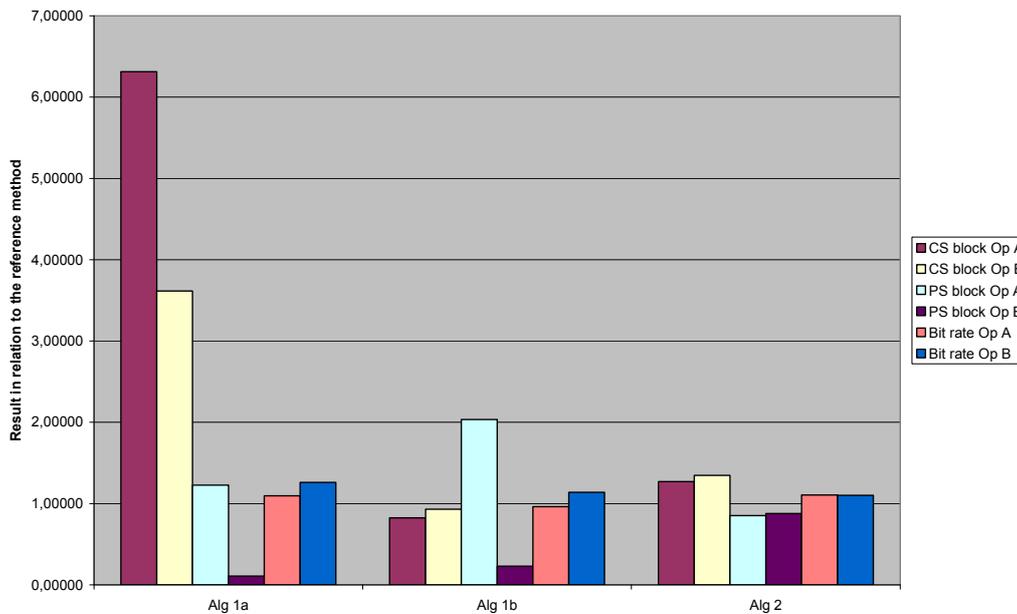


Figure 142. CS blocking, PS blocking and bit rate when two operators have unequal offered load.

2.14.2 Multicell RAN Sharing

2.14.2.1 Introduction

The increase in the demand of mobile multimedia services has caused the necessity of an efficient use of the available resources because of the scarce amount of usable spectrum. The current method of assigning spectrum to the different radio systems used by regulators is a fixed spectrum allocation (FSA) scheme. A block of radio spectrum is allocated to a particular Radio Access Technology (RAT) and this is often divided in allocations for the different operators of this RAT. This methodology provides easy regulation and better management of the spectrum but results in an inefficient use of the scarce resources.

Sharing resources between operators provides higher efficiency with respect to fixed spectrum allocation (FSA), especially when the traffic demand varies throughout the day. Fixed Spectrum Allocation cannot adapt to time varying capacity demands leading to a loss in the efficiency with respect to dynamic resource sharing due to the statistical multiplexing gain. [26][27] investigate the possibility for competing operator networks to cooperate without exchanging operational information and share a block of UMTS carriers simultaneously. Sharing the resources among operators may provide a reduction of investment costs especially in coverage limited areas (e.g. rural areas).

In WCDMA systems, the existence of RRM (Radio Resource Management [28]) strategies, which determine how the available radio resources of the system must be used and shared among the different users, may provide flexibility in the resource sharing among the different operators. In particular, [29] proposes an admission control where users that makes a connection request are queued and prioritised depending on the operators resources usage. Moreover, in [17] some sharing spectrum algorithms have been analysed. The proposed algorithm in [17] reduces the bit rate of non-Real time users in order to make room of new users demanding a connection requests. However, the algorithm is applied in a single cell, without considering the neighbouring cells. In a real network, certain areas may be more likely to have users of a certain operator. These non-homogeneities in the operator user distributions may cause lack of resources for a certain operator in a Node-B and resource availability in others. Therefore, taking into account the resource availability not only in a certain node-B but in the neighbouring cells, the efficiency in the use of the radio resources may be increased. In next section, different resource sharing algorithms will be compared for real time services.

2.14.2.2 Proposed sharing algorithms

When considering only real time services at constant bit rate CBR, the flexibility in the management of the radio resources is reduced. As an example, an overload situation cannot be solved by reducing the bit rate of certain users, because users are CBR. The proposed algorithms study different ways to increase the system efficiency by blocking connection requests or, if necessary by dropping certain number of established connections. In the following, the proposed algorithms are described.

Half power

When an operator A user makes an admission request, it is accepted if the total power devoted to this operator is lower than half of the maximum power, as shown in equation (38). Similar condition is checked for an operator B user admission request.

$$P_T^A + \Delta P_T \leq \frac{P_{T \max}^*}{2} \quad (38)$$

$$P_T^B + \Delta P_T \leq \frac{P_{T \max}^*}{2} \quad (39)$$

Where P_T^A and P_T^B are the base station transmitted power devoted to operator A and B, respectively, ΔP_T is the power increase estimation if the admission request is finally accepted and $P_{T \max}^*$ is the admission threshold.

Total power with dropping.

This algorithm is described in Figure 143. When an operator A user makes an admission request, a base station power availability check is carried out in this Node-B as shown in (40).

$$P_T + \Delta P_T \leq P_{T \max}^* \tag{40}$$

$$P_T = P_T^A + P_T^B$$

Where P_T is the base station transmitted power. If equation (40) holds, the admission request is accepted. If equation (40) does not hold (i.e. there are not available resources for accepting the connection request), then equation (41) is checked. By doing this, the base station power of operator B is compared to half of the total power of the base station. If equation (41) holds (i.e. operator B is using more than half of the total power), the connection request is admitted and certain number of Operator B users must be dropped in order to make room for the new Operator A user. If equation (41) does not hold, the connection request is rejected.

$$P_T^B \geq P_{T \max}^* / 2 \tag{41}$$

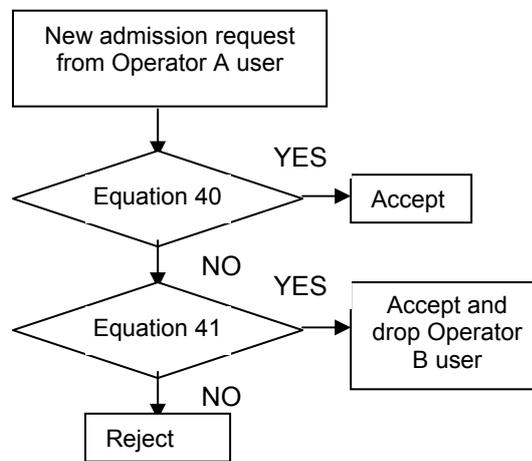


Figure 143 Total power with dropping algorithm

Multi-cell algorithm

This algorithm is described in Figure 144. The multi-cell algorithm not only considers the power availability of a single base station but the power availability of neighbouring cells. In this case, when there is not enough available power to accept the connection request (equation 40 does not hold), the total power devoted to operator B is compared with half of the total power as shown in equation (42).

$$\sum_{i=1}^{NumBS} P_T^B(i) \geq NumBS \cdot P_{T \max}^* / 2 \tag{42}$$

Where numBS is the total number of base station where the algorithm is executed. If equation (42) does not hold (i.e. Operator A is using more than half of the power of the system), then the admission request is rejected. Otherwise, it is admitted and certain number of Operator B users must be dropped in order to make room for the new Operator A user.

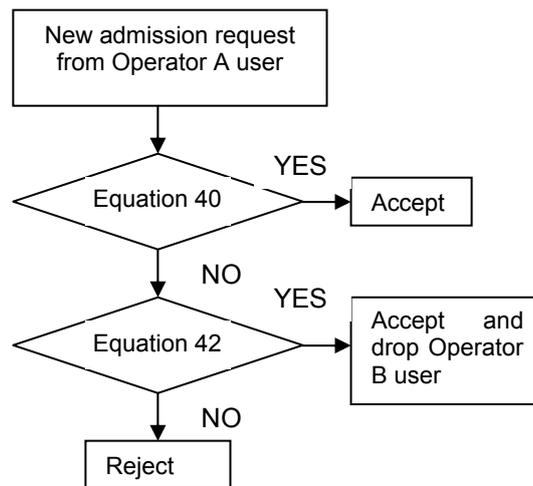


Figure 144 Multi-cell algorithm

It is worth noting that Figure 143 and Figure 144 present the algorithm behaviour when an Operator A user makes a connection request. Similar algorithm is considered when the admission request is made by an Operator B user.

2.14.2.3 Scenario Model

The considered cell layout consists of 12 omni-directional cells with base spacing of 1000m. The different base stations have been numbered as shown in Figure 145. A total number of 160 users or 180 users have been considered. In both cases, 50% of them have been distributed uniformly in all the scenario. These users may be from operator A (with probability p_A) or operator B (with probability p_B). The rest of users have been located in the hotspot 1 and hotspot 2. In the simulations, different percentage of users from Operator A and B in hotspot 1 and 2 have been considered. These users move at 3km/h with random movement taking into account the mobility model of [30].

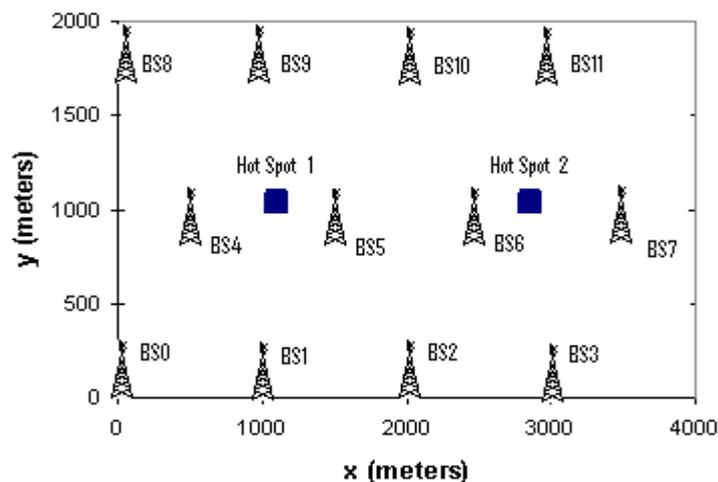


Figure 145 Scenario model

The simulation considers CBR 64kbps conversational services. The characteristics of the radio access bearer are given by a Transmission Time Interval (TTI) of 20 ms, a Transport Block Size (TB) size of 640 bits [31]. The characterization of the physical layer has been made by means of a link level simulator, which feeds the system level simulator with the transport Block Error Rate (BLER) statistics for each average (E_b/N_0). This characterization

includes a detailed evaluation of all the processes involved in the physical layer, like the estimation of the channel, antenna diversity, rate 1/3 turbo coding as well as the 1500 Hz closed loop power control. Similarly, these results at link level are used later to execute the outer loop power control (i.e. to compute the required E_b/N_0 , given a BLER requirement)[32]. Table 31 shows the simulation parameters.

Table 31 Simulation parameters

Parameter	Value
Chip rate W	3.84 Mcps
Frame duration	10 ms
BS parameters	
Cell type	Omnidirectional
Maximum DL power P_{\max}	43 dBm
Pilot and common control channels power P_c	30 dBm
Thermal noise	-106 dBm
Shadowing deviation	6 dB
Shadowing decorrelation length	20 m
Orthogonality factor	0.4
Measurement period (T)	1 s
Handover parameters	
Active Set maximum size	1
Time to trigger HO	0.5 s
Traffic model	
Call duration	120 s
Offered bit rate	64kb/s (CBR)
Activity factor	1
Call rate	29 calls/h/user
QoS parameters	
Block Error Rate (BLER) target	1 %

2.14.2.4 Results

In the following, the proposed algorithms will be compared. The case where no admission control is carried out is included as reference. Table 32 shows the Base Station 5 and Base Station 6 admission and dropping percentages when 50% of the users are from Operator A and 50% from operator B for the different algorithms. Moreover, the Grade of Service is presented. The Grade of Service is defined as $GoS = P_b + 10P_d$ where P_b is the blocking probability and P_d is the dropping probability. As shown in Table 32, the *Multi-cell* algorithm provides the best Grade of Service, because it provides higher admission probability than the *Half power* algorithm and maintains the dropping probability in a low value. As shown in brackets the gain in Grade of Service of the *Multi-cell* algorithm with respect to the *Half power* is $G=1.75\%$. Moreover, it can be observed that the *Total power with dropping* algorithm provides a poor Grade of Service because the droppings forced by this algorithm increases the dropping probability.

Table 32. 50% of users from Operator A and 50% from Operator B

	No admission control	Total power with dropping	Half power	Multi-cell
Admission (%)	100	91.20	83.89	89.35
Dropping (%)	6.63	1.53	0.16	0.67
GoS	66.39	24.08	17.7	17.39 [G=1.75%]

In the following, the percentage of users from Operator A and B in both hotspots have been changed. In hotspot 1, 25% of users will be considered to be from Operator A and 75% from Operator B (i.e. $P_{HS1,A} = 25\%$ and $P_{HS1,B} = 75\%$). On the contrary, in hotspot 2, 75% of users will be considered to be from Operator A and 25% from Operator B (i.e. $P_{HS2,A} = 75\%$ and $P_{HS2,B} = 25\%$). In Table 33 and Table 34 the admission and dropping percentages for Base Station 5 and Base Station 6 are presented. Moreover, these figures are shown for each Operator (OP.A and OP.B). As shown in Table 33, the *total power with dropping* algorithm reduces the admission probability of the Operator which has more users (Operator B in Base Station 5 and Operator A in Base Station 6). Moreover it increases the dropping probability of these users, resulting in a poor Grade of Service. With the *half power* algorithm the dropping probability is zero, but the admission probability is too low, especially for the users of the majority Operator (Operator B in Base Station 5 and Operator A in Base Station 6). The *Multi-cell* algorithm provides higher admission probability than the *half power* algorithm. Moreover, it provides lower dropping probability than the *Total Power with dropping* algorithm. Moreover, when setting the *Multi-cell* algorithm the admission and dropping percentage are more similar for the different Base Stations and Operators providing higher fairness between users.

Table 33 Admission percentage for $P_{HS1,A} = P_{HS2,B} = 25\%$ and $P_{HS2,A} = P_{HS1,B} = 75\%$

	NO ADM. CONTROL	TOTAL WITH DROP	HALF POWER	MULTICELL
BS5 OPA	100	99.01	99.44	89.40
BS5 OPB	100	89.37	74.56	89.08
BS6 OPA	100	90.72	78.33	90.85
BS6 OPB	100	98.51	98.54	90.09

Table 34 Dropping percentage for $P_{HS1,A} = P_{HS2,B} = 25\%$ and $P_{HS2,A} = P_{HS1,B} = 75\%$

	NO ADM. CONTROL	TOTAL WITH DROP	HALF POWER	MULTICELL
BS5 OPA	5.57	0.78	0	0.69
BS5 OPB	6.33	3.90	0.27	0.79
BS6 OPA	6.94	5.19	0.04	0.36
BS6 OPB	6.04	0.68	0.06	0.57

In Table 35, the obtained GoS is presented and compared with the case where 50% of users from Operator A and 50% from Operator B. The obtained gain with the *Multi-cell* algorithm with respect to the *Half power* algorithm is shown in brackets for both cases. As shown, the *Multi-cell* algorithm provides higher gain when the percentages of users per operator in each hotspot are different.

Table 35 GoS for different user distribution

	NO ADM. CONTROL	TOTAL WITH DROP	HALF POWER	MULTICELL
50 - 50	66.39	24.08	17.7	17.39 [G=1.75%]
25 - 75	64.60	39.14	17.59	16.03 [G=8.86%]

2.15 LOCATION AWARE RESOURCE RESERVATION

2.15.1 Introduction

The demand of wireless multimedia services is growing in recent years. For that reason, the efficiency in bandwidth utilisation has become an important objective in the development of mobile communication systems. On the other hand, recent developments in the positioning

technology in the context of WCDMA systems (e.g. based on Time-Difference-of-Arrival [75] or using the Global Positioning System GPS [76]), provide strong assurance that accurate position measurements will become a viable reality in the near future. The location information obtained by these techniques may provide better predictions of future resource availability, and therefore, it can be exploited to develop more advanced RRM (Radio Resource Management) strategies that increase the system efficiency.

RRM algorithms (admission control, congestion control, power control, etc. [28]) determine how the available radio resources of the system must be used and shared in an efficient way among the different users. In particular, the admission control must decide whether to accept or reject connection requests depending on system load estimation. In a given cell, an admission request may come either from a user that generates a new connection or a user connected to a neighbouring cell that demands a handover to this cell. In wireless networks, it is well-known the existence of a trade-off between minimizing the blocking probability (i.e. the probability of rejecting a new connection request) and minimizing the dropping probability of users in handover, i.e. the probability of cutting down a current connection because the user QoS (Quality of Service) requirements cannot be guaranteed. In terms of quality of service perceived by the user, it is better to reject an admission request instead of dropping an established connection. Therefore, certain priority in the assignment of the radio resources must be given to handover users in order to reduce the dropping rate. Several works dealing with this problem can be found in the open literature. In [77] a call admission control is proposed in order to reduce the number of dropped calls. The algorithm defines a certain reservation region in order to reduce the dropping probability rate at expenses of increasing the blocking probability. Similarly, [78]-[82] propose reservation algorithms in order to assure service to users in handover. In particular, [78][79] propose adaptive reservation algorithms to control the size of the reservation capacity according to the number of soft handover attempts. By doing this, the reservations are carried out in a more efficient way with respect to fixed reservation strategies [80].

The knowledge of the location and mobility pattern of the users will provide certain information to estimate the future need and availability of the radio resources. In particular, in scenarios with users moving along a road, these location estimations can be more accurate because main road users have usually a straight mobility pattern. Therefore, more accurate predictions of handover requests can be done, and consequently, certain radio resource mechanisms can be triggered in order to assure available resources for the handover procedure, increasing the efficiency in the use of the system resources. In this respect, [81][82] propose handoff prioritization algorithms based on predictions and estimations of the future mobile locations.

On the other hand, another feature of WCDMA systems is the ability to support different types of services and user profiles. Then, it is usual that some users (i.e. business users) should receive from the system a certain degree of priority with respect to other users (i.e. consumer users) because of contractual commitments. Under this framework, in this section we propose a location-aware radio resource reservation algorithm that, by making use of location information, is able to provide an efficient use of the radio resources to ensure the QoS constraints of business users while introducing, in case, minimum degradation in the performance of consumer users. The results obtained with the proposed resource reservation algorithm will be compared to the case where no reservation is carried out in terms of blocking probability and dropping probability. In order to optimise the existing trade-off between blocking and dropping probabilities, the impact of the main parameters of the proposed algorithm on the system grade of service (GoS) will be analysed.

2.15.2 Resource Reservation Algorithm

The scenario considers a main road with different base stations close to it, as shown in Figure 146. Users are distributed both in the main road and in the rest of the scenario. Moreover, two kinds of users will be considered: business users (higher priority) and consumer users (lower priority). The main objective of the proposed reservation algorithm is to assure service to business users moving along a main road while at the same keeping the service of consumer users at a satisfactory level. To this end, the proposed algorithm defines a certain reservation region around each station, starting at the reservation distance D [meters], as shown in Figure 146. The reservation distance D is always higher than the cell radius R . When a business user in the main road with an established connection reaches the reservation point (i.e. the distance between the user equipment UE and the Base Station BS is lower than D), certain resource reservation will be made to this user. The proposed algorithm considers accurate positioning measures to determine that the user enters the reservation region, which could be obtained with any of the existing location techniques.

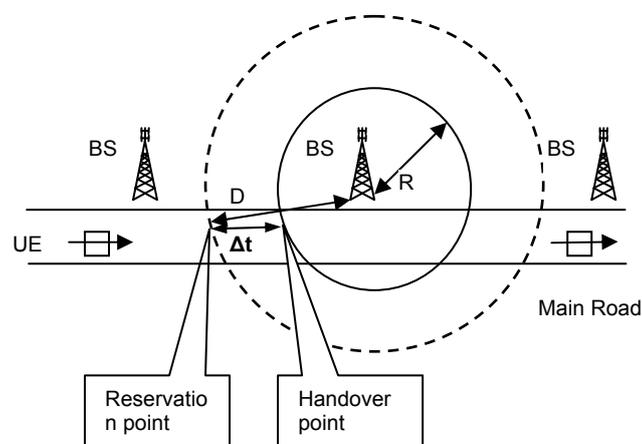


Figure 146 Example of resource reservation

Although the algorithm developed hereinafter could be easily extended to include other radio resources, the proposed algorithm makes the reservation in terms of the number of required downlink OVSF (Orthogonal Variable Spreading Factor [83]) codes, which are the “hard” physical resources consumed in a WCDMA system, and therefore they should be available at the new cell in order for a handover to be admitted. On the contrary, the availability check for other “soft” resources like e.g. transmitted power can be relaxed for handover users in a natural way recalling the soft-capacity in WCDMA systems. The reason is that a handover rejection may lead to a power increase in the new cell, due to an increase in interference, if the user remains connected to the old cell [28]. For that reason, the proposed algorithm illustrated in Figure 147 and Figure 148, considers only code availability check.

Each transmission in the downlink direction makes use of a channelisation code selected from the OVSF code tree. The number of available codes coincides with the Spreading Factor (i.e. there are 4 codes with SF=4, 8 with SF=8, and so on). When a main road business user reaches the reservation distance of a given cell, OVSF code availability check ($C_i + \Delta C < C_{max}$) will be carried out in this cell as shown in Figure 147. C_i is the total amount of codes already used by all the i users connected to this base station, ΔC denotes the resource that must be given to the user which is being reserved and C_{max} is the maximum number of codes. All the quantities refer to the number of codes with SF=512 (i.e. the minimum bit rate), so that if a user transmits with higher bit rate, in terms of code occupation, it is equivalent to occupying a higher amount of codes (e.g. if a user transmits with SF=32 it is equivalent of occupying 16 codes with SF=512). If there are enough available codes in this

cell, (i.e. $C_i + \Delta C < C_{max}$), a code reservation ΔC is carried out for the current user in order to assure resources for the future handover request. When the user finally requests the handover, it is accepted in the new cell, as shown in Figure 148.

On the other hand, if there is no code availability to make the code reservation to this user, the algorithm will firstly reduce the admission threshold C_{max}' for users demanding a new connection request in this base station, as shown in Figure 147, in order to make room for the reservation in the future during the time Δt before the handover, see Figure 146. Due to the dynamics of the system, a reduction in the admission threshold C_{max}' may provide enough available codes for the incoming user by blocking certain number of new connection requests, as long as some users end their connections.

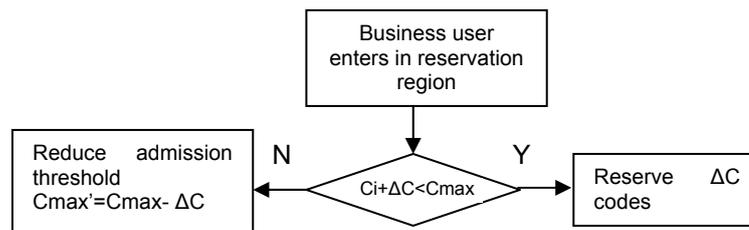


Figure 147 Reservation algorithm for Business users

Finally, when the handover is requested, the user will be accepted, either if there are ΔC available codes or if it is possible to make room for the user by dropping some less priority users (i.e. consumer users), as shown in Figure 148.

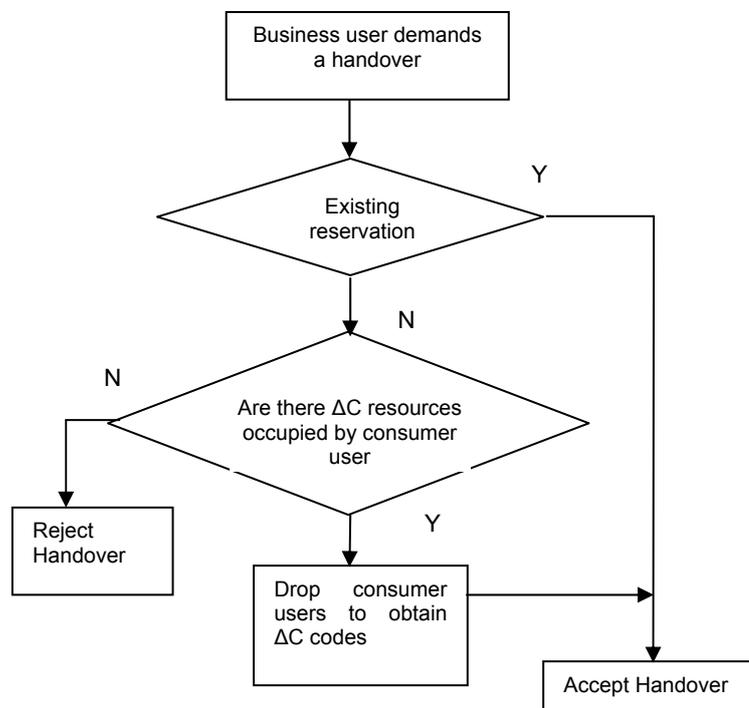


Figure 148 Handover algorithm for business users.

2.15.3 Simulation model

The considered cell layout consists of 7 cells located along a main road with base spacing of 1800m. The base stations have been numbered as shown in Figure 149. Two types of users are considered. On the one hand, a group of conversational consumer users (64kbps CBR) have been located in a rectangular region (i.e. a building) whose position and user mobility pattern can be chosen at the beginning of the simulation. In our simulations, these users will move randomly at 3km/h inside the building. On the other hand, several business conversational (64kbps CBR) or streaming users (384kbps CBR) will be considered. These users have been distributed uniformly along a main road, as shown in Figure 149. These users move following a straight trajectory (from left to right in Figure 149) with speed 50km/h. The main simulation parameters are shown in Table 36.

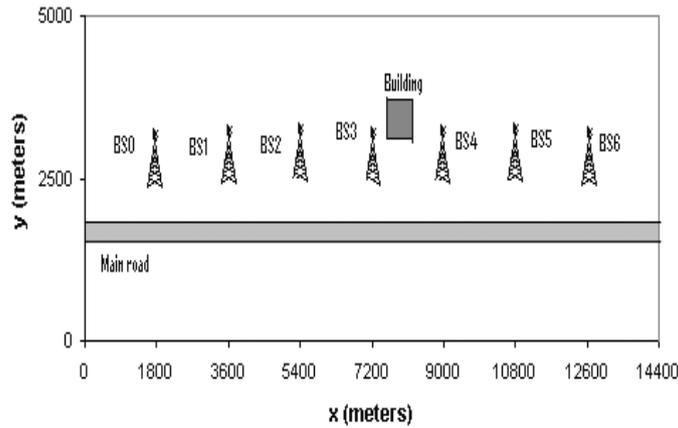


Figure 149 Cell layout

Table 36 Simulation parameters

Parameter	Value
Chip rate W	3.84 Mcps
Frame duration	10 ms
BS parameters	
Cell type	Omnidirectional
Maximum DL power P_{max}	43 dBm
Pilot and common control channels power P_c	30 dBm
Thermal noise	-106 dBm
Orthogonality factor	0.4
Handover parameters	
Active Set maximum size	1
Replacement hysteresis	1dB
Time to trigger HO	0.5 s
QoS parameters	
Eb/No target	4.36dB

2.15.4 Results

In order to evaluate the proposed reservation strategy, different algorithms are presented for comparison purposes:

- *Reference algorithm*: In this case, no resource reservation for business users moving along the road is done. Moreover, the dropping of consumer users is not considered when there are not available resources to accept the handover of a business user.
- *Consumer dropping based algorithm*: In this case, no resource reservation for business users moving along the road is done. However, when there are not enough resources to

accept the handover request of one of these users, a number of consumer users are dropped to make room for the business users.

- *Resource Reservation algorithm*: In this case, a reservation to business users moving along the main road and reach the reservation point is done. If necessary (see Figure 148) a number of consumer users are dropped to make room for business users.

In Table 37, a comparison of the different algorithms is presented in terms of blocking and dropping probability for business and consumer users. A total number of 125 conversational users (64kbps CBR) have been considered. 15% of these users are hotspot users (inside the building). The rest of users move along the main road. As shown, the *Consumer dropping based algorithm* is able to reduce the dropping of business users, with respect to the *Reference algorithm* at the expense of a high increase in the dropping of consumer users. This dropping can be reduced with the proposed reservation strategy. Moreover, the impact of the reservation distance in the *Resource Reservation algorithm* is analysed. As shown in Table 37, a too low value of the reservation distance will cause a high dropping probability of consumer building users. On the other hand, for high values of the reservation distance D (m), high number of users will fall in the reservation region, which will reduce the admission threshold C_{max} . This will reduce the dropping probability of consumer users at expenses of increasing the blocking.

Table 37. Comparison of the different algorithms with 125 users

		Blocking probability (%)	Dropping Business users (%)	Dropping Consumer users (%)
	Reference Algorithm	2.01	2.13	≈ 0
	Consumer dropping based algorithm	2.08	≈ 0	5.32
Resource Reservation algorithm	D=920	2.69	≈ 0	3.9
	D=1100	5.84	≈ 0	2.19
	D=1300	10.44	≈ 0	≈ 0
	D=1500	13.09	≈ 0	≈ 0
	D=1700	22.03	≈ 0	≈ 0

In order to account for the trade-off between blocking degradation and dropping improvement, let define the grade of service as $GoS(\%) = Pb(\%) + 10 * Pd(\%)$ where $Pb(\%)$ is the blocking probability and $Pd(\%)$ is the dropping probability. As shown in Figure 150, there is an optimum value for the reservation distance that minimises the system GoS. For comparison purposes, the GoS obtained with the *Reference Algorithm* (considering 125 users) is 23.31% while in the *Consumer dropping based algorithm* is 55.28%. Notice that this value is higher than the obtained with the *Reference Algorithm* because of the increase in the dropping of consumer users as a consequence of the reduction of business users dropping. By comparing these values with Figure 150, it can be observed that a reservation distance between 1300 and 1700 metres provides lower GoS than the *Reference Algorithm*.

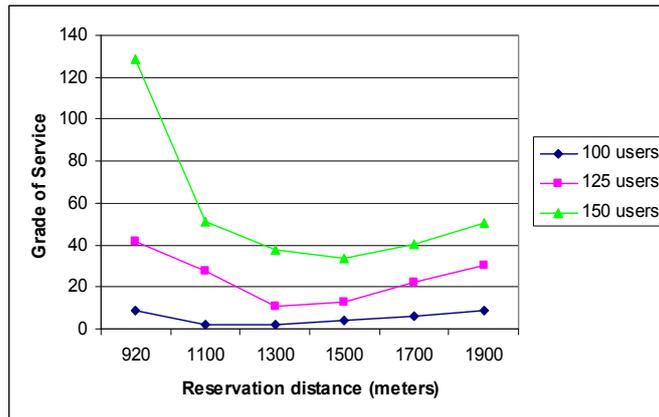


Figure 150. Grade of Service for different reservation distance (mean call duration 120seconds)

Further, certain user characteristics such as the mobile speed or the user call duration will impact on the optimization of the reservation distance. For example, the higher the mobile speed is, the sooner the reservation must be done in order to have enough time to obtain available resources for a main road business user before the handover process starts. Similarly, and focusing on the impact of the call duration, Figure 151, shows the obtained GoS for different reservation distances when the mean call duration is reduced to 20 seconds. In this case, the minimum GoS is obtained for lower values of reservation distance with respect to the case when the mean call duration is 120 seconds (see Figure 150). It is worth noting that a too high value of reservation distance may cause that a main road business user with reservation for a given cell may end its current connection before the handover is eventually made effective (i.e. false reservation). In this situation, the reserved resources for this user have been wasted, reducing the system efficiency. The false reservation probability is shown in Figure 152. Higher reservation distance causes a higher false reservation probability, particularly for short call durations.

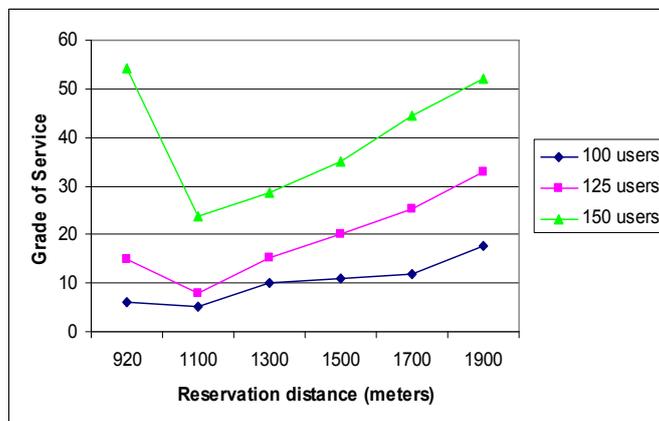


Figure 151 Grade of Service for different reservation distance (mean call duration 20seconds)

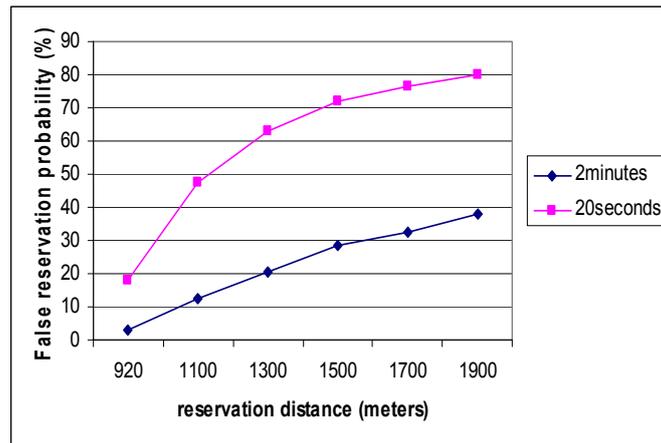


Figure 152 False reservation probability for different reservation distance

In the following, the impact of the service bit rate is presented. To this end, the main road business users are considered to be streaming (384kbps CBR). Consumer building users are conversational (64kbps CBR). When the 384kbps users moving along the main road reach the reservation region, they will demand a higher resource reservation than in the previous scenario, where main road users transmitted at 64kbps. Then, in order to optimise the system efficiency, the resource reservation must be carried out sooner (farther from the base stations) in order to have more time to make room for the high bit rate users before they demand the handover. Figure 153 and Figure 154 show the optimisation of the reservation distance in order to minimise the system GoS. 20 and 30 streaming users (384kbps) have been distributed along the main road. As shown, higher values of the reservation distance are needed with respect to the case of services of 64kbps for the main road users. Moreover, the impact of the call duration is presented, by comparing Figure 153 and Figure 154. As mentioned before, higher mean call duration requires higher reservation distance. Finally, Table 38 summarises the optimization of the reservation distance for different call duration and service bit rate. As stated before, higher call durations require higher reservation distances. Also, the gain $G(\%)$, in terms of GoS, of the proposed algorithm with respect to the *Reference Algorithm* is presented in brackets. It can be observed that significant gains are achieved. The gain is even higher for shorter call durations, since the reservation procedure blocks new call attempts for a shorter period of time.

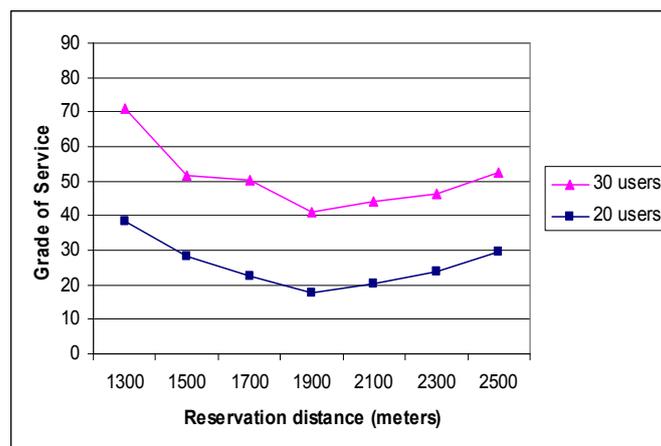


Figure 153 Grade of Service for different reservation distance (call duration 20seconds)

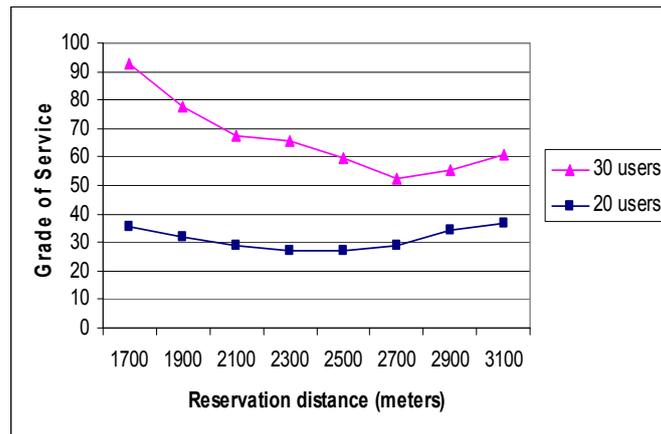


Figure 154 Grade of Service for different reservation distance (call duration 2minutes)

Table 38 Optimum value of the reservation distance for different mean call duration and service bit rate

	64kbps	384kbps
20 seconds	1100m [G=71%]	1900m [G=67%]
2 minutes	1300m [G=55%]	2700m [G=49%]

2.15.5 Conclusions

A resource reservation algorithm that makes use of location-aware techniques in order to assure service to high priority users has been proposed. The proposed algorithm optimizes the existing trade-off between the dropping probability and the blocking probability in a WCDMA system. The reservation of certain resources for handover users reduces the number of dropped connections at expense of certain increase in the blocking probability of new connection requests. The proposed algorithm takes advantage of the predictability in the movement of users along a main road in order to determine the most adequate instant of time when the resource reservation for handover users should be made. A too large reservation region may cause that a handover user ends its connection before starting the handover procedure, resulting in a high false reservation ratio. Moreover, the blocking probability of new connection requests would increase because a large number of resources would be devoted to reservations for users inside the reservation region. On the other hand, a too small reservation region increases the number of handover failures (i.e. the dropping ratio) because there is not time enough to obtain the available resources. Then, an optimization of the reservation distance has been made by minimizing the system GoS. Moreover, the impact of the user call duration and service bit rate on the proposed algorithm has been discussed. It has been shown that, for shorter call durations, the reservation distance must be lower in order to reduce the false reservation probability. Similarly, higher service bit rate require higher reservation distance because more time is needed in order to obtain the required resources.

3 RRM ISSUES FOR GERAN

3.1 INTRODUCTION

The evolutionary path from GSM towards UMTS encompasses the exploitation of the General Packet Radio Service (GPRS). QoS management functions enhancing initial best-effort data services need to be integrated to be able to guarantee subscriber and application specific QoS requirements. These QoS functions are based on the aggregation of flows belonging to the same service class and the prioritized admission control and scheduling of these aggregate flows in the radio network.

The radio resource management (RRM) procedures include the functions related to the management of the common transmission resources, e.g. the physical channels and the signalling channels. In general, purpose of these RRM procedures is to establish, maintain and release radio resource connections that allow a point-to-point dialogue between the network and a mobile station with a given QoS.

The focus of the resource allocation problem is placed in Section 3.2 in the admission control sub-problem whose purpose is to calculate which network resources are required to provide the quality of service (QoS) requested, determine if those resources are available, and then reserve those resources. Admission control is performed in association with the Radio Resource Management functions in order to estimate the radio resource requirements within each cell [84]. The AC protocol aims to maximize the number of admitted or in-session traffic sources supported over the wireless medium while guaranteeing their QoS requirements and ensuring that the new connection does not affect the QoS of the connections currently being carried out [85].

In turn, Section 3.3. is devoted to set the value of some parameters, which are not imposed by GPRS specifications, in order to optimize the network and the streaming client performances.

3.2 ADMISSION CONTROL

As stated earlier, different admission control criteria can be taken into account. In our simulations we use an admission control based on available resources, i.e. the available number of available PDCH, and the multiplexing capabilities of the system. The flow chart shown below represents a possible AC procedure when a new incoming GPRS call is generated. First, the system tries to assign the available resources, if any, to the incoming user. These available resources come in the form of free dedicated PDCHs, commutable PDCHs or commutable TCHs. If none of these are free, the system will accept the GPRS connection via multiplexing different TBF in the same slot as long as the system supports this feature. For this purpose, a threshold of maximum number of multiplexed PDCHs per timeslot is implemented (TBF/slot).

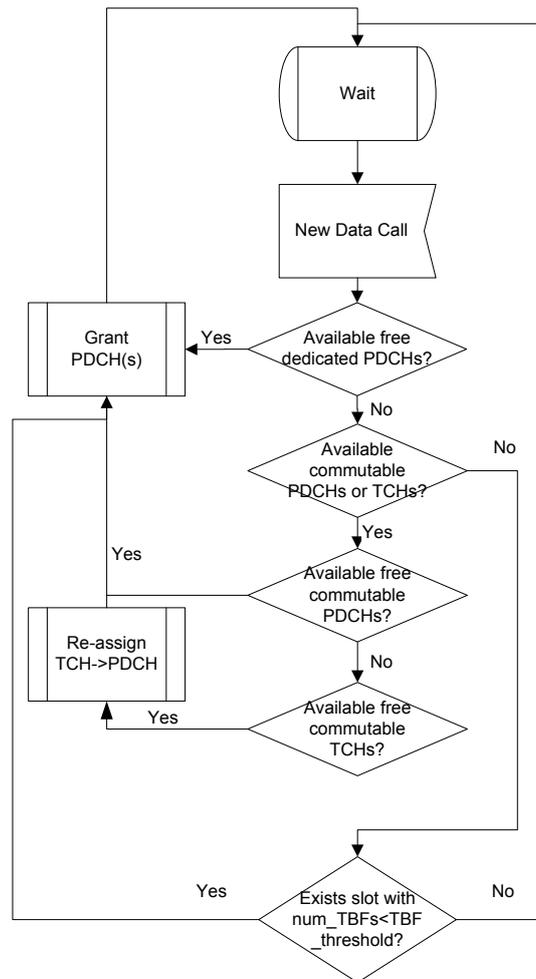


Figure 155 Resource assignment and admission control.

One carrier frequency per cell is used both for circuit switched GSM traffic and for data GPRS traffic. The air interface transmission error model decides whether a received data or control block is error-free or not. For this purpose a set of curves is used mapping C/I values to corresponding block error rate (BLER) values. All four coding schemes, CS-1 to CS-4, can be used. LLC and RLC/MAC are operating in acknowledged mode and multislot capabilities allow up to four downlink slots for one user.

The system measures comprise the downlink average packet delay as a function of number of users per site. This delay considers the time spent from packet generation at the BSS to the reception of the ACK at the BSS. For an admission control measure a maximum allowed number of multiplexed TBFs per slot threshold is used. When varying this threshold, delay performance will also vary, as we will see next.

The following figures show the average packet delay for a set of different number of users per site. Figure 156 below, is plotted considering GPRS users competing for two dedicated PDCHs and one commutable PDCH.

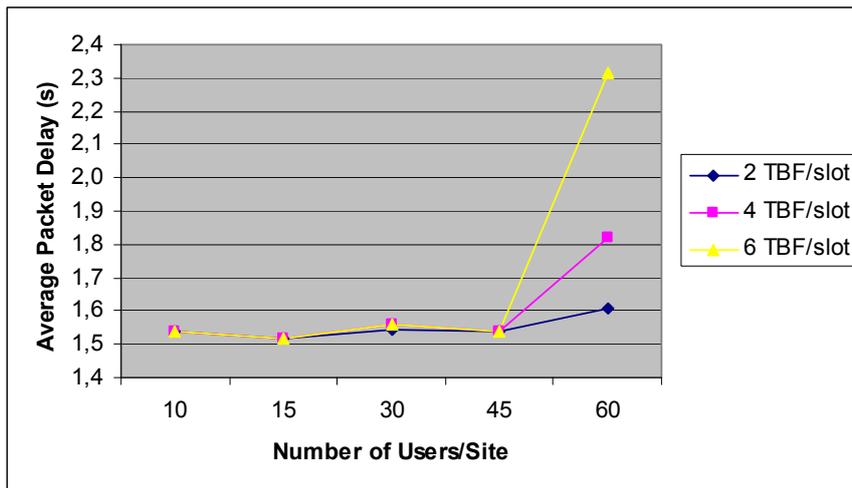


Figure 156 Average Packet Delay for different TBF/slot thresholds (3PDCH's)

For a low number of users per site the system is able to allocate resources regardless the chosen TBF/slot threshold, therefore the experienced delay remains almost constant. As the number of users increases, a system with a conservative (low) TBF/slot threshold will experience lower average packet delay than if a higher threshold is chosen. Higher thresholds will allow more users in the system at the expense of degrading the overall packet delay in the system. This trade-off should be further investigated together with the components of the delay, i.e. to quantify the individual delay contributions from different stages of the packet transmission.

Figure 157, was plotted considering GPRS users competing for a pool of 5 dedicated PDCHs and one commutable PDCH, meaning an increase of designated resources for GPRS users.

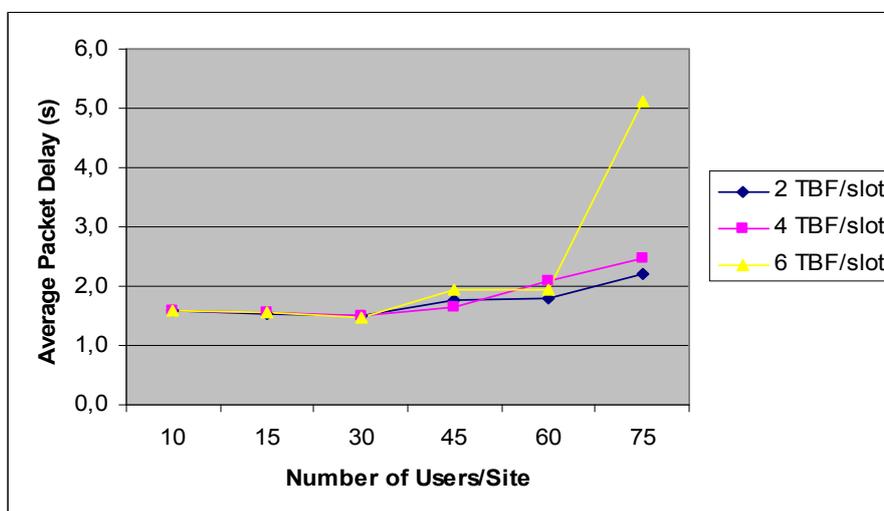


Figure 157 Average Packet Delay for different TBF/slot thresholds (5PDCH's)

Compared to Figure 156, we can see that delay remains almost constant for a higher number of users per site due to an increase in the PDCH pool. However, the system will eventually start congesting when users increase in number thus increasing the average packet delay.

3.3 VIDEO STREAMING OVER GPRS

Simulations with streaming application aim to set the value of some parameters, which are not imposed by GPRS specifications, in order to optimize the network and the streaming

client performances. Choosing the proper bit rate and the right dimension of the buffer are important issues to be solved if we want to guarantee a good usage of radio and client resources.

Secondly, we have to take into account the way these parameters affect the overall end-to-end QoS as it is perceived by the end user. For example, some statistical measured parameters that are relevant in defining the quality of a network delivering a streaming service are: the average number of blocks during a single session, the percentage of lost frames (and/or bytes) and other quantities related to the global network behaviour, e.g. the percentage of interrupted sessions, the percentage of users which are fully satisfied by the service supply and so on.

In particular we found very useful to introduce a new parameter that synthetically defines the satisfaction degree of a user on a scale from 0 to 1; this parameter is linked to the percentage of lost bytes and to the number of blocks. The general idea is that interruptions during a session are perceived as a major clue of a poor quality system, whereas the percentage of lost bytes even if important has a lower weight because it does not prevent from enjoying the video. We also define a "fully satisfied" user of a single session the user whose session has no interruptions at all and with a lost bytes probability $\leq 10^{-5}$ (user satisfaction = 1.0).

The first problem to solve for the right characterization of a specific end-to-end streaming service offered by a GERAN network is the definition of the average bit rate for the video codec. The range of useful values for this parameter must be chosen taking into account both that the capacity of the Radio Access Bearer cannot be exceeded and the codec bit rate must be as high as possible in order to reach the highest video quality.

In our simulation experiments each user may use a terminal with multislot capabilities so that four time slots are allocated in the downlink (and one in the uplink). Since the coding scheme that we adopted is CS-2, the maximum net RLC/MAC data rate is $12 \times 4 = 48$ kbps. When we consider the application layer this value must be of course reduced because of the overheads, segmentation, etc. introduced by the different protocol layers.

The situation of blocking for higher bit rates shows that the blockage is nearly irrelevant for codec bit rate less or equal to 28 kbps whereas for greater bit rates it grows with a linear law. As a consequence, most of subsequent simulation experiments are done by using codec bit rate NOT greater than 32 kbps, which is assumed to be the limit value for bit rates of the application codec.

Another important degree of freedom, which has been tuned in our simulation experiments, is the RTP packet size, that is the amount of segmentation introduced by RTP protocol.

The application throughput is nearly constant and equal to its maximum value when the dimension of RTP packet is greater than 400 bytes. If we reduce this dimension up to 100 bytes the throughput decreases from its maximum value and the effects of a too fine segmentation in packets are too strong. This is of course more evident when the codec bit rate is higher, whereas for smaller dimension of RTP packets the dimension of the packet may be still reduced if the overhead is still acceptable.

On the other hand, the selection of adequate RTP packet size too involves the analysis of delay variation (jitter) because as the RTP packet size decreases the jitter decreases as well. In that sense it is an advantage the selection of very small sizes. However, it is worth recalling that it also produces a significant reduction of offered throughput (as explained before), which is not desirable.

Experiments has shown that by configuring a dimension of RTP packet size equal to 500 bytes will provide high offered throughput while maintaining low jitter level.

In our simulations experiments, the optional functionality for compression of RTP/UDP/IP headers is also supported in order to reduce the effects of the considerable header overhead for voice and video applications. The header compression scheme applied is the called RObust Header Compression RoHC, which allows reducing the RTP/UDP/IP header size from 40 bytes to 3 bytes.

The major feature is the diminution of the delta range between offered throughputs and required throughputs. This means that the obtained application throughput is closer to the codec bit rate than the first case without RoHC option. It is evident that the process of compressing protocol headers saves the bandwidth and exploits the radio resources efficiently.

Another important parameter to be considered for the support of video streaming services is the limitation of the receiver initial buffering time. Since the determination of service acceptability or unacceptability significantly depends on the delay variation called jitter, the size of this buffer will be critical to decide whether a given jitter is acceptable or not.

Therefore, the compensation of jitter effects will be done throughout the receiver buffer. This element will store during a period as much as necessary data before reproduction for jitter compensation and to reduce effects of disparities between reproduction bit rate and offered bit rate. For that reason, it is necessary to preserve the limit of this initial buffering time because the selection of very small values can result on periods of empty buffer (reproduction blocks), and choices of a lot of seconds can imply expensive implementations (memory requirements, etc).

Furthermore, from the network viewpoint, the functionalities provided by this mechanism along with the RLC retransmissions reduce the requirements of radio resources allocation for streaming sessions in terms of delays and jitter by satisfying the requested QoS levels.

Hence, taking into account that codec bit rates from 24 up to 32 kbps have been defined as the acceptable bit rates for video streaming applications over GRPS networks, so the subsequent experiments only involve the variation of initial buffering time for these specific bit rates.

In order to determine the minimum initial buffering time, we must assume an acceptable level for user satisfaction. From the results we appreciate it can be chosen according to the requested bit rate. Thus, at 24 kbps, the percentage of fully satisfied users is always greater than 95 percentage due to the fact that the network is almost able to offer this reproduction rate. Consequently, the jitter effects are easily compensated with a minimum buffering time of 1 second.

On the other hand, at 28 kbps the number of blocks per session is more elevated, giving a smaller percentage of user satisfaction. However, these are still very good levels to provide the video streaming service. Therefore, to achieve more than 95% of fully satisfied users preserving a small number of numbers of blocks per session, it is enough to configure the initial buffering time to 3 seconds.

Finally, for codec bit rate of 32 kbps less than 1 block per session is obtained when the buffering time is either 4 or 5 seconds. Now, taking into account the desired levels of user satisfaction (about 90%) we can conclude a safe value for the initial buffering time must be 5 seconds.

Another considerable parameter affecting the optimal setting of the pre-jitter buffer length is the speed of the user with an active video streaming session. In this sense, it is useful to consider that the results described in the previous section were achieved in a low-mobility scenario (mean value of user's speed equals to 3 km/h). If an higher velocity is considered (i.e. 50 km/h), the buffered data fluctuations increases due to the different values of spatial shadowing affecting the transfer delays of data packets.

Nowadays more mobile users are demanding the video streaming solutions, which encourage mobile operators to meet the needs, the requirements and looking for a high quality user' media experience. Since this fundamental consideration is taken into account on the considered radio access network's characterization, a high load environment will be simulated.

Based on previous network and service configuration parameters the initiative consists on setting them, and subsequently adding an average number of active users per cell in order to observe the performance behavior. Therefore, because of the limited radio resources on GPRS networks, the offered throughput per user in downlink progressively decreases as the average active users increase. The simulation was executed with mobile users requesting for a video streaming sequence at 24, 28 and 32 kbps.

If also voice users are considered in the simulations and a shortage of traffic channels in the CS domain occurs, a request for PDCH pre-emption is sent by the system to the PS domain. According to this standard mechanism of the GPRS/EGPRS system, the mean value of the throughput offered to a video streaming session (PS domain) decreases with the increase of the number of voice users (CS domain). The decrease level of the mean value of the throughput offered to a video streaming session involves also a certain decrease of the QoS.

4 RRM ISSUES FOR WLAN

4.1 INTRODUCTION

The IEEE 802.11 standard for wireless local area networking presented in 1999 opened new possibilities to local area networking due to its flexibility and convenience. First implementations of WLAN technology focused on networks with best effort traffic like e-mail, web browsing or ftp. However rapid deployment, popularity and effectiveness of this new technology with fast growth of multimedia application made WLAN users interested in supporting more sophisticated applications like audio or video streams. Although IEEE 802.11 is working well with best effort traffic it is unable to guarantee real time traffic requirements like delay, jitter or error rate. To answer this necessity Quality of Service (QoS) enhancements became a prominent research issue. Performed investigations resulted in a numerous proposals [88],[89],[90] with final conclusion that QoS provisioning can be supported by 802.11 Medium Access Control (MAC) layer by adding to it service differentiation mechanism. To develop and standardise a unique QoS framework for IEEE 802.11 a QoS subgroup (Task Group E) has been formed.

On the other hand, wireless LANs of IEEE 802.11 family are "non-blocking" systems: every new user entering into the system tries to access the shared medium for transmitting and receiving data. Consequently, the quality of service of all the users in the BSS degrades, in terms of throughput, delays, jitter and transmission errors. This is particularly true for the Wireless LANs of IEEE 802.11 family, in which the way of work of the Medium Access Control (MAC) layer, according to the so called Distributed Coordination Function (DCF) access method, make the system's performances very sensitive respect to the number of stations and their traffic profiles. On the basis of these considerations, it should be evident that without any policy for blocking the admission of new users when specific load conditions

accur, wireless LAN systems cannot succeed to guarantee whatever profile of Quality of Service and only best effort services can be supported.

In this framework, the WLAN-related issues dealt in EVEREST so far and reported below correspond to:

- An analytical model for an admission control algorithm, accompanied by results on admission control regions for a mix of services
- A short to medium term proposal for QoS enhancements on 802.11b based on the Hierarchical Token Bucket algorithm
- A more long term evaluation for QoS enhancements on 802.11e

4.2 ADMISSION CONTROL FOR IEEE 802.11A/B/G

This section summarizes the results coming from the work carried out to introduce an appropriate Admission Control policy for IEEE 802.11 family of standards [1]. The AC policy can be applied in order to make the wireless LANs able to support real-time data services (i.e. conversational and streaming classes) requesting a minimum amount of bandwidth.

The proposed Admission Control (AC) policy consists in three main steps:

- Step 1: throughput offered by the BSS to each real-time user considered by the EVEREST project is evaluated by means of an analytical model (4.2.1).
- Step 2: results coming from the analytical model are exploited in order to identify the maximum number of users for each class (i.e. services mix) that the considered WLAN BSS can support ("Capacity region",4.2.2);
- Step 3: the algorithm in charge of the AC policy can be based on the capacity region related to the WLAN BSS (4.2.3);

4.2.1 Analytical model to estimate the performance of MAC 802.11 DCF for real-time services (step 1)

The analytical model used to derive the Admission Control policy is based mostly on [91]; the extensions made are: full distinction of traffic patterns in uplink and downlink, and the exact calculation of a parameter that was only approximated in the precursor work.

The analytical model has been applied considering the real-time services envisaged by EVEREST scenario [24]; the bit rates and packet lengths, as well as the requested bandwidth for the above mentioned services are summarized by Table 39:

Table 39 Video Telephony and Video Streaming characteristics

Service	Bit rate Packet length	Requested QoS (guaranteed bit rate)
Video Telephony	UL:64/DL:64 kbps UL:1024/DL:1024 bytes	UL:58/DL:58 kbps
Video Streaming Business	UL:16/DL:128 kbps UL:128/DL:2048 bytes	UL:8/DL:112 kbps
Video Streaming Consumer	UL:16/DL:64 kbps UL:128/DL:1024 bytes	UL:8/DL:58 kbps

According to the considered services and the used parameters, [1], the analytical model is able to provide the results depicted, as example, in the graph below Figure 158.

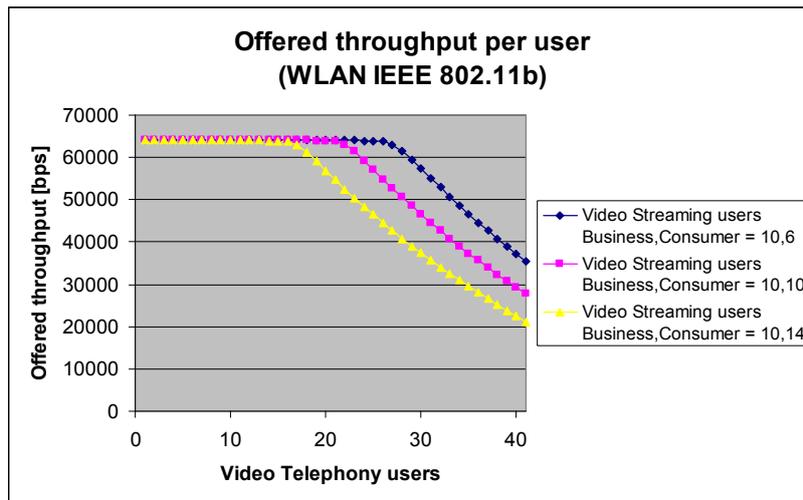


Figure 158 Offered throughput per user (UL/DL) versus number of video telephony users when users of other services are present.

In the next section, it will be described how the results provided by the analytical model like these depicted in Figure 158, can be exploited to determine the capacity region of the wireless LAN system.

4.2.2 Capacity region for the WLAN hot-spot (step 2)

The second step of the Admission Control strategy consists in the identification of the service mixes compliant with a minimum level of offered throughput (uplink and downlink) per user (i.e. the “Capacity region”). The capacity region describes the maximum number of users for each class (i.e. service mixes) that can be supported by the wireless LAN system, exploiting all the available radio resources. Figure 159 provides an example of the capacity region offered by a IEEE 802.11b hot-spot, when 10 video streaming consumer users are present into the system:

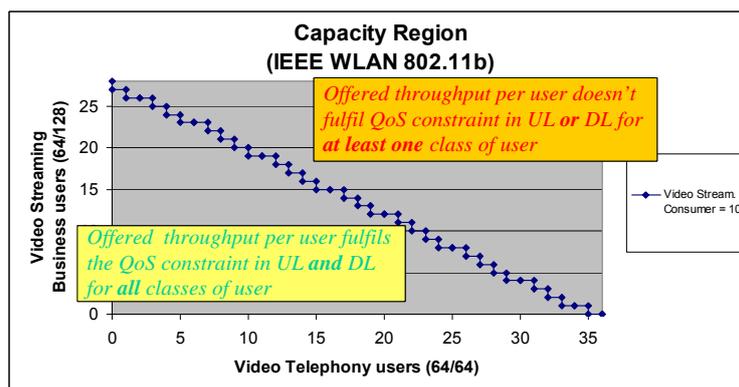


Figure 159 Example of capacity region offered by an 802.11b hot-spot when video streaming business (64/128 Kbit/s) users and video telephony users (64/64 Kbit/s) are considered (10 video streaming consumer users are supposed).

As depicted in Figure 159, the blue line split the plane into two regions: if the load point (i.e. number of users of the two considered services) are above the curve, the quality of service constrain is not satisfied since the throughput that the wireless LAN system can offer is below the minimum specified for at least one class of user. On the contrary, if the point identified by the total amount of users for the two classes is under the curve, the quality of service constrain is fully satisfied for each class of users. The above mentioned example

considers only two classes of users for simplicity matters. In the case of EVEREST real-time services, the already specified (Table 39) three classes of service must be considered and the following capacity region can be obtained:

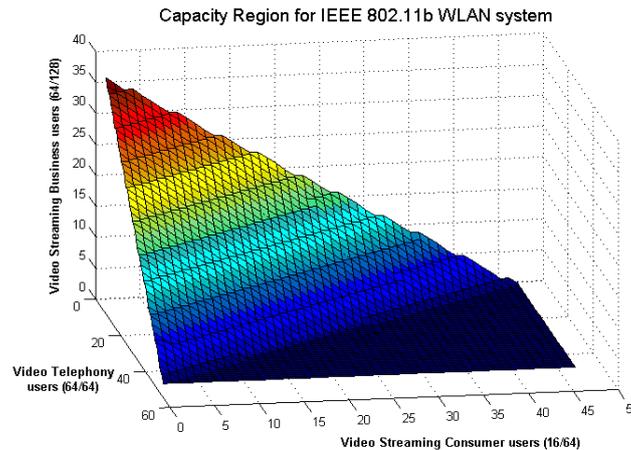


Figure 160 Capacity region offered by an 802.11b hot-spot for the real-time services envisaged by EVEREST scenario.

4.2.3 Capacity region based Admission Control (step 3)

The capacity region reported in the previous section (4.2.2), can be used by an Admission Control algorithm to keep the number of active users for each class of services within the range of values able to respect the QoS constrains.

More in detail, the Admission Control algorithm (step 3), based on the capacity region, should perform the following steps:

- ✓ For every stations requesting a new association to the wireless LAN Access Point, retrieve the type of service requested by the user;
 - ✓ If the acceptance of the new user brings the load work point of the WLAN BSS outside the QoS region (according to the a-priori evaluated capacity region), deny the request of association;
 - ✓ If the new user can be admitted without compromising the minimum level of offered throughput per each class of users, accept the request of association and update the load work point of the system (i.e. number of active users within the systems).
- ✓ For every stations requesting a deassociation from the wireless LAN Access Point, update the load work point of the system (i.e. number of active users within the systems);

It is clear how the results provided by the analytical model can be exploited by an Admission Control algorithm so that the maximum number of users the hot-spot can manage will be never exceeded. This can be realized preventing a new user from entering into the system when the maximum amount of users has been reached, so that the offered throughput per each user remains above the specified threshold. From an operative point of view, it is very important to remark that the IEEE 802.11 family of standards foresees explicitly that a station should be authorized by the Access Point before exchanging data, by means of the "association procedure". According to this procedure, a station that wish to enter into the wireless LAN system, sends a "request of association" frame to the Access Point, waiting for the response. Taking advantage of this fact, the response of the request of association can depend on the decision of the above mentioned AC algorithm. In this way, the wireless LAN system can be used keeping the control of how many users can use the system, respecting a minimum QoS level in terms of guaranteed bandwidth.

4.2.4 Validation of the analytical model for the performance evaluation of IEEE 802.11a/b/g WLAN

This section summarizes the most relevant results coming from the work carried out in order to validate by means of simulations the analytical model for the performance evaluation of WLANs IEEE 802.11a/b/g. The analytical model was presented in [1], also describing how to exploit the results in order to derive an admission control policy for real-time services offered through a WLAN hot-spot. Within the context of the work, also the behaviour of the RTS/CTS mode of the Distributed Coordination Function (DCF) was investigated. Moreover, additional performance statistics respect to the throughput offered by the hot-spot was collected for the real-time services envisaged by the EVEREST project.

4.2.4.1 Introduction

In recent years, much interest has been involved in the design of wireless networks for local area communication [92],[93]. Study group 802.11 was formed under IEEE project 802 to recommend an international standard for Wireless Local Area Networks (WLANs). The final version of the standard has recently appeared [94], and provides detailed Medium Access Control (MAC) and Physical layer (PHY) specification for WLANs.

In the 802.11 protocol, the fundamental mechanism to access the medium is called Distributed Coordination Function (DCF). This is a random access scheme, based on the Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) protocol.

Retransmission of collided packets is managed according to binary exponential backoff rules. The standard also defines an optional Point Coordination Function (PCF), which is a centralized MAC protocol able to support collision free and time bounded services. The analytical model presented in [1] is able to evaluate the performance of the DCF scheme.

DCF describes two techniques to employ for packet transmission. The default scheme is a “two-way” handshaking technique called Basic Access mechanism. This mechanism is characterized by the immediate transmission of a positive acknowledgement (ACK) by the destination station, upon successful reception of a packet transmitted by the sender station. Explicit transmission of an ACK is required since, in the wireless medium, a transmitter cannot determine if a packet is successfully received by listening to its own transmission.

In addition to the Basic Access, an optional “four-way” handshaking technique, known as Request-To-Send/Clear-To-Send (RTS/CTS) mechanism, has been standardized. According to this scheme, before transmitting a packet, a station operating in RTS/CTS mode “reserves” the channel by sending a special Request-To-Send short frame. The destination station acknowledges the receipt of an RTS frame by sending back a Clear-To-Send frame, after which normal packet transmission and ACK response occurs. Since collision may occur only on the RTS frame, and it is detected by the lack of CTS response, the RTS/CTS mechanism allows to increase the system performance by reducing the duration of a collision when long messages are transmitted. On the contrary, in the case of services that use short length packets, this mechanism could downgrade the overall performance due to the overhead of RTS/CTS frame transmission. As an important side effect, the RTS/CTS scheme designed in the 802.11 protocol is also suited to eliminate the so called problem of Hidden Terminals [95], which occurs when pairs of mobile stations result to be unable to hear each other.

4.2.4.2 Previous work

A number of papers have studied the throughput performance of 802.11 DCF MAC for both the BASIC and RTS/CTS access modes through simulation [96], or analytical models with simplifying assumptions on the DCF MAC layer operations and/or the traffic conditions in order to enable mathematical analysis. A detailed mathematical model for the DCF has been

developed by Bianchi [97] and slightly extended by Wu et al. [98]. The key approximation made in these models that enables a relatively simple Markov chain analysis is the assumption of independent transmissions by different flows, as well as constant and independent collision probabilities, regardless of the number of erroneous transmissions already experienced.

In [99] the work mentioned above is extended in two directions. First, authors further elaborate on Bianchi's packet level model [97] by integrating the various modeling enhancements on the physical and MAC layer proposed by other authors into one single DCF performance model, still allowing analytical treatment. The second extension covers the practical situation that the number of active stations is not constant (as in e.g. [97],[98]) but varies in time due to the random user behavior, i.e. the initiation and completion of flow transfers. In order to enable mathematical analysis of flow throughputs and transfer delays in the system with the extensions mentioned above, in [99] an integrated packet/flow level modeling approach is proposed. In particular, from the flow level point of view, the WLAN is considered as a queuing system with Poisson flow arrivals and a Processor Sharing (PS) type of service discipline, which reflects the IEEE 802.11 DCF MAC design principle of distributing the transmission capacity fairly among the active flows.

The Bianchi's model [97] has been extended in [100] by eliminating the hypothesis of saturation condition and that regarding the equivalence of mobile stations in terms of the offered load to the MAC layer:

In saturation condition each station has, in any time, at least one packet ready to be sent. This means that the analysis is independent of packet flow arrivals, because whichever is the statistical distribution of data arrivals at the access sublayer, in any case there is always a packet in head of line. Then, the saturation throughput is independent on the frequency arrival of the packets at the MAC from the upper layers. An analysis without the hypothesis of saturation condition enables the evaluation of throughput with respect to the traffic pattern produced; this, besides analyzing the wireless LAN network in several load conditions, is a key point for analyzing networks with stations having different traffic patterns.

In our previous contribution [1], we proposed a further extension of [100] that consisted in the full distinction of traffic patterns in uplink and downlink and in the exact calculation of a parameter that was only approximated in the precursor work.

4.2.4.3 Overview of the Simulation work

4.2.4.3.1 Main characteristics of the simulations

The model presented in [1] has been validated by means of event-driven simulations of the WLAN IEEE 802.11 architecture.

In the IEEE 802.11 standard, two operation modes are foreseen ('ad-hoc network' or 'infrastructure'): in the simulations only the infrastructure architecture has been considered; this means that in the scenario there is always at least an Access Point that handles communications between STAs.

According to the IEEE 802.11 standard, the architectural view of the two lowest layers of the ISO/IEC basic reference model of Open Systems Interconnection (OSI) includes the MAC of the Data Link Layer, the PHY and the Station Management Entity (SME).

With respect to the MAC layer, in the carried out simulations it has been taken into account the DCF (Distributed Coordination Function) mode, whether with basic access method or

with the optional RTS/CTS handshaking technique, whereas the optional PCF (Point Coordination Function) has not been considered.

With respect to the features of the physical layer, the following items have to be remarked:

- In the simulator both standards 802.11b and 802.11a are considered with all physical modes up to 54 Mbit/s; the optional transmission with Short Preamble is not implemented.
- The radio link chain is not simulated and it is taken into account in terms of BER (Bit Error Rate) or PER (Packet Error Rate) versus C/I relations coming from link simulators; in particular, typical 802.11b figures of PER versus C/I reported in the literature was used in order to decide if a frame has been received correctly.

Moreover, all procedures related to scanning, authentication, association, rate switching, handover are implemented in the SME.

4.2.4.3.2 Comparison with the mathematical model

Since the simulations have been carried out according to the 802.11 standard, some approximations generally used in the analytical models have been overcome; in particular in this paragraph we outline some parameters that affect the performance offered by the WLAN hot-spot and are the key factors for explaining the differences between the simulation and the analytical model performance:

- ACKTimeout is defined in the IEEE 802.11 standard: a source STA, after transmitting an MPDU that requires an ACK frame as response, shall wait for an ACKTimeout amount of time without receiving an ACK frame before concluding that the transmission failed. This means that a source STA after a failed transmission shall invoke its backoff procedure for contending the channel upon expiration of ACKTimeout interval. This time interval, defined in the standard greater than SIFS+PLCP-Preamble+ PLCP-Header (202 μ s), gets the network performance worse but is not take into account in the analytical models.
- CTSTimeout is defined in the IEEE 802.11 standard for the stations using a RTS/CTS mechanism: a source STA, after transmitting a RTS frame that requires a CTS frame as response, shall wait for a CTSTimeout amount of time without receiving it before concluding that the transmission failed. This means that a source STA after a failed RTS/CTS transmission shall invoke its backoff procedure for contending the channel upon expiration of CTSTimeout interval. This time interval, defined in the standard greater than SIFS+PLCP-Preamble+PLCP-Header (202 μ s), affects the network performance in case of RTS/CTS access and is not take into account in the analytical models.
- EIFS is the amount of time used by the DCF when the PHY has indicated to the MAC that a frame transmission was begun that did not result in the correct reception of a complete MAC frame with a correct FCS value⁷. The EIFS interval shall begin following indication by the PHY that the medium is idle after detection of the erroneous frame, without regard to the virtual CS mechanism. The EIFS is defined to provide enough time for another STA to acknowledge what was, to this STA, an incorrectly received frame before this STA commences transmission. EIFS for the standard 802.11b equals 364 μ s because is derived in the standard as the sum of SIFS plus DIFS plus the length of time to transmit an ACK control frame at 1Mbit/s. In the analytical models EIFS is not used and a STA after an incorrectly received frame shall invoke its backoff procedure for contending the channel upon expiration of DIFS (50 μ s) instead of EIFS (364 μ s).

⁷ The reception of an erroneous frame can be due to a collision of two or more frames or simply to the channel loss resulting from path loss and fading effects.

- The beacon is a management frame, periodically sent by the AP for the synchronization of all STAs that are members of the BSS in an infrastructure network. A STA operating in passive scanning mode shall listen to each frequency channel for the beacon frames in order to choose the AP to which start the authentication and the association procedures. As a logical consequence this frame is very important also for the handover procedure. Beacons shall be generated for broadcast transmission by the AP once every "BeaconPeriod" time unit; this time is not defined in the standard but the typical value used by the most important manufacturers is 100 ms. Moreover, all broadcast frames shall be transmitted at one of the rates included in the BSS basic rate set (for 802.11b standard these rates are 1Mbit/s or 2 Mbit/s). The transmission of the beacon is not foreseen in the analytical models.
- All control frames, and consequently all ACK, RTS and CTS frames, as defined in the standard, have to be transmitted at one of the rates in the BSS basic rate set in order to be understood by all STAs in the cell. In the software simulator this rule is maintained whereas in the analytical model, even if apparently there is no restriction on the rate transmission for the control frames, generally these are transmitted at the same rate as the data frames. This means that in the analytical models, when the 11Mbit/s rate is used for transmitting all frames, the performance of the network is better than that specified by the standard.
- As defined in the standard, after transmitting a data or a management frame, if the source STA does not receive an Ack frame, it shall attempt to retransmit the failed frame after performing the backoff procedure and the contention process. At each frame is associated a counter of the retransmission attempts; if the maximum retry limit is reached the frame shall be discarded and retransmissions shall cease. The default maximum retry limit in the standard is 7 whereas in the analytical models usually the frame is retransmitted without limits until it is correctly received by the destination station.

4.2.4.4 Simulated scenarios

Simulations have been carried out in order to analyze the real-time services envisaged by Everest, summarized in Table 40; the comparison between these results and those calculated from the analytical model proposed in [1] will be shown in the next paragraph. All analyses are with ideal channel with no loss due to the path or fading effects; this means that the frames will be discarded only when there is a collision.

The most important parameters that will remain fixed in all simulations are shown in Table 41, whereas the parameters for the cases considered in this report are summarized in Table 42. In particular:

- Case 1 indicates the analysis of the scenarios with the parameters (data rate for control and management frames, EIFS, maximum number of retries) defined as in the standard and with typical value for the Beacon dimension and beacon interval;
- Case 2 indicates that the analysis has been done with some parameters as in the model [1]: EIFS equals DIFS, ACK frames transmitted at 11 Mbit/s, maximum number of retries very high, beacon transmitted at the highest data rate and with a large time interval;
- Case 3, 4 and 5 are intermediate analyses in order to see the effect of each parameter on the overall performance of the system. In particular these cases differ from case 1 for the following parameters:
 - Case 3: ACK frames transmitted at 11 Mbit/s instead of 2 Mbit/s
 - Case 4: ACK frames transmitted at 11 Mbit/s instead of 2 Mbit/s and beacon transmitted at 11 Mbit/s and every 500ms (instead of 100ms). The dimension of the beacon is reduced to 1 byte in order to have only the MAC header and the PLCP header. The AP in any case shall apply the basic medium access rules specified in the standard for transmitting it.

- Case 5: EIFS equals DIFS.

The traffic pattern is a Poisson-distributed process as used in [100] with an inter-arrival frequency depending on the services (bit rate in uplink and downlink).

Table 40 Video Telephony and Video Streaming characteristics

Service	Bit rate [kbit/s]	Packet length [byte]	Requested Qos (Guaranteed bit rate) [kbit/s]
Video Telephony	UL: 64 DL: 64	UL: 1024 DL: 1024	UL: 58 DL: 58
Video Streaming Business	UL: 16 DL: 128	UL: 128 DL: 2048	UL: 8 DL: 112
Video Streaming Consumer	UL: 16 DL: 64	UL: 128 DL: 1024	UL: 8 DL: 58

Table 41 - System parameters

Parameter	Values
B: Overall bandwidth	22 MHz
R: Rate Data	11 Mbit/s
Wmin: Minimum Contention Window	31
Wmax: Maximum Contention Window	1023
L: Buffer Length	10 SDU
HMAC: Length of the MAC header	272 bit
ACK: length of the Ack frame (without header)	112 bit
PLCP Header	48 μ sec
PLCP Preamble	144 μ sec
Slot time	20 μ sec
SIFS	10 μ sec
DIFS	50 μ sec
AckTimeout	202 μ sec

Table 42 - Different cases analyzed

	Rate Ack [Mbit/s]	Rate Beacon [Mbit/s]	Beacon Period [msec]	Beacon Dimension [byte]	Maximum Number of Retransmission	EIFS [μ sec]
Case 1	2	2	100	50	7	364
Case 2	11	11	500	1	700	50
Case 3	11	2	100	50	7	364
Case 4	11	11	500	1	7	364
Case 5	2	2	100	50	7	50

4.2.4.5 Results

4.2.4.5.1 Throughput performance with Basic Access

For the real-time services envisaged by Everest, summarized in Table 40, the comparison for case 1 and case 2 between the simulation results and those calculated from the analytical model proposed in [1] are shown in Figure 161, Figure 162 and Figure 163. As depicted in the figures, the downlink results for case 2 (with most parameters as equal as in the model) are very close to those from the analytical model [1]. The small differences between the two are related to some approximations of the analytical model (for example the Acktimeout and the beacon). With the parameters set according to the standard (Case 1), the performance in downlink of the all services cited in Table 40 deteriorates; however, considering the maximum number of users that can have guaranteed the QoS, the difference is very small:

only 3 users for the Video Streaming Consumer service and 2 users for the Video Streaming Business and Video Telephony services.

In uplink, the agreement between analytical model and software simulation is very good for both cases 1 and 2.

In order to investigate the effect of each parameter on the hot-spot performance the results for case 3, case 4 and case 5 for the Video Streaming Business service are shown in Figure 164. The throughput performance strongly depends on the rate used for transmitting ACK frames and on the EIFS time interval. On the contrary, the beacon transmission features (data-rate, size, beacon period) do not have significant impact on the performance.

In Figure 165 the performance of the hot-spot taking into account three classes of users is shown; in particular the figure shows the throughput offered to each Video Streaming Business user with a fixed number of Video Streaming Consumer (4) and Video Telephony (4) users. As before, you can see a very good agreement between the mathematical model and the Case 2 software simulations and a difference of 2 users between Case 1 and Case 2 at the requested QoS level.

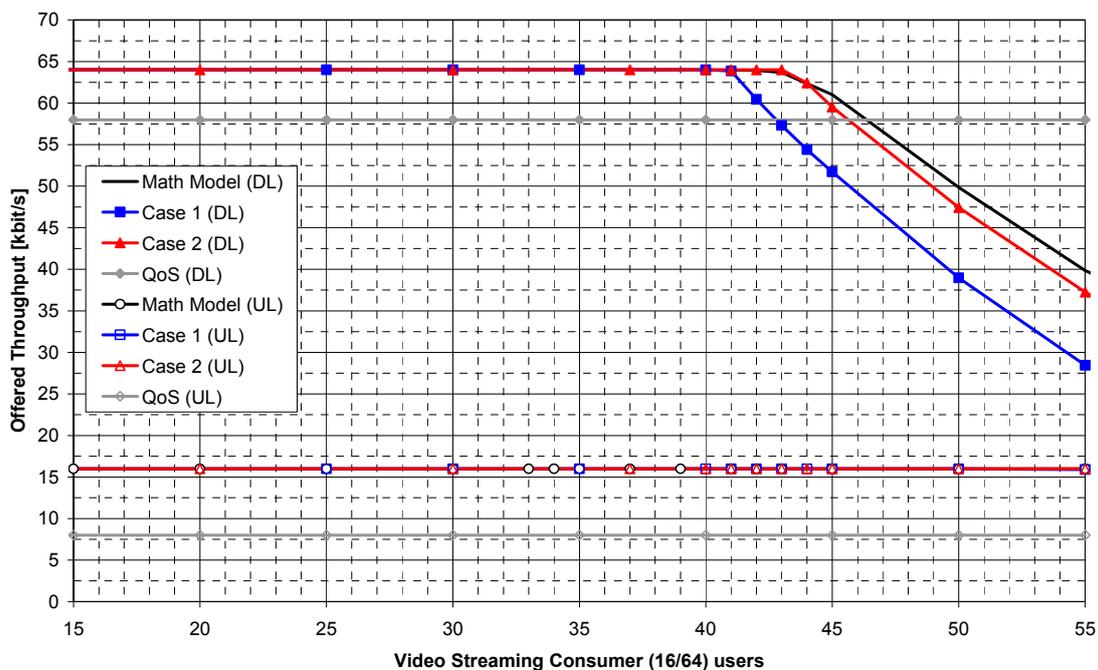


Figure 161 Offered throughput per user (UL/DL) versus number of Video Streaming Consumer users

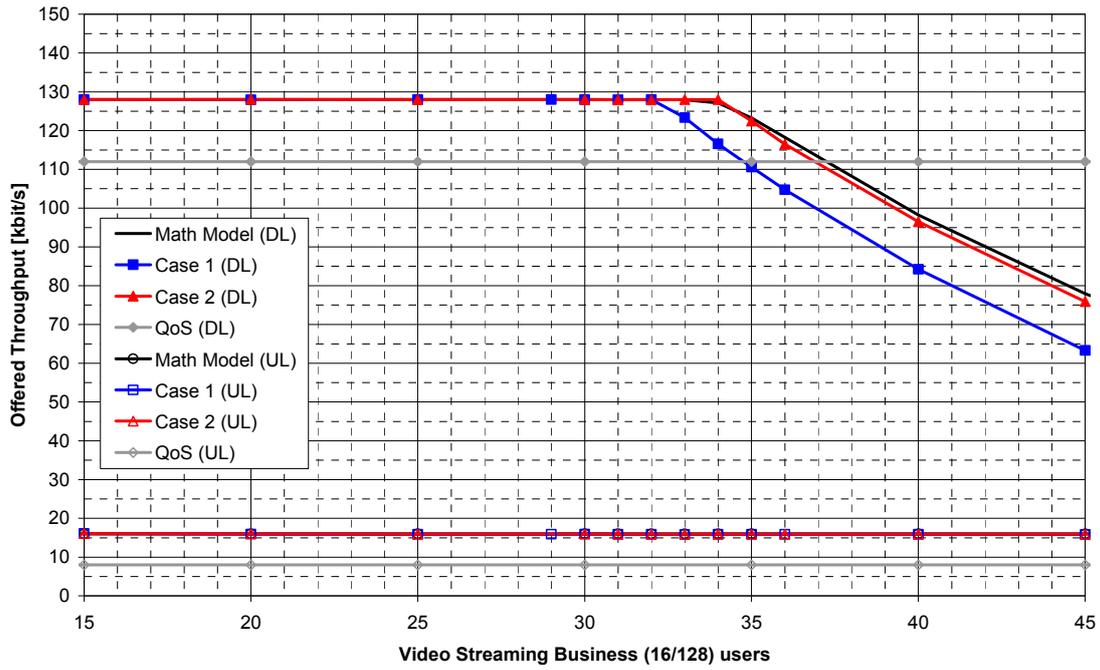


Figure 162 Offered throughput per user (UL/DL) versus number of Video Streaming Business users

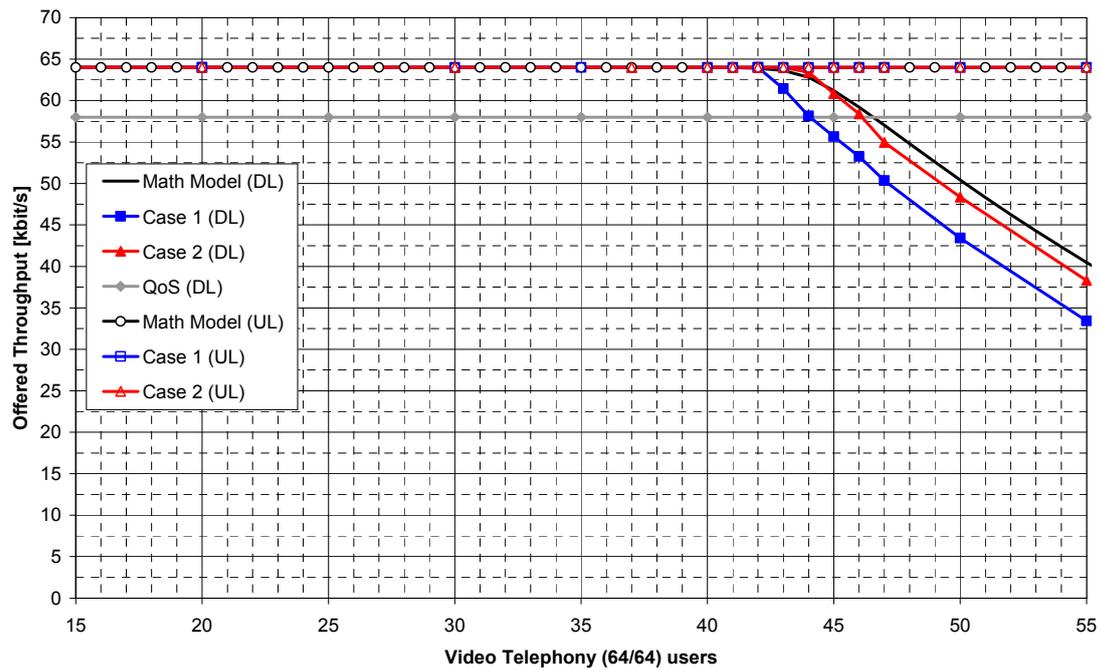


Figure 163 Offered throughput per user (UL/DL) versus number of Video Telephony users

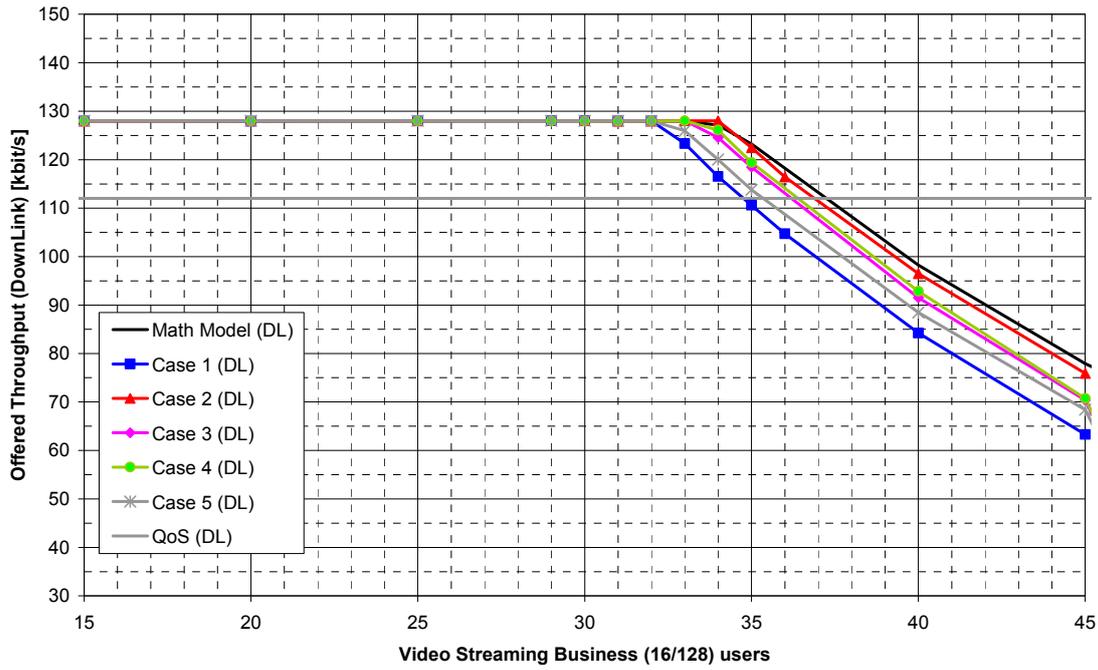


Figure 164 Offered throughput per user (DL) versus number of Video Streaming Business users

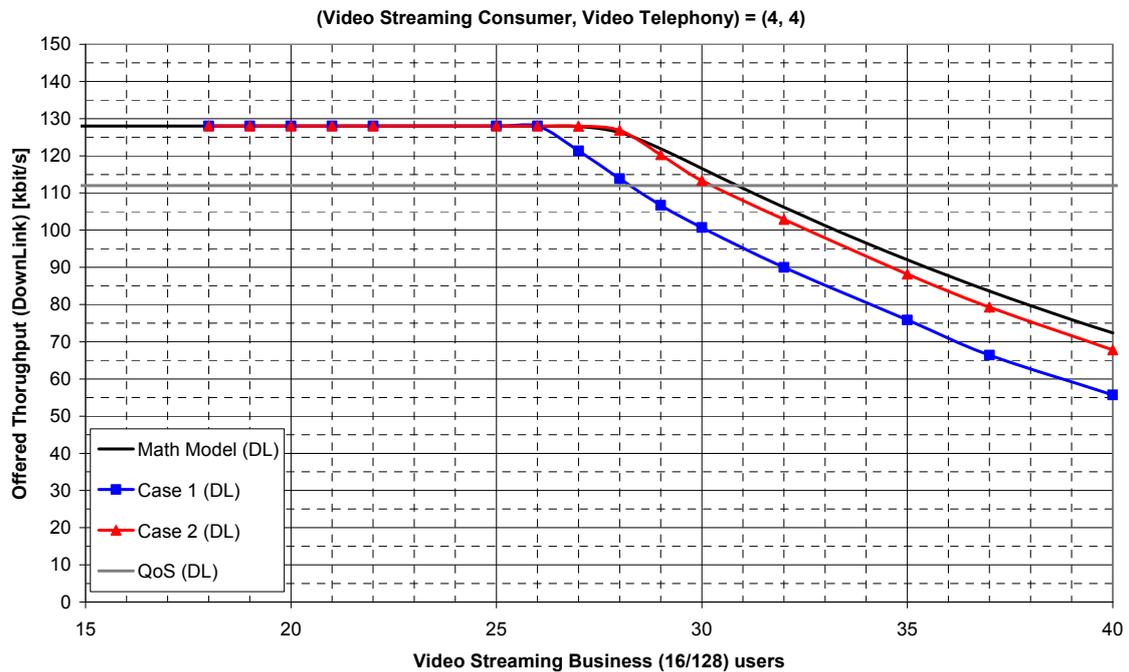


Figure 165 Offered throughput per user (DL) versus number of Video Streaming Business users with a fixed number of Video Streaming Consumer and Video Telephony users

4.2.4.5.2 Throughput Performance with RTS/CTS handshaking

The throughput performance obtained with RTS/CTS mechanism for Video Streaming Consumer and Video Streaming Business services is shown in Figure 166 and in Figure 167 respectively. As before, the curves in black are obtained from the analytical model, whereas from simulations those in blue refer to case 1 and those in red to case 2; also with the RTS/CTS mechanism there is a very good agreement between the results of the analytical model and those of case 2. With RTS/CTS handshaking, in order to guarantee the QoS constraint, with no other class of users, no more than 15-13 Video Streaming Consumer users (case 1-case2) or 8 Video Streaming Business users could be admitted.

Figure 168 shows the performance of the hot-spot with three classes of users using RTS/CTS procedure; in particular, the figure shows the throughput offered to each Video Streaming Business user with a fixed number of Video Streaming Consumer (4) and Video Telephony (4) users.

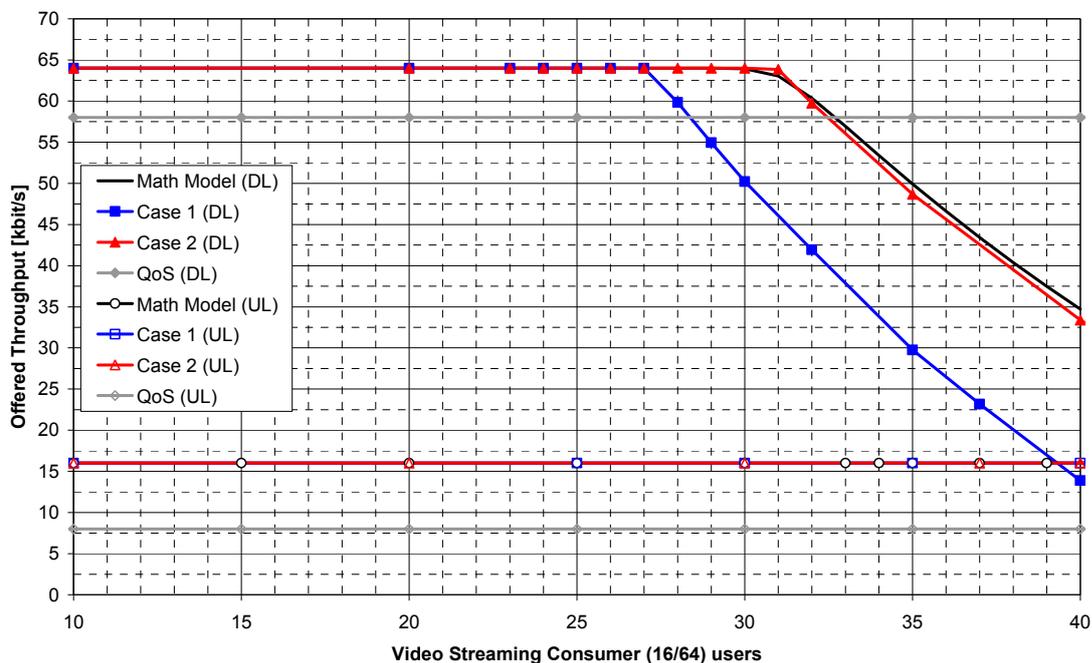


Figure 166 Offered throughput per user (UL/DL) versus number of Video Streaming Consumer users with RTS/CTS handshaking

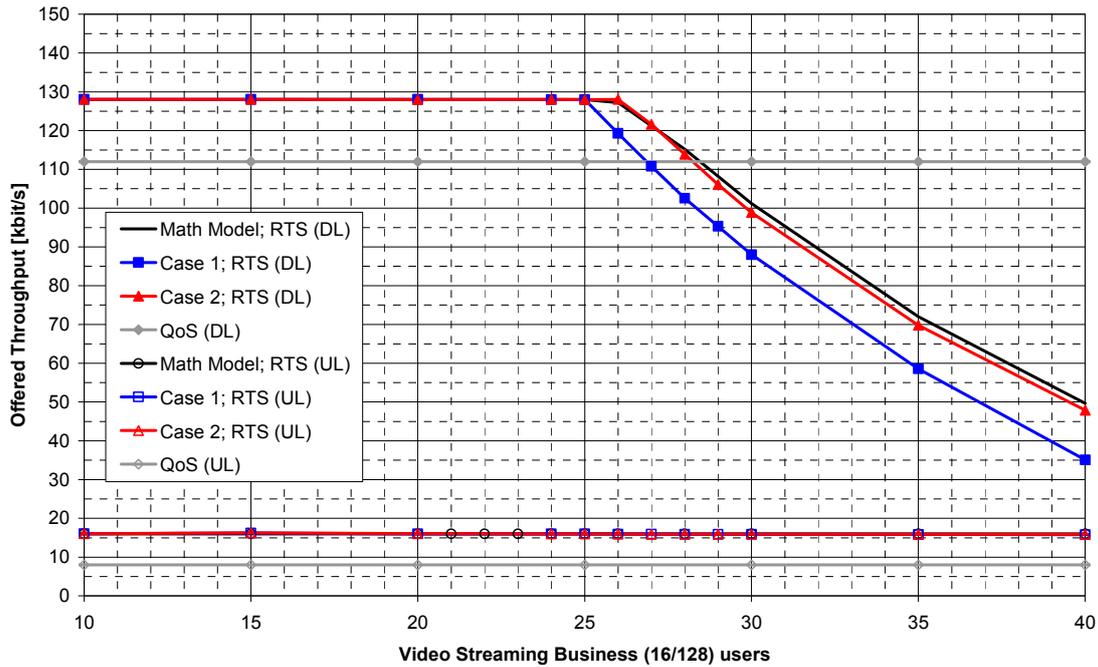


Figure 167 Offered throughput per user (UL/DL) versus number of Video Streaming Business users with RTS/CTS handshaking

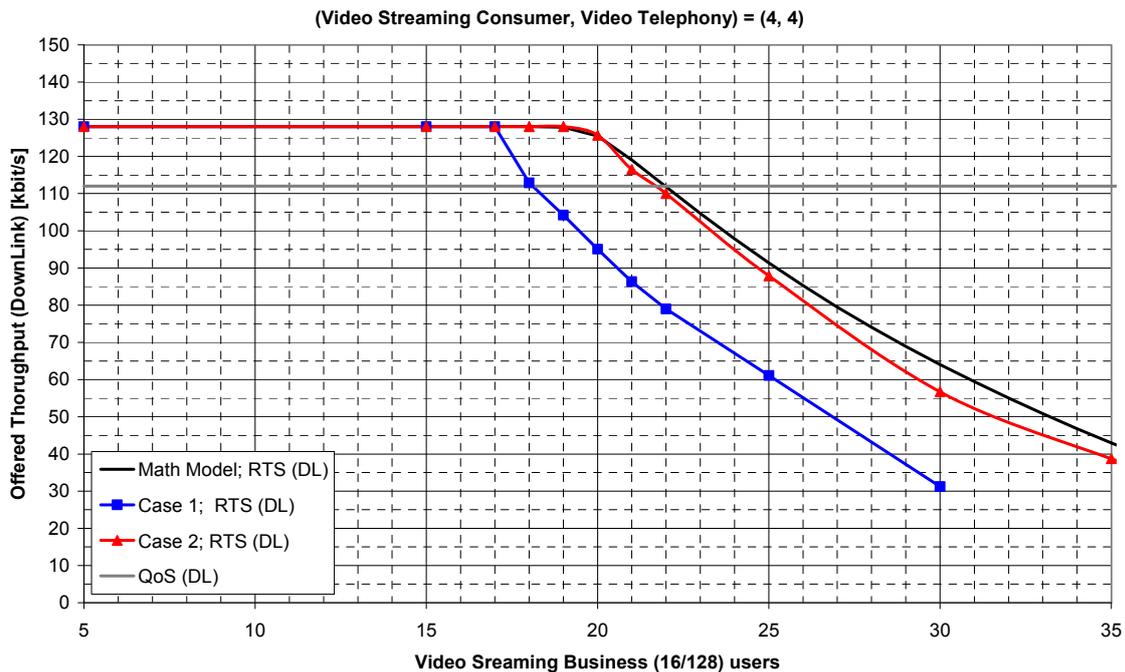


Figure 168 Offered throughput per user (DL) versus number of Video Streaming Business users with a fixed number of Video Streaming Consumer and Video Telephony users

4.2.4.5.3 Other performance parameters derived from simulations

In Figure 169 and in Figure 170 is shown the average delay of a frame obtained for case 1 and case 2 in uplink and downlink for Video Streaming Consumer and Video Streaming Business service respectively. As expected, the average delay for case 2 is lower than that for case 1, due to the reduced time interval of EIFS and the higher data-rate used for transmitting ACK frames. Moreover taking into account that for all services the packet length

in downlink is greater than the uplink (Table 40), the downlink average delay is larger than the uplink one.

The average delay increases with the number of users and in particular when the number of retransmissions becomes very high for each packet the average delay increases suddenly and becomes larger than 1 second.

This phenomenon is shown with an expanded scale in Figure 171 for the Video Streaming Business service.

The average delay for Video Streaming Business service in case of RTS/CTS mechanism is reported in Figure 172: the transmission of the two control frames (RTS and CTS) before each data frame causes an increase of the average delay.

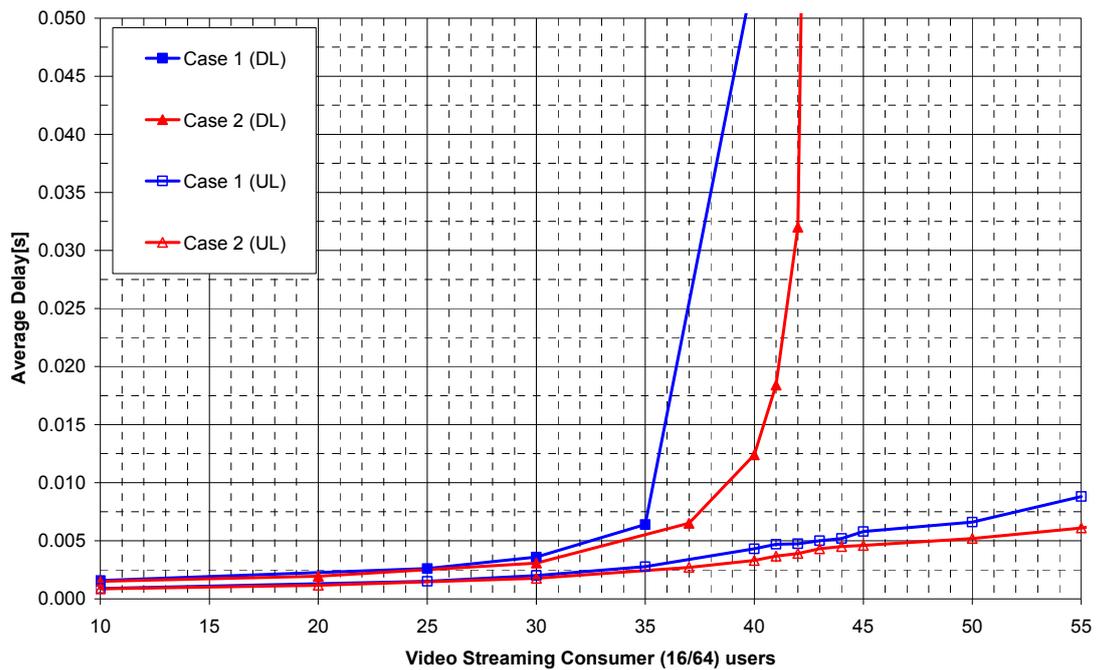


Figure 169 Average Delay (UL/DL) versus number of Video Streaming Consumer users

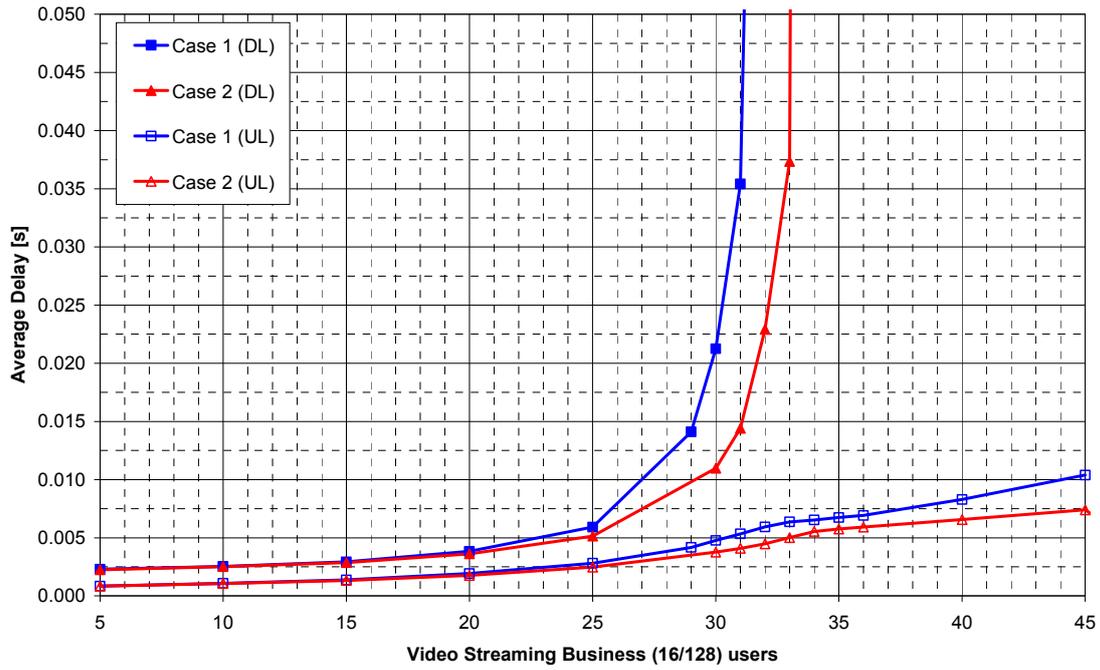


Figure 170 Average Delay (UL/DL) versus number of Video Streaming Business users

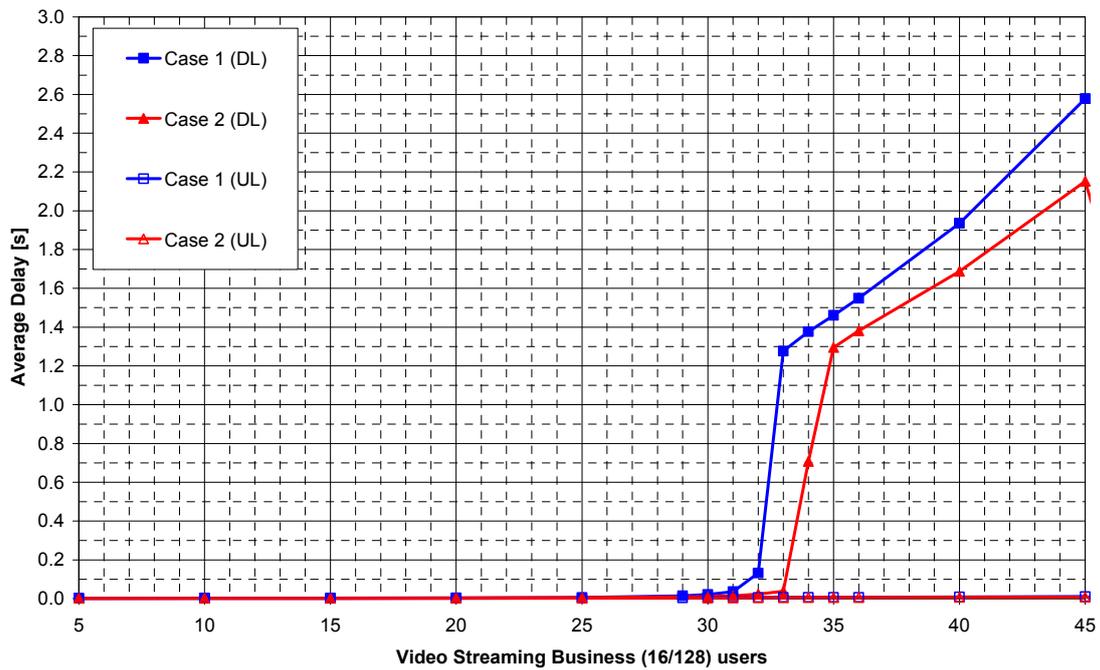


Figure 171 Average Delay (UL/DL) versus number of Video Streaming Business users

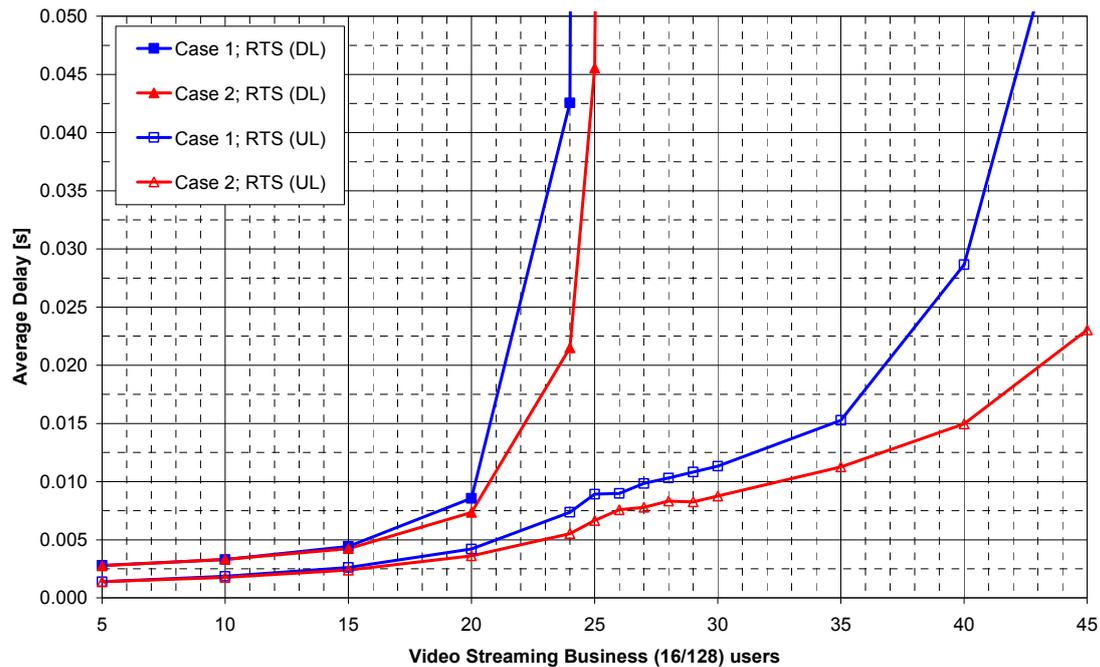


Figure 172 Average Delay (UL/DL) versus number of Video Streaming Business users with RTS/CTS handshaking

4.2.4.6 Conclusions

In this work the analytical model for WLANs IEEE 802.1b/a/g presented in [1] has been validated by means of simulations. In particular the behavior of the Distributed Coordination Function (DCF) was investigated with Basic Access and with the RTS/CTS handshaking and the effect of the different factors not considered in the analytical model, such as the AckTimeout, the EIFS interval, etc., was highlighted. Moreover, for the services envisaged by the EVEREST project the results have been used to derive an admission control policy for real-time services offered through a WLAN hot-spot. Within the context of the work, additional performance statistics as respect to the throughput offered by the hot-spot have been collected; due to the rapid growth of the average delay as respect to the number of users, the admission control policies applied by the network, have to take into account this parameter especially when applied to users that require time-bounded services.

4.3 ADMISSION CONTROL FOR IEEE 802.11E CONTENTION ACCESS

The incorporation of Quality of Service (QoS) issues to communication networks involves treating some traffic preferentially to others, what implies the ability to reject traffic.

The process that manages the rejection and admittance procedures within the network is referred to as admission control. Besides the simple rejection/admittance actions any admission control algorithm ensures that admittance of a new flow into a resource limited network will not degrade the QoS assurance of already admitted flows, while at the same time optimizes the network resource usage. Hence, admission control is an important component of QoS based resource management schemes.

Concerning admission control algorithms for wireless local area network (WLAN) IEEE 802.11e Enhanced Distributed Channel Access (EDCA) we can classify them into two main groups namely: measurement-based and model-based. Each of this group can be further split into centralized and distributed sets. The measurement-based algorithms takes the

decision about admittance/rejection of a new flow based on the continuous measurements of system parameters such as retry count or delay. In contrast, in model-based schemes the admission control decision is made according to the system status evaluated by means of some established metrics derived in analytical way. Graphical representation of admission control family for IEEE 802.11e EDCA with some corresponding example algorithms is shown in Figure 173.

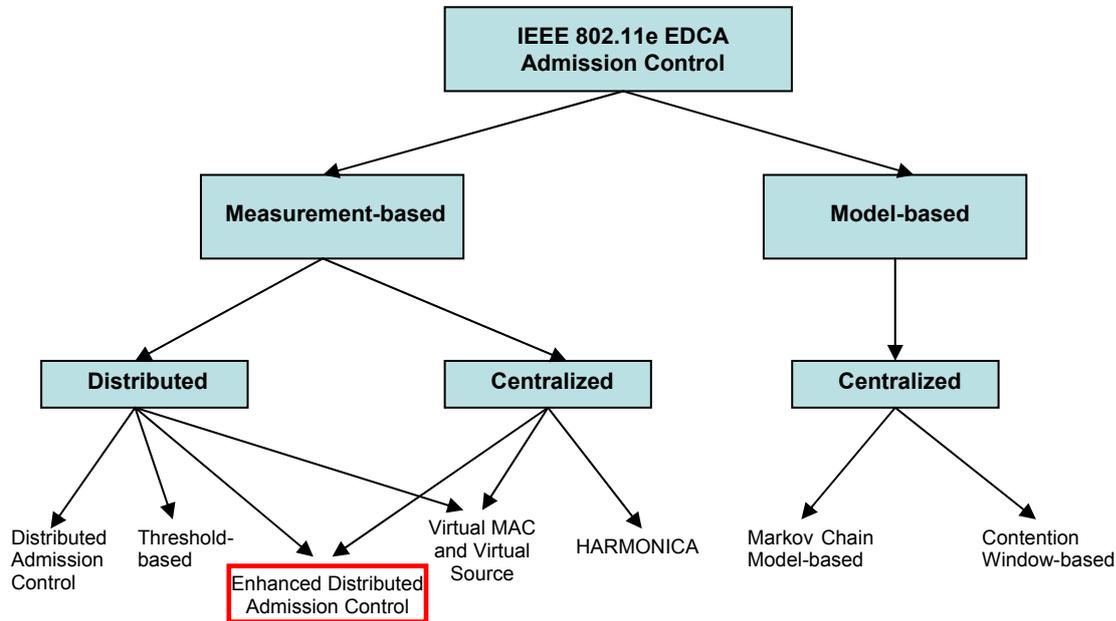


Figure 173 IEEE 802.11e admission control categories with examples

Before developing an enhanced admission control solution, first in this section the aforementioned mentioned schemes are studied and the summary of their main characteristics and limitations is presented in Table 43.

Table 43 Review of main characteristics and limitations of admission control algorithms for IEEE 802.11e EDCA

	Main features	Limitations	Special features
Distributed Admission Control [101][102]	Admission based on time estimation of time available, actually used for packet transmissions and wasted on collisions.	- uncertainty with resource estimation in heavy load conditions - no restrictions on burst traffic entrance	
Threshold based Admission Control [103]	Uses relative occupied bandwidth and average collision ratio to measure traffic condition in the system and depending on its state rejects/admits new flows.	- difficulty for setting the threshold values - possible loop effects when flows are stop and resume in parallel - instantaneous metrics are not guaranteed	
Virtual MAC and Virtual Source Admission Control [104]	Virtually emulates application and MAC processes to evaluate wireless link and verify whether flow can be admitted or not.	- needs some extra processing in each node - ignores the effect of incoming flow on existing flows	-does not consume any channel bandwidth
HARMONICA: Enhanced QoS	Relies on admission control policies and Link-	- difficulty for setting the optimal increment and	

Support with Admission Control for IEEE 802.11 Contention-based Access [105]	layer Quality Indicator (LQI) parameters, which include drop rate, link layer end-to-end delay and throughput	decrement of the channel access parameters - admitted flow may be rejected if stable state is not reached	
Markov Chain Model-based Admission Control [106]	Admission control is done according to the predicted achievable throughput for each flow calculated as explained in [97]	- model derived under saturation conditions - lack of consideration of multiple flows per station - lack of virtual collision model	
Contention Window-based Admission Control [107]	Adjusts contention window values to meet throughput requirements of each flow. If there exist a set of CW values that satisfy this condition new flow is admitted	- model derived under saturation conditions - lack of consideration of multiple flows per station - lack of virtual collision model - not optimum only based on CW adjustment	

Having analysed proposed schemes we decided to enhance the DAC model proposed by TGe of IEEE 802.11 as system traffic can be quite good estimated by its time metrics and this estimation can be easily done for each access category without nearly any modification to current draft. In the proposed Enhanced DAC (EDAC) algorithm we reused the idea of time estimations and enhanced it by introducing restrictions on burst traffic and by adding some centralize part to it, which will solve the case of parallel channel access of two or more stations⁸.

4.3.1 Enhanced Distributed Admission Control Algorithm

Similarly to the DAC mechanism, the proposed Enhanced Distributed Admission Control algorithm (EDAC), [108] is composed of two parts: one executed in AP, and the other in each station.

The station component of the algorithm is reduced to minimum and its responsibilities include:

1. To compute the total occupation time (TOT) per beacon interval;
2. For a given AC, to decide the acceptance or not of the new call depending on the TOT value and a transmission time budget (TTB).

The TOT parameter is determined according to following equation:

$$TOT = BeaconInterval - \sum_{AC=0}^3 TTB[AC] \quad (43)$$

If the value of TOT is lower than some threshold value (for instance 70%) the station directly admits its new flow. In case the TOT value is greater than the specified limit the station verifies whether its entering flow load is lower or equal to the TTB[AC]. The load should be calculated as an average of uplink and downlink load per beacon interval. The admission conditions for real and non-real time applications are summarised below:

For real time traffics:

$E(\text{Traffic Load UL\&DL}) \leq TTB[AC] \rightarrow \text{Accept}$
otherwise $\rightarrow \text{Reject}$

⁸ This is crucial point in a heavy load system conditions

For non-real time traffics:

TTB[AC] > 0 → Accept
otherwise → Reject

When the new traffic is of interactive or background nature, they can enter the system if the TTB of their AC is greater than zero. The reason is that this type of traffic could be delayed when needed (congestion situations), by means of contention parameters, without losing the QoS attributes.

From the point of view of admission control functionality, the AP role is to evaluate the TTB value for each AC according to the below expression and send it in a beacon frame.

$$TTB[AC] = \frac{tx_left[AC]}{SPF[AC]} \quad (44)$$

where,

tx_left[AC] - specifies the amount of time that has been left unused during the last beacon interval per access category.

SPF[AC] – it is a surplus factor representing the ratio of total time spend on all transmissions of a packet (with corresponding ACKs) to its actual length with employed channel speed.

Latest Drafts of IEEE 802.11e do not provide separate fields for sending available bandwidth information for each AC, only aggregated available bandwidth is sent. Therefore the QoS Basic Service Set (QBSS) load element's field "available admission capacity" should be replaced with four fields containing available admission capacity for each AC as shown in Figure 174.

Beacon Frame (Draft 4.4)

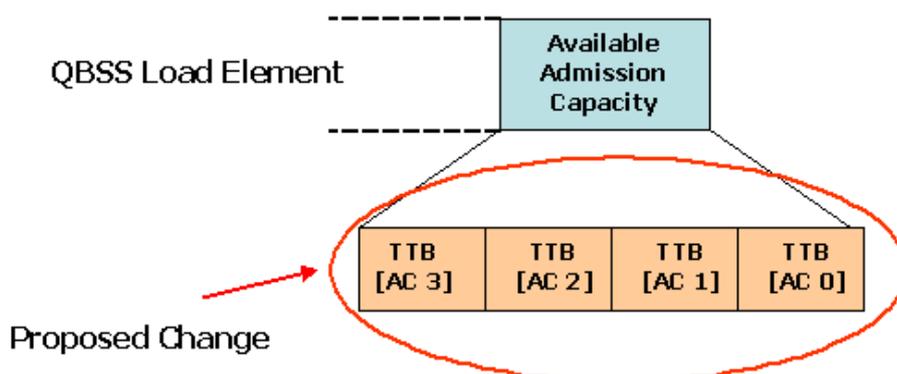


Figure 174 Proposed change of Available Admisión Capacity blocsk of QBSS Load Element

The SPF parameter role is to increase the computed value of the on-air bandwidth of transmitted frames by some additional factor, which takes into account the bandwidth wasted on collisions. The value of SPF may be constant or dynamic. The constant value may result in some inefficiency because too large values may provoke overestimation of used bandwidth and as a result a system efficiency decrement whereas too small values may allow entrance of too many stations. In the case of dynamic estimation of SPF a main problem concerns to the correct calculation of number of retransmissions per AC as well as to determine the AC classes that have suffered collisions. Moreover, the involved AC packets may experience following retransmission situations:

- retransmitted once or many times with finally positive outcome
- retransmitted in the next beacon interval
- retransmitted many times and finally discarded

In our experiments we chose the dynamic SPF estimation and we determine it as:

$$SPF[AC] = \frac{used_time[AC] + wasted_time[AC]}{used_time*[AC]} \quad (45)$$

where,

used_time[AC] – is a total used time by transmissions from specified AC observed by AP including all correct transmissions from and to AP and collisions experience by packets originated from AP;

wasted_time[AC] – is a time wasted on collisions from each AC;

used_time*[AC] – is a total used time by only correct packet transmissions from and to AP

The wasted_time[AC] parameter is 0 if no collision has occurred in the last beacon interval. In case of collision, first, collision_count parameter is incremented by 1 and next collided_pq[AC] flags⁹ are initialized (e.g. a value equal to -1). When the first retransmission packet is received by AP or if a packet from AP has taken part in the collision the collided_pq[AC] flag is set to enabled (e.g. a value equal to 1) and wasted_time[AC] is set equal to the duration of the packet. The collided_pq[AC] flag is used to not count twice the time wasted on collision if packets of the same AC has collided. Actually, if packets of the same AC has collided, only the largest packet duration is considered to estimate wasted_time[AC] parameter. When the second retransmission packet is obtained by AP with different AC then the above procedure repeats. The whole process repeats if a second collision takes place and at least two retransmission packets have been received after first collision. However, if after a collision only one packet or no packet is received and next collision occurs, then we move to the so called PRIMITIVE case as the previous calculations are not valid any more. In the PRIMITIVE case a waste_time[AC] parameter for given AC is incremented by duration of each received retransmitted packet independently of whether other packets of the same AC has collided or not.

As performed waste_time[AC] calculation is done on per beacon interval bases therefore after reception of a beacon frame we come back to the initial state.

The flow diagram of waste time calculation is shown in Figure 175.

Moreover, we introduce one more restriction on packets which retransmission takes pace in the next beacon interval. If such a retransmission occurs before the first collision happens the received packet is no treated as retransmitted packet, hence does not increase the wasted_time[AC] parameter value.

⁹ Indicating whether retransmission packet from each AC were received or not after a collision

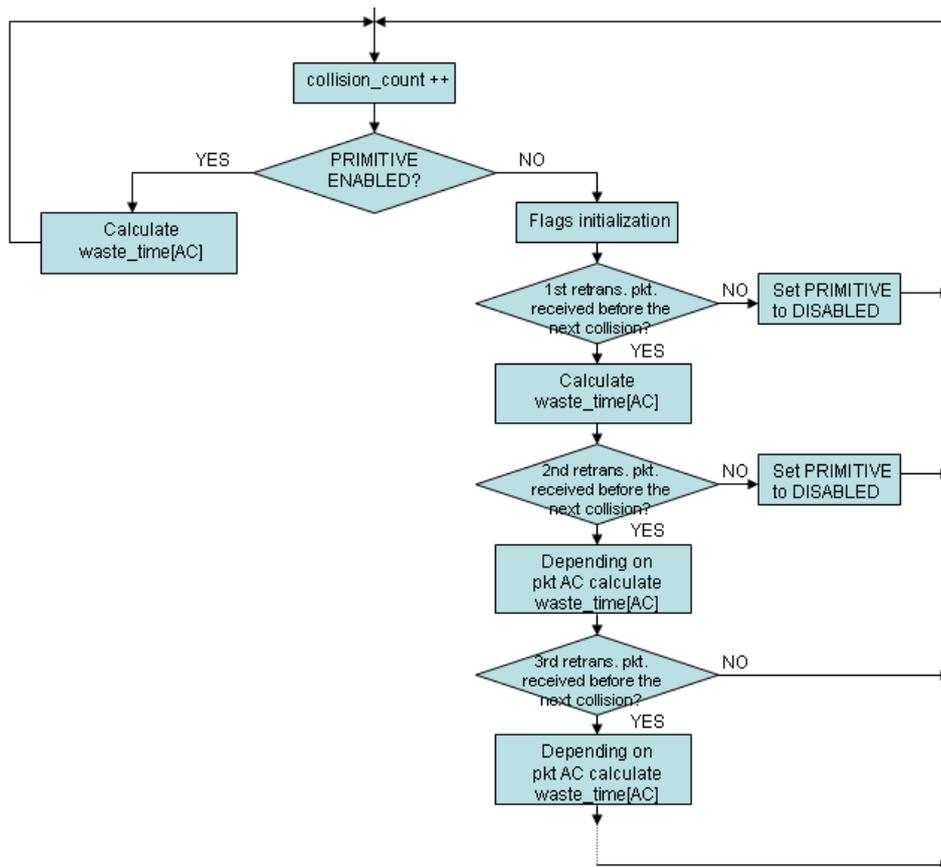


Figure 175 Wasted time calculation diagram

The expression for calculating the $tx_left[AC]$ time is shown below, where the AP node uses two new parameters: Transmission Time Threshold ($TTT[AC]$) and Load[AC].

$$tx_left[AC] = MAX(TTT[AC] - Load[AC], 0) \quad (46)$$

The $TTT[AC]$ is a crucial parameter of EDAC algorithm, which represents the maximum amount of time that may be spend on transmissions of a specific AC per beacon interval. Its value may be constant or may change dynamically. The constant value of $TTT[AC]$ is an optimum solution only for a given service mix but, when service mix distribution changes with the time the constant threshold is ineffective. The dynamic tuning of $TTT[AC]$, which may depend on parameters like: current load per AC, contention window size, number of stations, type of application, etc., is an adequate solution for a system with varying number of users. In performed studies a dynamic adjustment of TTT is implemented being proportional to the used service mix distribution. The detail description of realized TTT tuning is given in performance evaluation paragraph.

The Load[AC] attribute corresponds to the total time occupied by transmissions from each AC per beacon interval. Its value is computed differently for real time and burst type traffics. For real time traffic the below formula is applied:

$$Load[AC] = MAX(tx_time[AC] * SPF[AC], beacon_load[AC]) \quad (47)$$

where

$tx_time[AC]$ – is a set of counters that indicate the total time on-air of the frames during beacon interval;

$beacon_load[AC]$ – corresponds to the average load that could be introduced during a beacon interval by all admitted streams within each AC;

In the case of non-real time applications, as the load is not known a priori, due to its bursty nature, the AP establishes a minimum load average, known as average guaranteed rate (GuaranteedRate). By means of the GuaranteedRate and the transmission time threshold, the AP is able to estimate the load of active burst flows and control their number. Hence, the Load[AC] parameter is calculated as follows:

$$Load[AC] = GuaranteedRate * admitted_strm_num[AC] \leq TTT[AC] \quad (48)$$

where

admitted_strm_num[AC] – refers to the number of all admitted streams within each AC.

Besides the distributed part of EDAC mechanism, the AP also performs some centralized decisions, taking care of the case when two or more flows try to enter the system in the same beacon interval and there is no sufficient time for placing all of them. According with aforementioned explanation all the flows will satisfy the condition and will be admitted. Therefore to avoid this problem, the station only admits its new traffic if an ACK is received after sending the first packet (trial packet). However, the lack of the ACK is considered by a station as a packet loss thus, it will try to retransmit it. Then, to limit the number of retransmitted trial packets each station rejects its new traffic after three missing acknowledgement, obviously assuming that a new beacon frame with updated parameters is not received earlier.

To manage this new centralized situation, the AP needs a continuous control of TTB[AC] time. In addition, to know at each moment the minimum occupation time, the implementation of two new tables (the number of admitted stations and the number of stations accepted in the current beacon interval) is also needed in AP.

4.3.2 EDAC performance evaluation

To assess and validate the effectiveness of EDAC algorithm a single QBSS cell was assumed with an increasing number of mobile stations. Moreover, the following service mix distribution was assumed: voice 50%, video 16% and web 34%. The voice traffic is generated following the specification G.729 A/B for VoIP application, with transmission rate of 24kbps [109]. To model the video stream a Group of Pictures (GOP) of 12 was used with 25 frames per second and 128 kbps transmission rate in downlink and CBR of 16 kbps in uplink [1]. The traffic model for web traffic considers the generation of activity periods (i.e. pages for www browsing), where several information packets are generated and a certain thinking time between them exists, reflecting service interactivity. The specific parameters are: time between pages: avg. 4sec. UL/5.17sec. DL; average number of packets arrival per page 25 (UL/DL); number of bytes per packet: 1000 bytes maximum 60000 (truncated Pareto distribution); time between packets arrival: avg. 0.03125 UL/0.015625 DL exponentially distributed [24]. Table 44 summarizes the main traffic generation parameters.

Table 44 Traffic generation information

Voice	CBR	24 kbps
Video	GOP 12	25 fps avg. 128/16 kbps
Web	Activity periods (web browsing model)	512/256 kbps

Moreover, it is assumed that each station operates with IEEE 802.11b physical layer with channel rate of 11Mbps. The EDCA contention parameters used for each AC are presented in Table 45.

Table 45 EDCA contention parameters

AC	AIFSN	CWmin	CWmax
1	3	31	1023
2	2	15	31
3	2	7	15

The beacon interval is equal to 100 ms and, at the beginning, TTT[AC] time is the same for each traffic and equal to 30 ms. The remaining time (10 ms) is used to absorb the traffic fluctuations. The TTT[AC] value is adjusted dynamically being proportional to the service mix distribution. Moreover, to limit the number of collisions due to the small size of CW, for voice traffic the maximum TTT[AC3] is determined as a function of CW size and station number resulting equal to 54 ms.

Firstly aggregated throughput for voice traffics with and without admission control is analysed and the results are shown in Figure 176. We observe that without the EDAC algorithm the throughput starts to oscillate when system load is very heavy (18 stations). For analysed scenario, when EDAC scheme is used the maximum number of voice station that can be admitted is 8¹⁰. The proposed algorithm stops further admission of voice stations and prevents the system from entering into the saturation state. Certainly, without admission control all stations are allowed to enter the system. Therefore when the number of voice stations is higher than 8, the system becomes saturated (overloaded) and the collisions between packets increase. As a result, the aggregated throughput for voice traffic fluctuates, but, in average, it does not increase beyond the throughput corresponding at 8 stations.. In consequence, in the saturation case, without EDAC algorithm the QoS requirements for voice traffic cannot be guaranteed. On the other hand, with admission control mechanism, when the system reaches the heavy load state no more stations are admitted and already active stations are protected.

¹⁰ For the assumed service mix, voice is the 50% of the total. From figure 3 a maximum of 16 mobile stations could be admitted. Then the maximum number of admitted voice services is 8.

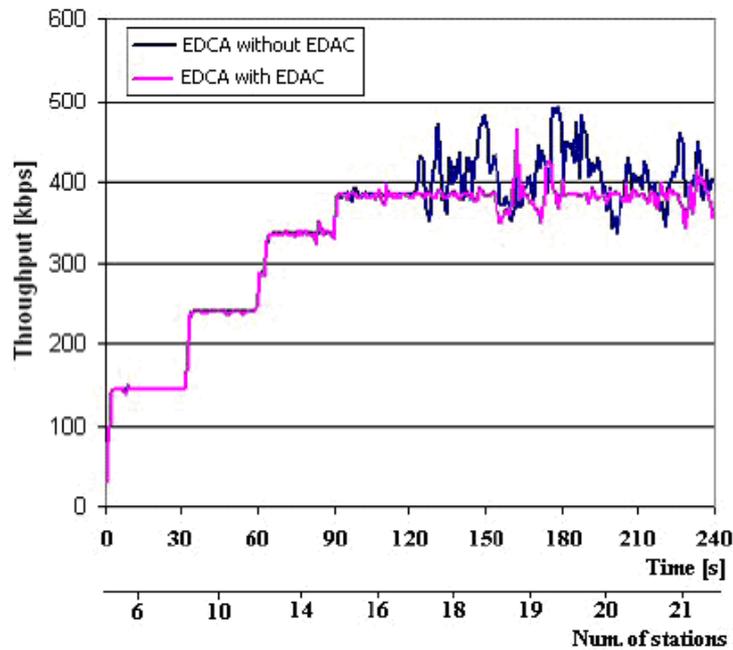


Figure 176 Aggregated throughput for voice traffic with and without EDAC algorithm

Next, Figure 177 demonstrates the cumulative distribution function of MAC delay for video traffic with and without admission control. Comparing these two plots we may clearly see that without admission control the MAC delay of 95% of packets is lower than 4s, whereas in case of EDAC mechanism MAC delay for the same case is lower than 0.11 s. Accordingly, uncontrolled admission of video stations in the set-up scenario without EDAC algorithm provoke a significant increase of experienced delay of these stations and, in consequence, the loss of their QoS expectative.

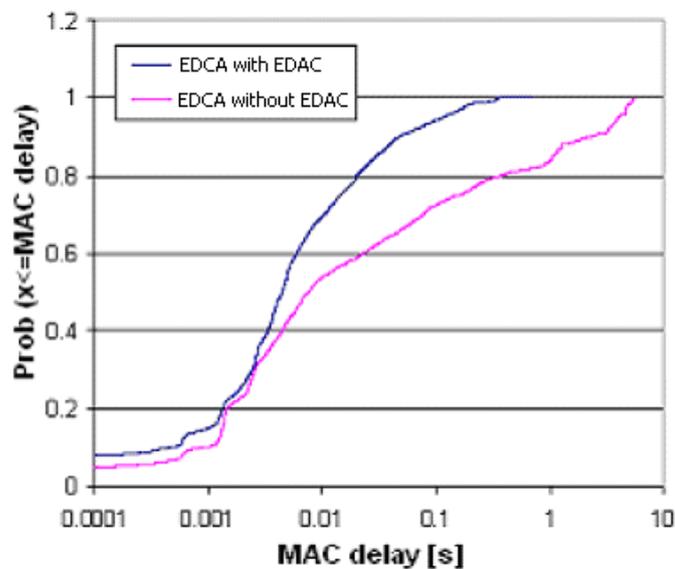


Figure 177 CDF for video MAC delay with and without EDAC algorithm

Now, the dynamic tuning of TTT[AC] value for analysed set-up is studied. At the initial state value of this parameter is the same for each access category. However, when a new station enter the system, the AP needs to recalculate values of TTT for each AC to make them proportional to current bandwidth requirements of each traffic class. Therefore, when the transmission time limit of AC “i” is superior to the initial TTT[i] value and the total occupation

time (TOT) is lower than some threshold (for instance 80% was considered in performed simulations) the AP first assigns required transmission time to the class "i", and in the next step divides the remaining free time between other traffic categories proportionally to the provided service mix distribution as shown in Figure 178. This tuning of TTT[AC] values is repeated up to the moment when the system reaches heavy load state, for instance 80% of total possible bandwidth.

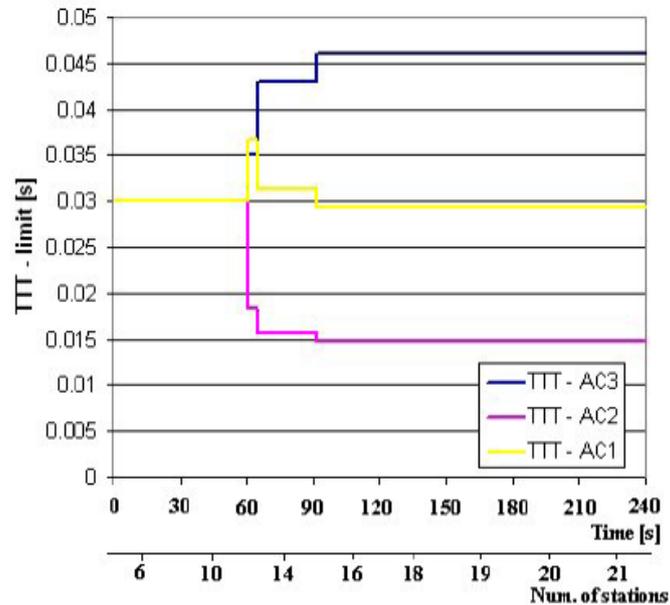


Figure 178 TTT[AC] time tuning with increasing number of stations

The total occupation time for each AC seen by AP is presented in Figure 179. We can clearly see that this parameter experiences great oscillations. This behaviour is conditioned by elevated number of collisions resulted from the high system load and, also, by the uncertainty problem in exact estimation of wasted bandwidth on collisions, which is reflected by means of the SPF parameter. In consequence, such a big oscillation may provoke problems in the admission decision algorithm. To limit this negative effect, the AP at each beacon interval knows besides the peak total occupation time the minimum average time, the `beacon_load[AC]`, of the transmissions from each AC. Implementing this minimum average time, a lower band limit on the total occupation time from each AC is established. This value does not change during a beacon interval¹¹ and it is only updated when a new station is admitted to the system.

¹¹ Notice that the AP compares the current value of the used bandwidth with this minimum averaged load, selecting the highest.

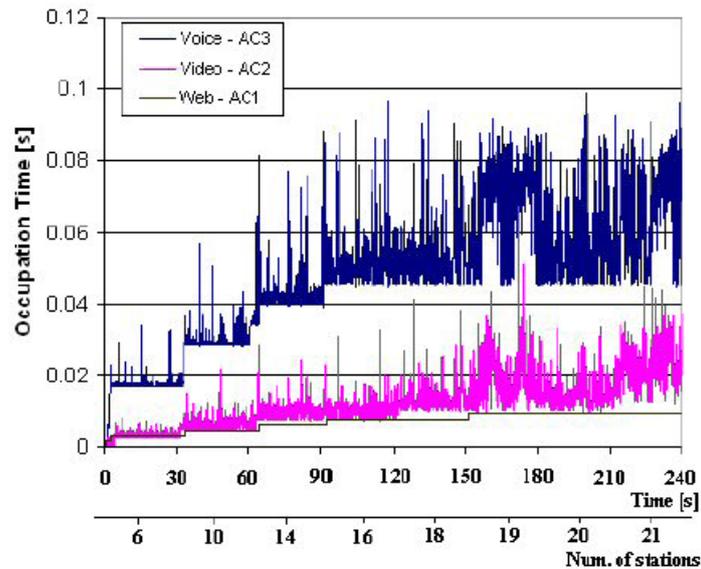


Figure 179 Total occupation time for each AC

4.3.3 Conclusions

The proposed Enhanced Distributed Admission Control algorithm is able to protect already active flows of continuous nature (conversation and streaming) and allows controlling low priority bursty traffic by means of minimum average guaranteed load. It provides a dynamic control of time spend on transmissions from each access category managed by Transmission Time Threshold parameter and controlled by AP. Moreover, the proposed algorithm only introduces slight changes in the current draft of IEEE 802.11e standard. In that sense its applicability is guaranteed.

4.4 SERVICE PRIORITY – QoS ENHANCEMENTS IN 802.11b

Despite the important role of wireless LANs based in particular on 802.11b, the lack of quality of service (QoS), security, etc. are important factors to consider. Nevertheless, the Task Group E within the IEEE 802.11 Working Group is developing a project (802.11e) whose purpose is to enhance the current 802.11 MAC to expand support for applications with Quality of Service (QoS) requirements [101]. In the meantime, and because of the plausible concerns about the time needed to stabilize and agree the standard in a first step and the time to market for 802.11e products in a second step, there have been proposals to provide QoS in 802.11b at MAC level, such as Distributed Fair Scheduling [110], Blackburst [111], Aad's Differentiation Scheme [112], etc. These proposals modify parameters that define how a station (STA) access the wireless medium, i.e. they modify parameters of either the fundamental access method of the IEEE 802.11 MAC called Distributed Coordination Function (DCF), or the optional access method, which is the Point Coordination Function (PCF) [113]. In fact, some of these proposals are being included in the future Hybrid Coordination Function (HCF) in 802.11e [101].

4.4.1 Hierarchical Token Bucket

The purpose of this section is to introduce new mechanisms allowing to provide appropriate quality of service in WLAN with the novelty that, instead of focusing on MAC layer, the proposed solution is set at IP level with Hierarchical Token Bucket (HTB), which exercises control over the transmissions, queuing and dequeuing packets in a determined and configurable way. HTB is a very complete and useful traffic shaper that has been successfully tested on wired environments [114] and [115]. This section extends the use of

this algorithm in a WLAN environment, and the proposed solution at IP layer could be incorporated into MAC layer as an enhancement to IEEE 802.11e.

One of the characteristics of current IEEE 802.11 products is the link adaptation, which consists on downgrade the bit rate transmission to a lower value when repeated unsuccessful frames transmissions are detected. This behaviour is shown to be very efficient for a standalone host. Nevertheless, link adaptation may seriously degrade the WLAN global performance, which penalizes fast hosts and privileges slow stations [116]. The HTB mechanism proposed can be a solution to the mentioned problem of stations transmitting at different rates. Furthermore, using Hierarchical Token Bucket less aggressive medium access behaviour can be achieved, with the corresponding positive influence on the throughput standard deviation.

HTB is based on hierarchical classes where three class types exist: root, inner and leaf. Root classes are suited on the top of the hierarchy and all traffic goes out through them. Inner classes have father and daughter classes. Finally, leaf classes are terminal classes, so they have father classes but not daughter classes. In leaf classes, traffic from upper layers is injected following a classification which must be performed using filters, so it is possible to difference kinds of traffic and priorities, which should have different treatment. In this way, before traffic enters in a leaf class, it needs to be classified through filters with different rules, which can filter by kinds of services, IP addresses or even network addresses. This process is known as classifying process. Furthermore, when traffic has been classified, it is scheduled and shaped. In order to perform these tasks, HTB uses the concept of tokens and buckets to control the bandwidth use in a link. To adjust the throughput, HTB generates tokens at necessary cadence and de-queues packets from the bucket only if tokens are available.

As an example, different bit rate transmissions are considered. STA1 is 11 Mbps bit rate and STA2 is 2Mbps. When STA2 transmits UDP traffic at 2 Mbps, it uses five longer time than STA1. For this reason, if HTB constricts STA2, available bandwidth to share is almost four times larger the bandwidth HTB has constricted. For example, when HTB limit of STA2 is reduced from 1300 Kbps to 900 Kbps, STA1 HTB limit is upgraded from 900 Kbps to 2400Kbps, so a 50 % gain on available throughput follows. Additionally, the standard deviation reduces significantly (Table 46). In TCP traffic, gain is almost 53 %. Applying this concept, a plain and sustained throughput with low standard deviation to stations or even to differentiated services can be achieved.

Table 46 Measured Throughput and Standard Deviation for 11 and 2 Mbps rates and UDP Traffic.

HTB Status	STA rate (Mbps)		HTB Limit (Kbps)		Average Throughput (Kbps)		Std Deviation σ (Kbps)	
	STA1	STA2	STA1	STA2	STA1	STA2	STA1	STA2
Without HTB	11	2	-	-	965	1303	164	62
With HTB	11	2	900	1300	900	1301	42	27
	11	2	2400	900	2404	901	56	11

4.4.2 Comparison of the legacy DCF and DFS, DRRR service differentiation schemes

Considered fair scheduling based SDS mechanisms namely, Distributed Fair Scheduling [110] and Distributed Deficit Round Robin [117] are advantageous over the legacy DCF

access mechanism as attempt to fairly allocate bandwidth among traffic classes for that reason a comparative analysis between them was performed.

Simulation environment is composed of one AP and eight stations and is identical for each service differentiation model. In proposed set-up we employ neither hidden terminal nor beacon frame. The channel model used is error free. Packet generation is the constant bit rate (CBR) with packet size of 1000 Bytes and packet interarrival time of 0.08 seconds. Data packets are transmitted at 1Mbps.

To not make a simulation process too complex, due to fine parameters adjustment, performed evaluation investigate the ability of fair bandwidth distribution of each scheme among flows with equal priority.

In the simulation we considered following DRR parameters. As each flow requires 100kbps a quantum rate is equal to 110 kbps (10kbps for header). The scaling factor is $7.5 \cdot 10^{-6}$. In DFS mode weight of each flow is equal to 0.125 and scaling factor of 0.01.

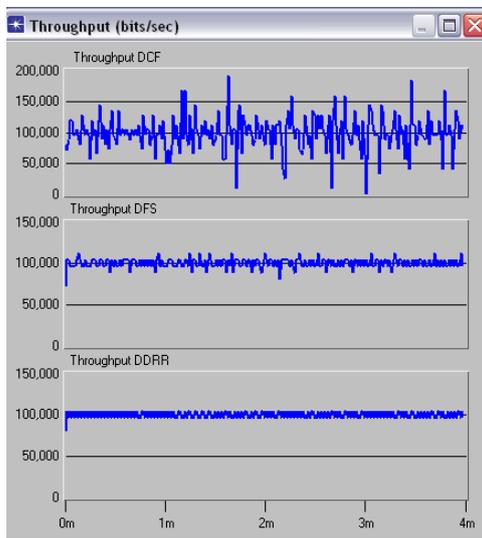


Figure 180 Throughput for DCF, DFS and DRR schemes

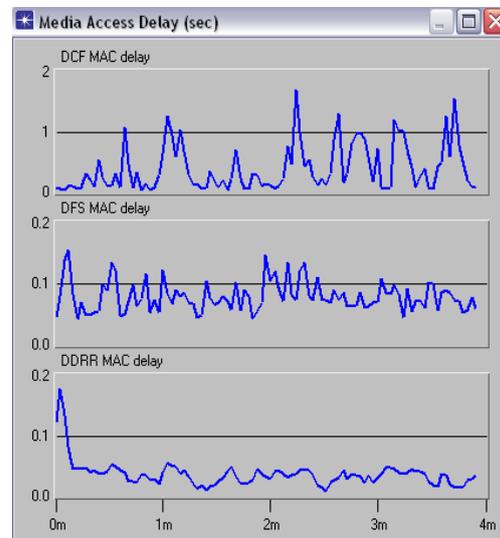


Figure 181 Media Access Delay for DCF, DFS, DRR schemes

From above figures we see that DRR and DFS mechanisms show better performance than legacy DCF access model. Concerning throughput variability DFS presents great reduction while in case of DRR it is negligible, as it does not use backoff algorithm.

In terms of MAC delay fair queuing schemes also demonstrate high improvement. Delay variability is decreased considerably in DRR and DFS.

Standard deviations and average values of throughput and MAC delay for each scheme are shown in Table 47 and Table 48 respectively.

Table 47 Standard deviation and average value of throughput for different schemes

	Average value (bits/sec)	Standard deviation (bits/sec)
DCF	98866.666	28261.556
DFS	99900.000	5932.116
DRRR	99933.333	4194.705

Table 48 Standard deviation and average value of MAC delay for different schemes

	Average value (sec)	Standard deviation (sec)
DCF	0.394282	0.389428
DFS	0.079129	0.024932
DDRR	0.036394	0.022943

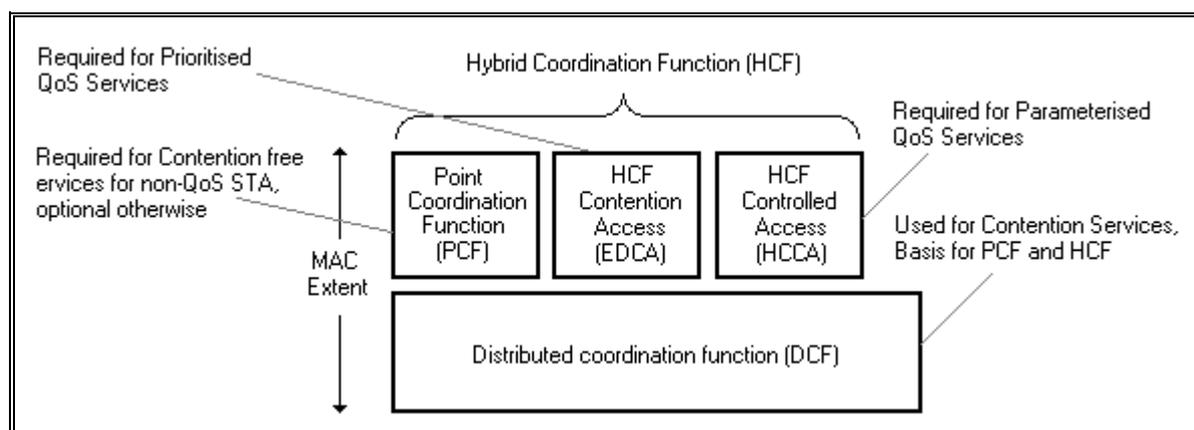
Performed analysis shows superiority of the fair scheduling schemes over the legacy DCF mode, however we should take into account that the study was done without priority differentiation.

4.5 SERVICE PRIORITY - QoS ENHANCEMENTS IN 802.11e

Despite the important role of wireless LANs for best effort traffic based in particular on 802.11b and g physical layers, limitations of the standard are apparent for mixed traffic scenarios because of the lack of quality of service (QoS) support. Nevertheless, the 802.11e proposals modify parameters that define how a station (STA) access the wireless medium, i.e. they modify parameters of either the fundamental access method of the IEEE 802.11 MAC called Distributed Coordination Function (DCF), or the optional access method, which is the Point Coordination Function (PCF) [113]. In fact, some of these proposals are being included in the future Hybrid Coordination Function (HCF) in 802.11e [101].

The Task Group e (TGe) within the IEEE 802.11 Working Group is developing a specification (802.11e) whose purpose is to enhance the current 802.11 MAC to expand support for applications with Quality of Service (QoS) requirements [101]. In the meantime, and because of the plausible concerns about the time needed to stabilize and agree the standard in a first step and the time to market for 802.11e products in a second step, there have been proposals to provide QoS in 802.11b at MAC level, such as Distributed fair Scheduling [110], Blackburst [111], Aad's Differentiation Scheme [112], etc.

The enhanced MAC introduces an additional coordination function, HCF, which consists of two new access methods, the enhanced distributed channel access (EDCA) and the HCF controlled channel access (HCCA). EDCA corresponds to the legacy DCF, while HCCA corresponds to the legacy PCF. The MAC architecture of 802.11e is illustrated in Figure 182.

**Figure 182 MAC architecture in 802.11e**

Simulations have been made to evaluate 802.11e in terms of jitter, MAC delay, throughput and packet loss. The number of users in the network at a specific time has not been taken into account. Therefore, only applications have been used as a base for the traffic model. Each station represents one or more applications, which start with uniform distribution during

the initial stage and continue throughout the simulation. This renders the possibility to control the traffic load during the simulation and traffic, belonging to a certain access category, may easily be added or removed.

The simulation model developed for evaluation of 802.11e is based on OPNET Modeler 10.5 with the wireless module. The original WLAN MAC model, *wlan_mac*, shipped with OPNET Modeler has support for 802.11a/b/g standards. The additional 802.11e implementation was based on the existing WLAN MAC model.

The terminals, here stations, implementing 802.11e functionality are referred to as a QSTA. In order to be compatible with legacy stations, DCF must be implemented in an 802.11e network. PCF is support, but optional and will most likely not ever be used together with 802.11e since it is only rarely used in legacy 802.11. EDCA supports prioritised QoS, while HCCA supports parameterised QoS.

4.5.1 Enhanced distributed channel access (EDCA)

The enhanced distributed channel access, EDCA, is based on legacy DCF and supports prioritised quality of service. A problem with legacy DCF is that the MAC uses a single FIFO (First-In-First-Out) transmission queue. With this approach there is no way to differentiate traffic and provide priority to particular traffic types. EDCA implements four different access categories (ACs), voice (VO), video (VI), best effort (BE) and background (BK). Every AC implements a unique queue and to each AC, there is a specific EDCA parameter set, which defines the unique properties of the access category. Every AC can be seen as a virtual station, which competes individually for channel access. Traffic can be divided further into user priorities (UP). User priorities are simply a priority scheme within an AC. There are eight different user priorities, derived from 802.1D [118][117], which are shown in Table 49. When frames arrive from the higher layer, the MAC layer examines the TOS field in the IP header. The TOS value corresponds to the UP and the incoming frame is mapped upon an access category and queued in the appropriate queue, as illustrated in Figure 183

Table 49 User priorities and access categories used in 802.11e.

Priority	User (same as UP)	priority as 802.1D	Access category	Designation
Lowest	0		AC_BE	Best Effort
	1		AC_BK	Background
	2		AC_BK	Background
	3		AC_BE	Best Effort
	4		AC_VI	Video
	5		AC_VI	Video
Highest	6		AC_VO	Voice
	7		AC_VO	Voice

An additional timing interval is introduced in 802.11e in order to differentiate the four access categories. This timing interval is referred to as Arbitration inter frame space (AIFS) and is the time an AC must sense the medium being idle before seizing it. Each AC has specific AIFS time and contention window boundaries, see Table 50. The min and max values in Table 50 are values used in legacy DCF and it is the relation between these values, in the different ACs, that are of interest.

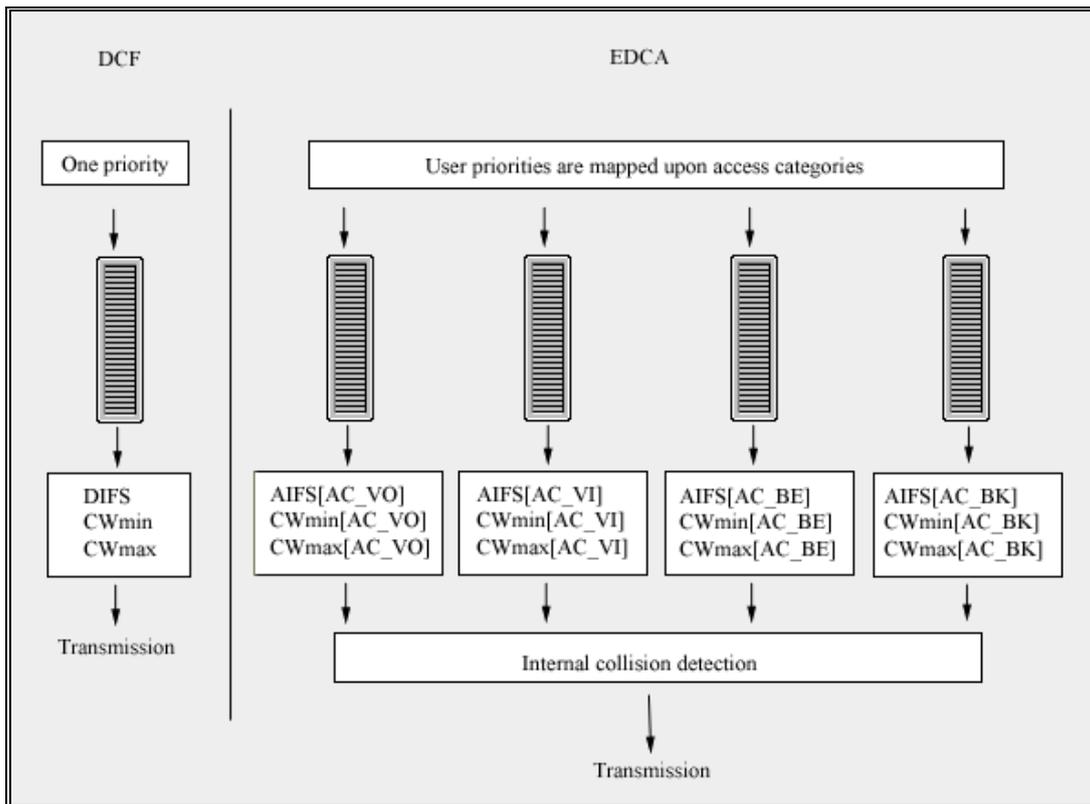


Figure 183 EDCA implements four unique access categories based on eight user priorities.

Table 50 Inter frame space timers and relation in contention window size for EDCA and legacy DCF.

AC	AIFS / DIFS	CW min	CW max
Voice	$2 \cdot aSlotTime + SIFS$	min/4-1	min/2-1
Video	$2 \cdot aSlotTime + SIFS$	min/2-1	min
DCF	$2 \cdot aSlotTime + SIFS$	Min	max
Best Effort	$3 \cdot aSlotTime + SIFS$	Min	max
Background	$7 \cdot aSlotTime + SIFS$	Min	max

In order to investigate the QoS enhancements specified in 802.11e a number of simulation scenarios with the developed WLAN MAC model performed. The prioritised QoS scheme, EDCA, was evaluated by studying throughput, jitter and delay in different network configurations.

In the first set of simulations, the traffic model is adapted to emphasize and evaluate some of the properties of the EDCA access mechanism. The packet size is set to 1500 bytes, which roughly corresponds to the maximum size of an Ethernet frame, which will be passed down to the MAC by the higher layer. Note that traffic mapped as voice may be a videoconference application, which, regarding QoS parameters, requires the same properties as a voice channel.

The parameters packet inter arrival time, distribution and traffic load are set equal for all access categories in order not to affect the sought results. See Table 51 for the traffic configuration.

Table 51 Traffic parameters.

Voice	Video	Best effort	Background
-------	-------	-------------	------------

Packet size (bytes)	1500	1500	1500	1500
Packet inter arrival time (ms)	20	20	20	20
Arrival distribution	Constant	Constant	Constant	Constant
Percentage of total traffic load	25%	25%	25%	25%

4.5.2 Prioritisation in EDCA

To differentiate access categories in the 802.11e WLAN MAC, additional inter frame space times and AC dependent contention windows were introduced.

The first simulation set up was designed to evaluate the general functionality of the EDCA prioritisation scheme. The result of a real life 802.11e implementation is highly dependent on the traffic characteristics used in the network, therefore, traffic parameters are configured as in Table 51. Each station in the simulation is running one application from each access category. In the following scenario, simulations were done with five, six and seven stations in the network, in order to simulate the network during heavy load and after saturation.

4.5.3 Simulation Results

Figure 184 shows the obtained throughput with five, six and seven stations. The results showed that the highest throughput with acceptable MAC delay and packet loss, for all access categories, was achieved with five stations and that the network is saturated a bit below 25 Mbps. When adding even more traffic to the network, the total throughput decreases. This is due to the fact that the network is saturated and lower prioritised access categories are dropping packets. With higher resolution when adding traffic it is possible to fine tune the maximum throughput and simulations have shown that a throughput close to 25 Mbps can be achieved.

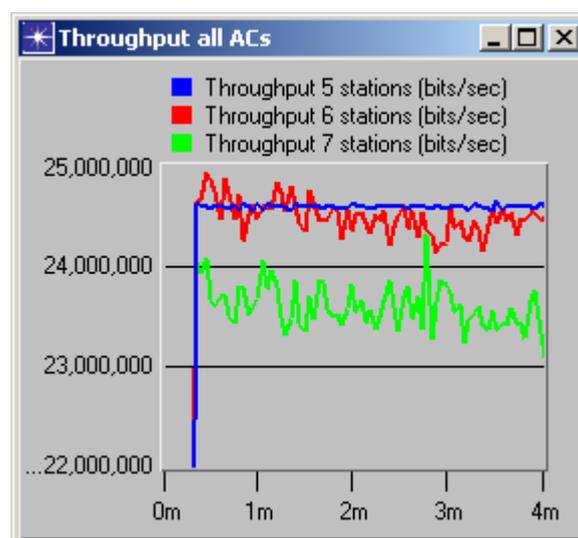
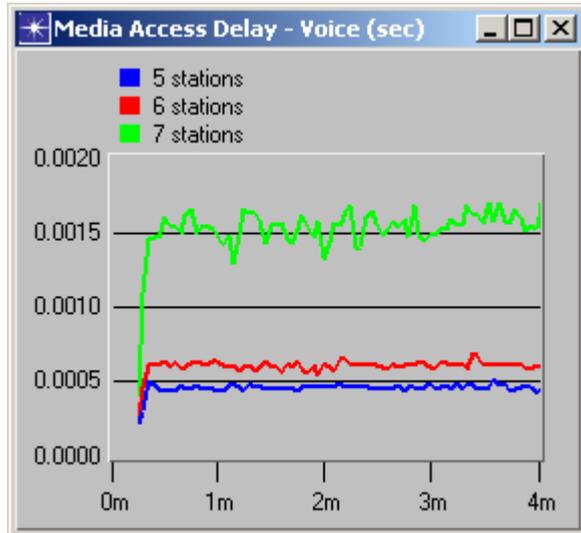


Figure 184 Maximal throughput achieved with EDCA

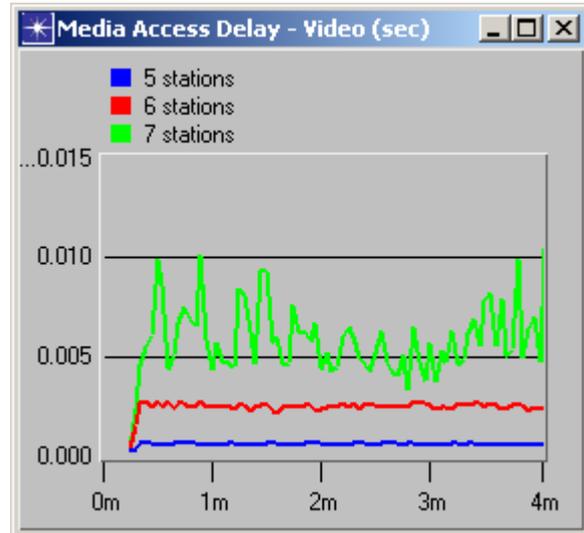
Table 52, show the MAC delay for the different access categories with five, six and seven stations. The numerical values of the average MAC delay and jitter is presented in Table 52. From the figures it is clear that the prioritisation scheme performs as expected. It can be seen that the MAC delay is acceptable for all access categories in the scenario using five stations. In the scenario with six and seven stations, the MAC delay is quickly rising for the lower access categories while the MAC delay is still very low for voice and video. Another observation, as shown by the figures and the table below, is that jitter for voice and video is very low and even with seven stations the jitter is within the limits for the QoS requirement.

Table 52 Average MAC delay and jitter from Figure 185.

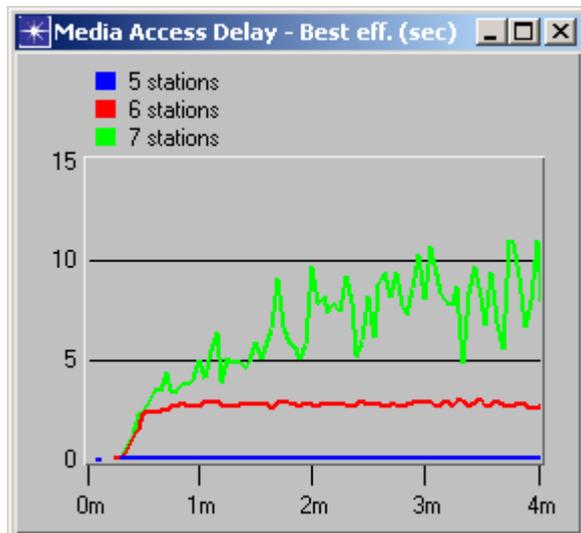
	Voice	Video	Best effort	Background
Average MAC delay (ms), 5 stations	0.45	1.0	2.5	10.0
Average MAC delay (ms), 6 stations	0.65	3.0	-	-
Average MAC delay (ms), 7 stations	1.5	6.5	-	-
Maximum Jitter (ms), 5 stations	0.04	0.06	0.56	10.0
Maximum Jitter (ms), 6 stations	0.07	0.34	310	3580
Maximum Jitter (ms), 7 stations	0.29	3.64	3580	8000



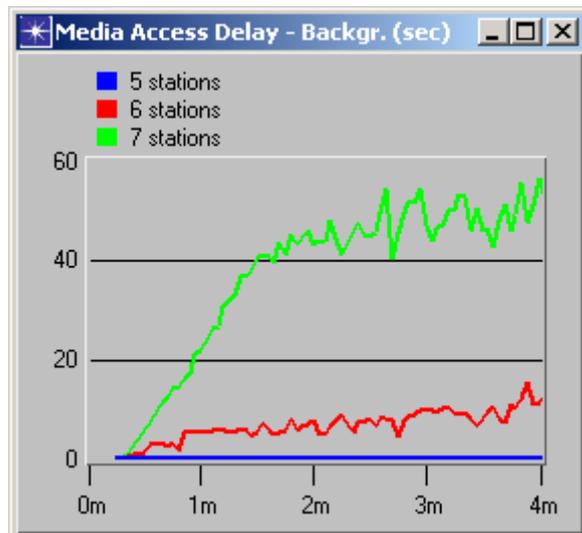
(a) MAC delay for voice traffic.



(b) MAC delay for video traffic.



(c) MAC delay for best effort traffic.



(d) MAC delay for background traffic.

Figure 185

The prioritisation between different access categories is also clear when observing the offered throughput AC by AC. In the scenarios where the network load was too high, the throughput decreased for the lower prioritised access categories. The reduced throughput in best effort and background access categories leads to full transmission buffers and therefore packets are dropped. As shown in the result, remains the throughput satisfactory for both voice and video in a saturated network.

The simulations showed that the EDCA prioritisation scheme works as expected in all aspects. The higher prioritised access categories get much lower MAC delay, jitter and are provided with their required throughput. In a saturated network the lower prioritised access categories are furnished with high MAC delays and dropped packets as a result of full transmission buffers, i.e., the performance for best effort and background applications are significantly reduced.

4.5.4 On the use of the EDCA Transmission Opportunity (TXOP) mechanism for improving the WLAN system performances

In this section the advantages of using TXOP when mobile stations operate a different bit rate are highlighted. To this end, first the influence on the system performances of the stations working at lower transmission rates is envisaged. Next, the TXOP mechanism is presented and the optimum TXOP limit is obtained for each type of traffic assumed in the study. Finally, an algorithm for dynamic adjust of the values of TXOP is presented and its performances evaluated.

4.5.4.1 Influence of stations working at lower transmission rates on system performance

As a result of varying radio channel properties a station changes its transmission bit rate at the physical layer to increase its resilience to experienced errors. However, when changing physical layer it also changes the bandwidth required for transmitting a packet, which, in case of downward shifting of physical layer, may provoke a congestion problem in the system as required bandwidth increases. Figure 186 shows the transmission times, assuming a basic access of one packet of 1kB and considering different bit rates at the physical layer. From this figure we can conclude that the time needed for transmission of the same size packet assuming 1Mbps bit rate in the physical layer is 7 times greater than when 11Mbps bit rate is considered.

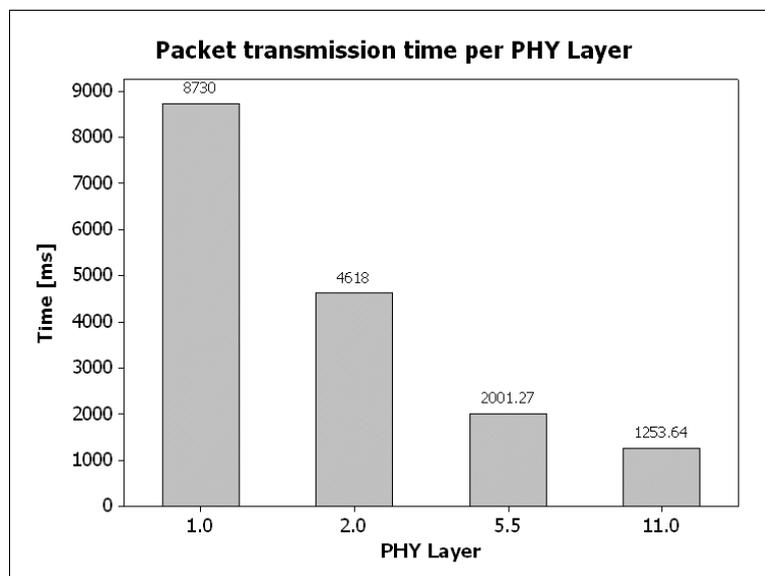


Figure 186 Packet transmission time with different physical layers

To demonstrate the great impact that this phenomena may have on system performance we have analyzed the following service mix distribution [16]:

- 5 voice stations (with an average bit rate of 24 Kbps in the downlink (DL) and 24 kbps in the uplink(UL)),
- 2 video stations (with an average bit rate of 128 in the DL and 16 kbps in the UL) and

- 3 web stations (with an average bit rate of 45.5 and 33.6 kbps in the UL)

It is assumed that some stations shifted their transmission rate from 11 Mbps to 1 Mbps whereas the rest of the stations are working at 11Mbps. The voice traffic is generated by means of G.729 A/B VoIP application with transmission rate of 24kbps. To model the video stream a Group of Pictures (GOP) composed by 12 pictures was used. The assumed rates are 25 frames per second and 128 kbps transmission rate in downlink and CBR of 16 kbps in uplink. The traffic model for web traffic considers the generation of activity periods (i.e. pages for www browsing), where several information packets are generated and a certain thinking time between them exists, reflecting service interactivity. The specific parameters are: time between pages: 4sec. in average for the UL and 5.17sec. for the DL; the average number of packets arrival per page is 25 for both UL and DL; the number of bytes per packet ranges from 1000 to a maximum of 60000 following a truncated Pareto distribution; whereas the time between packets arrival is in average 0.03125 for the UL and 0.015625 for the DL. In both cases follow a exponentially distributed statistic.

The effect on system performance of the bit rate shifting is presented in following figures.

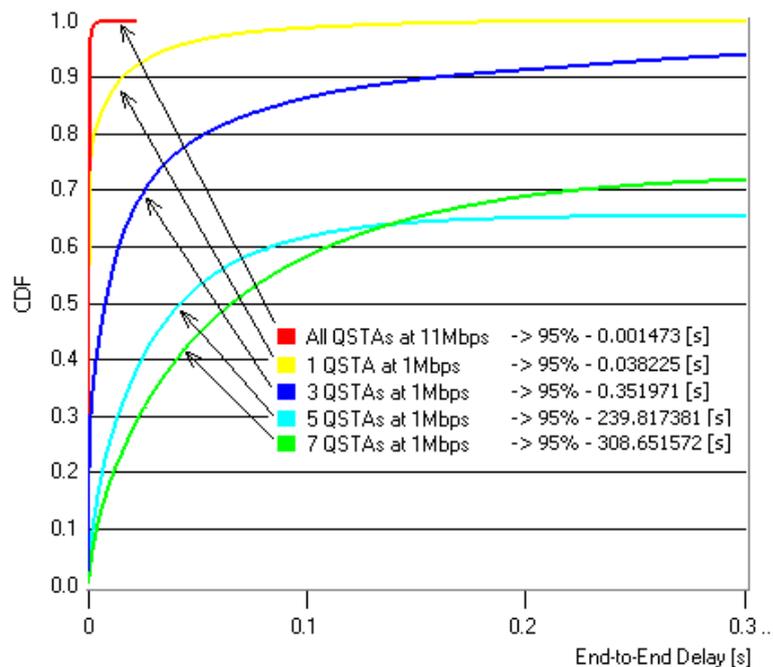
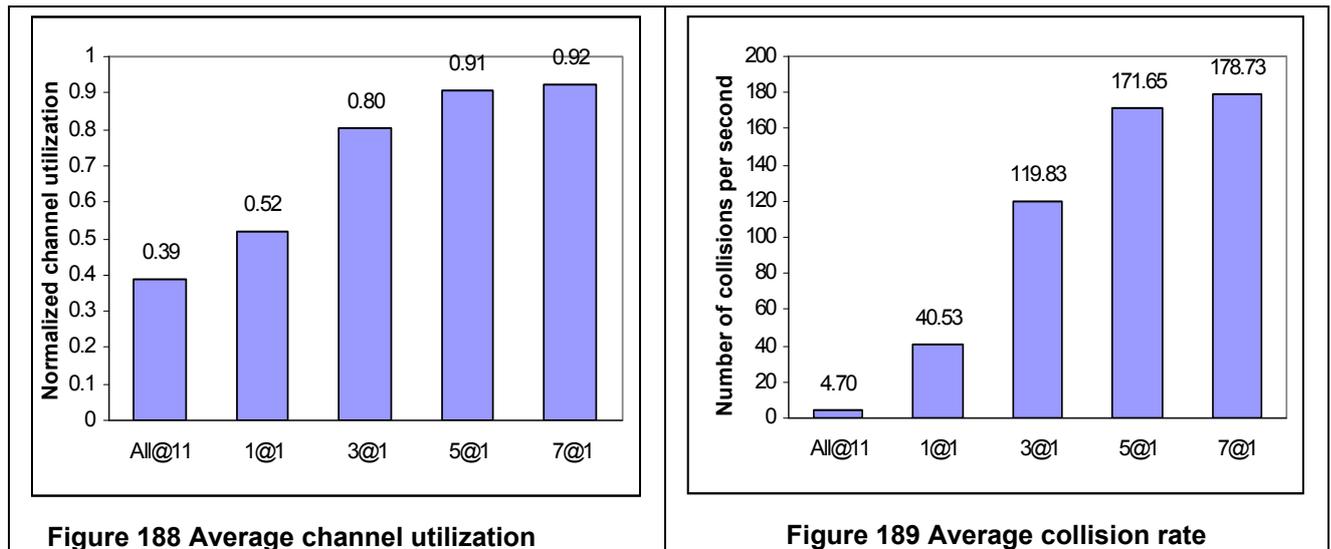


Figure 187 Cumulative distribution function of end-to-end delay at link layer level of aggregated voice streams



In performed experiments 5 cases were analyzed:

- All@11, when all stations transmit at maximum transmission rate of 11 Mbps;
- 1@1, only 1 web station transmits at 1 Mbps and the rest at 11 Mbps;
- 3@1, 3 web stations transmit at 1 Mbps and the rest at 11Mbps;
- 5@1, in this case 3 web and 2 video stations transmit at 1 Mbps and the rest at 11 Mbps;
- 7@1, in this case 3 web, 2 video and 2 voice stations transmit at 1 Mbps. and the rest at 11 Mbps;

From obtained results it is clear that even with one station working at low transmission rate the degradation on system parameters is significant. Comparing All@1 and 1@1 scenarios we see that the deterioration of all the analysed parameters is really drastic. For instance, the average collision number per second, Figure 189, increases nearly 10 times when 1Mbps is used and number of packets in voice AP queue augments even more than 40 times for the same condition, Figure 190. In case of the channel utilization, one web station at 1 Mbps boost the used bandwidth for more than 30 per cent, see Figure 188. If we analyse the system when users are working at 1Mbps at the physical layer we can observe that system parameters are getting more and more deteriorated and, with only 3 web stations using lower transmission rate, the system is on the edge of saturation region. In a situation corresponding to 5 stations transmitting at 1 Mbps a saturation state of the channel is reached and link layer delay for 95% of voice packets is lower than 239.812 seconds.

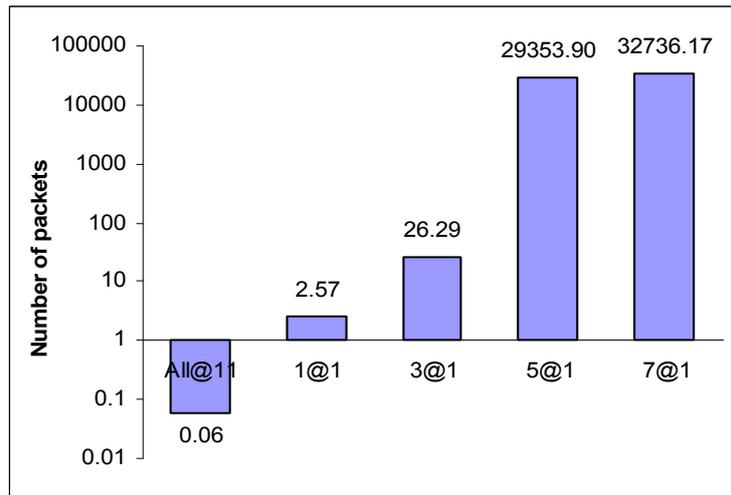


Figure 190 Average number of packets in AP voice queue

Accordingly, observed results demonstrates that transmission rate switching is a crucial issue and should be carefully considered as even one station that changes its transmission velocity to lower rate can degrade system performance.

4.5.4.2 Implementation of TXOP mechanism for system performance improvement

The new draft “e” [101] of the standard IEEE 802.11 provides a novel mechanism for packet transmission allowing multiple packet transmission by the station once the channel has been captured. This mechanism is called Transmission Opportunity (TXOP) and it is characterised by its start time and duration as shown in Figure 191. This enhancement may significantly improve the system performance as it optimizes the channel efficiency by allowing successive packet delivery. Moreover, it solves the problem related to the unknown transmission duration of the legacy IEEE 802.11 stations due to changes in the bit rate value (link adaptation mechanism). Furthermore, by applying different TXOP duration times to the different traffic types, some QoS control over these flows may be obtained. To analyze the effect of TXOP mechanism on system performance the same 5 scenarios used in the previous section were considered. In particular, the assumed TXOP duration for each type of traffic is: 14ms for voice, 6ms for video and 0ms for web (only one packet can be send per channel access for web traffic).

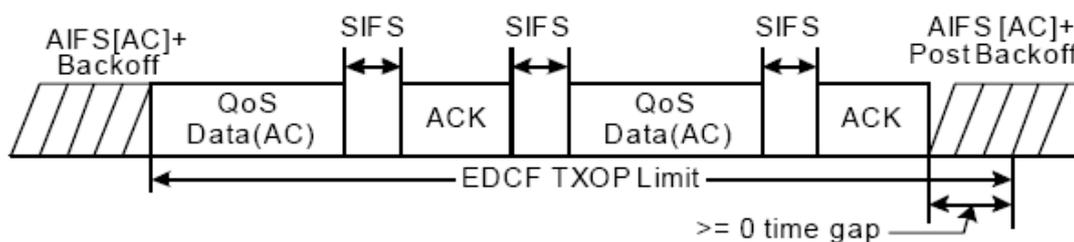


Figure 191 TXOP mechanism for EDCA channel access

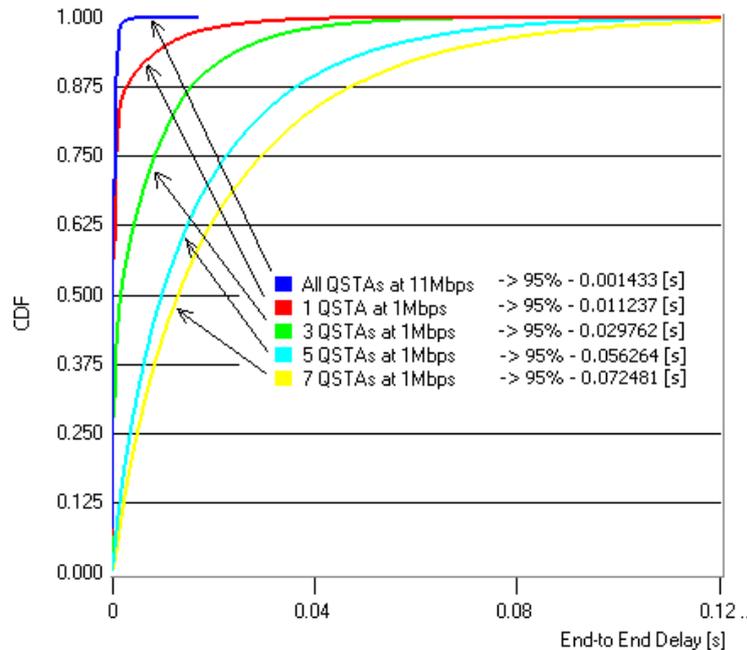


Figure 192 Cumulative distribution function of end-to-end delay at link layer level of the aggregated voice streams

Comparing the accumulative distribution function of the delay of the of aggregated voice streams with (Figure 192) and without (Figure 187) TXOP mechanism it could be realized that the TXOP mechanism has significantly enhanced the system performance. Certainly, Figure 192 shows how now, with a TXOP limit, the experienced link layer delay by voice streams is lower than the obtained without TXOP limit (see Figure 187). For instance in the case of (3@1 the obtained bound delay for the 95% of the cases is 0.0298 sec. with TXOP whereas without TXOP this limit reaches 0.3520 sec.). This improvement is due to the fact that when using packet bursting there is lower number of packets contending for a channel access. Moreover, also notice that the obtained delay values for voice traffic are within required limits (less that 100 msec.) even in deep saturation scenario 7@1 when TXOP mechanism is used.

In addition to that, the number of packets in voice queues also decrease as each voice station is allowed to send more than one packet after winning channel access, such as it is shown in Figure 195.

However, the channel utilization does not change significantly. For instance, assuming 3@1 the channel occupancy is 80% without TXOP and 72% 3@1 with TXOP (see Figure 188 and Figure 193) This result could be expected because the channel utilization value represents the current radio resource requirements of the system, which are the same in both cases. In nay case, the small difference between the obtained values is a result of a better channel utilization in the scenario with TXOP. Moreover, as a result of better bandwidth administration when TXOP is considered, the number of collisions decreases around 40 per cents as demonstrated comparing Figure 189 and Figure 194. That is, the system wastes 40% less bandwidth compared with the without TXOP case, resulting in an improvement of system parameters like end-to-end delay, shown in Figure 192.

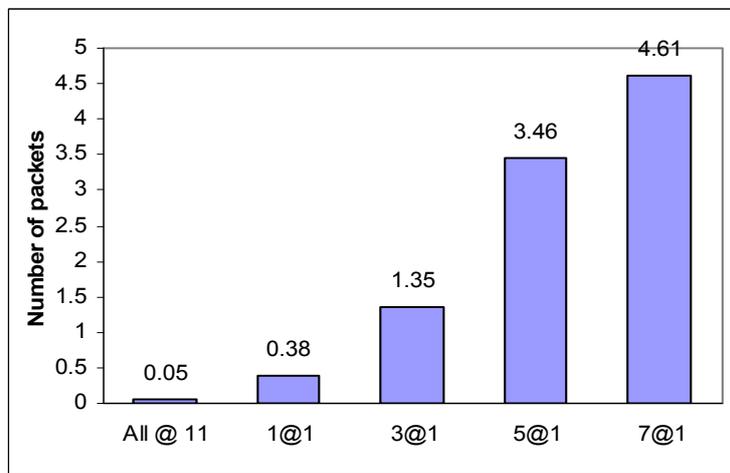
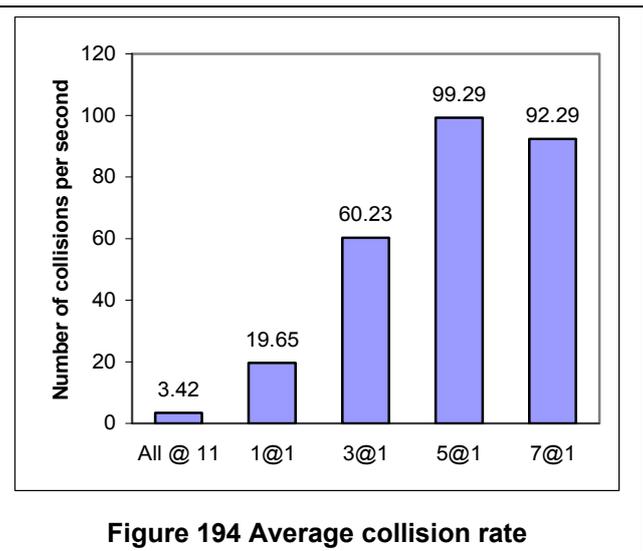
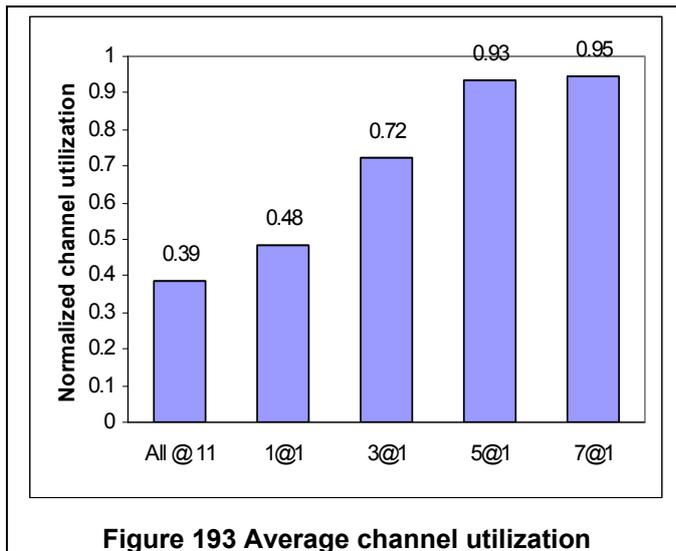


Figure 195 Average number of packets in AP voice queue

In summary, from the results we could conclude that that applying TXOP limits to high priority traffics like voice and/or video the system performance improves, allowing to satisfy the QoS requirements (e.g. like delay) for the highest priority traffics.

Once the benefit to apply the TXOP opportunity has been demonstrated, the question is: Is there any optimum TXOP limit for the different traffic types?

4.5.4.3 Optimum TXOP limit for each type of traffic

The system performance will definitely vary with the different values of TXOP limit as stations will be allowed to send different number of packets on a single channel access. Therefore we realize a study to find out if there is an optimum value of TXOP limit or not. In this analysis the scenarios assumed is one of the described in the previous sections, namely with the one with 3 web station operating at 1 Mbps physical layer. In particular, the dependence between the TXOP limits for voice and video traffics, assuming that the web flows and best effort flows can only send one packet per channel access, is investigated. Two statistics, the end-to-end delay on link layer level for 95% of packets and the number of packets in AP queues in 95% of cases, were used as system performance.

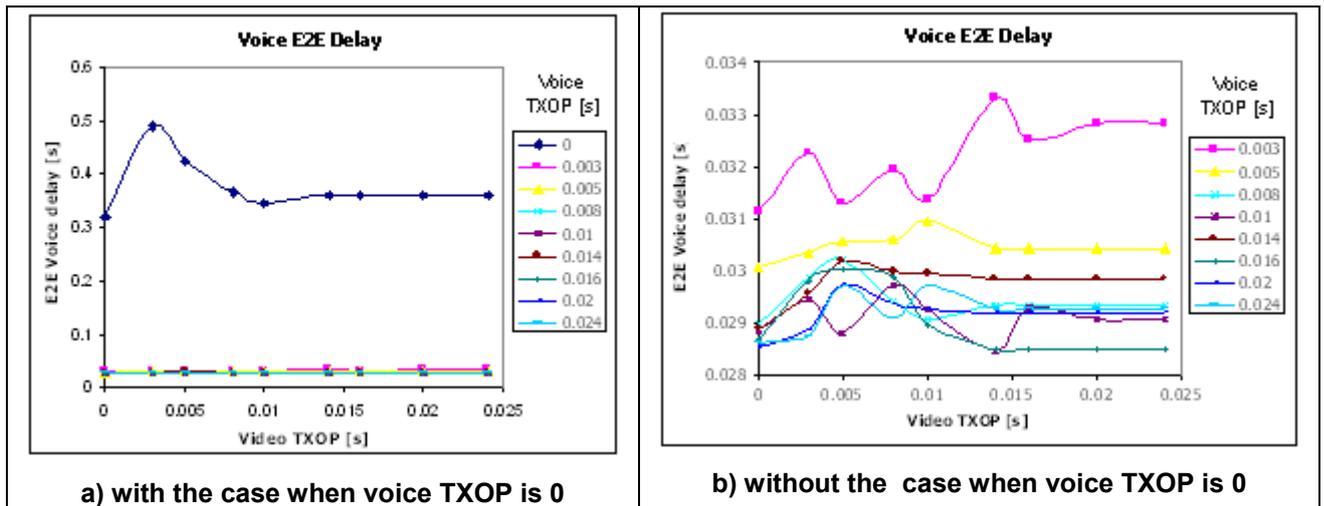


Figure 196 End-to-End delay at link layer level for 95% of voice packets for different values of voice TXOP limit and as a function of video TXOP limit

From Figure 196 we see that by assigning to TXOP limit for voice traffic a value greater than transmission time of two voice packets (3ms in our case) a decrease in end-to-end delay of 10 times is obtained. Moreover, the value of TXOP limit for video streams does not have significant influence on the voice link layer delay while TXOP duration for voice flows is greater or equal to 3ms. Moreover, applying a sufficiently big value to TXOP duration for video traffic (greater than 15 ms) a constant value of voice link layer delay can be reached. Nevertheless, this value has to be chosen carefully because too high value of TXOP for video streams may not give required outcome and even increase the link layer delay of voice flows in situations when the number of video flows is much higher than voice flows.

Analyzing Figure 197, we observe that, above a video TXOP duration of 5ms, the delay experienced by video flows is nearly constant and independent of voice TXOP limit.

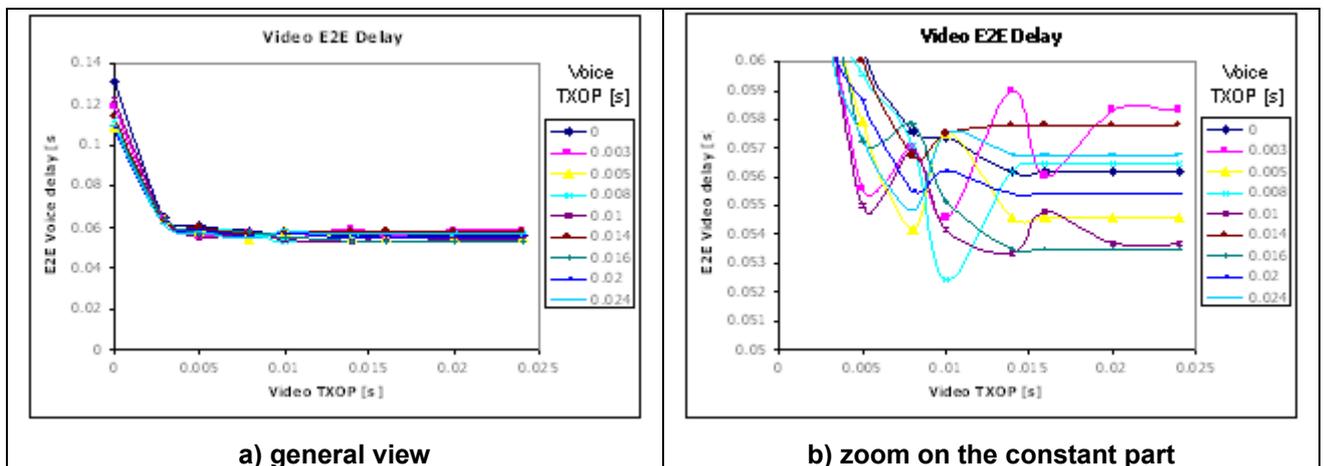


Figure 197 End-to-End delay at link layer level for 95% of video packets for different values of voice TXOP limit and as a function of video TXOP limit

Another interesting observation can be drawn when looking at end-to-end delay for web traffic, Figure 198. Even though, the best effort stations only send one packet each time that access the channel, they can achieve lower delay (even 50% lower) when one or both higher priority classes are allowed to send more than one packet after winning channel contention procedure.

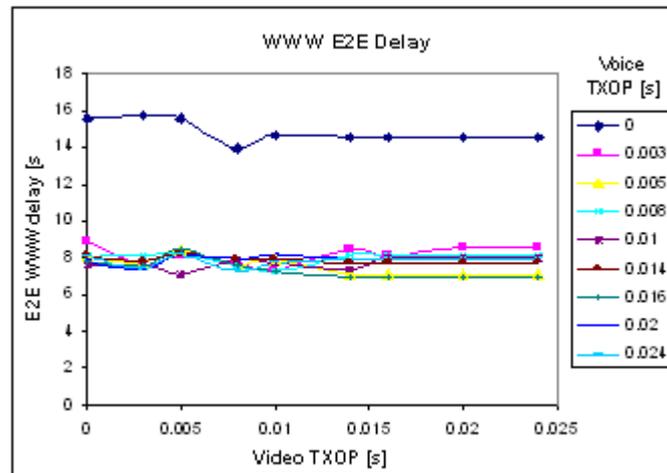


Figure 198 End-to-End delay at link layer level for 95% of web packets for different values of voice TXOP limit and as a function of video TXOP limit

Similar conclusions to the ones drawn from end-to-end delay can also be drawn in the case of the number of packets in AP queues. We decided to analyse the number of packets from AP queues as one of the metrics, as all the traffic has to go through the AP, which consequently has much higher load than other stations. Therefore, for each traffic class, the behaviour experienced by the packets in the AP provides a lower bound for behaviour experienced by the packets in other stations.

The obtained results resemble the link layer delay graphs presented above. Consequently, for a voice TXOP limit value greater than or equal to 3 ms, Figure 199, a significant reduction of number of packets in AP voice queue is seen (for instance assuming a video TXOP of 14ms the average number of packets in the queue is 2.65 packets with a voice TXOP higher or equal to 3 ms, whereas the average number of packets stored grow-up 94.83 packets without voice TXOP). Furthermore, when voice TXOP limit is greater or equal to 3ms, the influence on the average number of stored packets in AP voice queue is almost negligible.

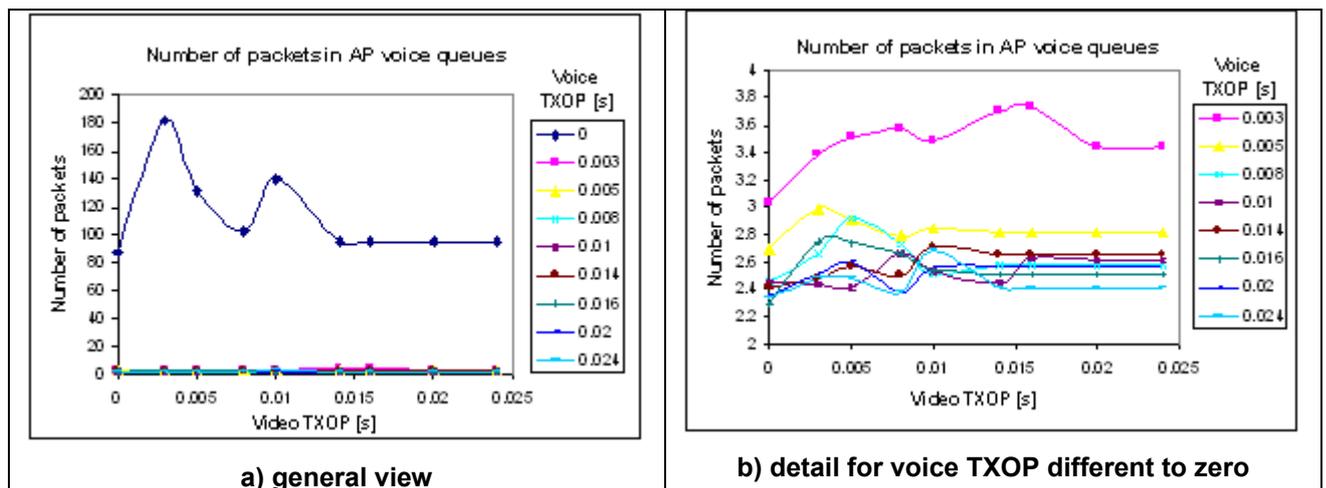


Figure 199 Number of packets in the AP voice queue in 95% of cases for different values of voice TXOP limit and as a function of video TXOP limit

In case of video queue, the minimum value of the packets in the queue is also reached when video TXOP is greater or equal to 5ms, Figure 200. This performance results from the fact that with TXOP of 5 ms the video stations are able to transmit all the packets of their queues independently of the value of TXOP for voice stations.

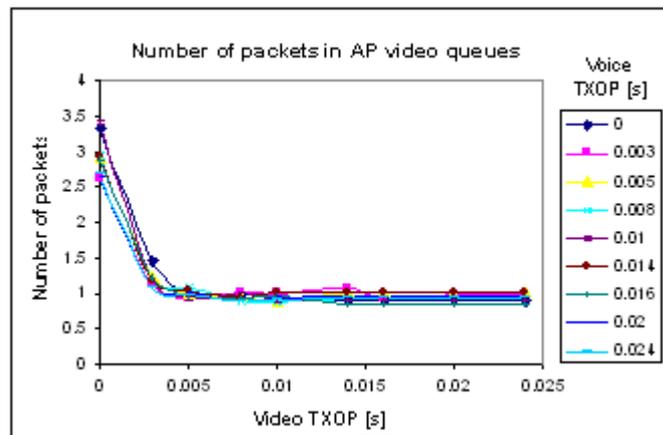


Figure 200 Number of packets in the AP video queue in 95% of cases for different values of voice TXOP limit and as a function of video TXOP limit

As could also be expected, the number of packets in the web queue is reduced when voice and video streams are allowed to send more than one packet per channel access as observed in Figure 201.

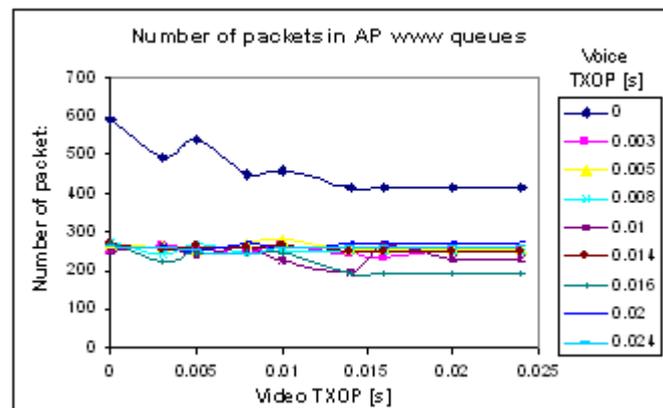


Figure 201 Number of packets in the AP web queue in 95% of cases for different values of voice TXOP limit and as a function of video TXOP limit

In summary, given a service mix distribution the optimum TXOP duration for high priority traffics that optimize system performance can be obtained. For the service mix assumed in this work, the optimum performance for voice traffic is obtained with a TXOP limit greater than 3ms whereas for video streams the TXOP duration should be greater than 5ms. Obviously, these values will be different if the service mix distribution varies.

4.5.4.4 Dynamic TXOP limit configuration

As explained in previous section, the major packet concentration is in the AP because it sends all downlink traffic. Moreover, usually the downlink traffic is superior to the uplink traffic due to the traffic asymmetry of the envisaged services. In order to adapt the TXOP limit dynamically in this section an algorithm that varies TXOP limits for each type of traffic based on the number of packets in AP queues is introduced. The rationality of the algorithm follows:

- Average number of packets in a queue can be calculated by means of Little's equation[119]:

$$N[AC] = T[AC] * \alpha [AC]$$

where:

N – is an average number of packets per AC queue;
 T – is an average MAC delay of packets at each AC;
 α is an average inter-arrival time of packets at each AC.

- Once, the average number of packets per queue and average packet length (Pq_Length[AC]), computed as the time needed for transmit it over physical layer are available, the TXOP limits can be calculate as:

$$TXOP_Limit[AC] = N[AC] * Pq_Length[AC] * \alpha [AC]$$

where:

$\alpha [AC]$ – is an upper limiting coefficient for TXOP limit. The value of the $\alpha [AC]$ parameter is 1 for non-saturation conditions and less than 1 for saturation conditions.

The above algorithm was implemented in the simulator in a following way. To obtain average number of packets in a queue, the AP estimates it by means of following equations:

$$q_total_{[i]} = q_total_{[i-1]} + q_size \quad (49)$$

where

q_total - is calculated with every packet arrival to corresponding transmission queue

q_size – is a size of a queue to which packet arrives including this packet

$$N = \left\lceil \frac{q_total}{pq_num} \right\rceil \quad (50)$$

being,

pq_num - is a number of arrived packet to corresponding transmission queue

N - is an average number of packets in a queue estimated every beacon interval, in our case every 100 ms.

After the estimation of the value of N, parameters q_total and pq_num are initialized to zero thus with first packet arrival $q_total_{[1]} = q_size = 1$.

To calculate TXOP limit, the packet length is needed. First of all, for simplicity, a constant voice packet size is assumed (e.g.60 Bytes). Thus, its transmission considering 11Mbps bit rate in the physical channel will take around 560 us. However, as voice is the highest priority traffic and it is a dominant service, according to the EVEREST assumptions, we decided for this service to allow more than one packet per channel access, because, as shown in Figure 196 and Figure 199 is advantageous for the whole system. Consequently, the minimum TXOP duration assuming a three times packet transmission duration is 0.00168 seconds. This will guarantee at voice flows some basic improvement. Hence, TXOP limit for voice is calculated using expression (51):

$$TXOP_Limit = 3 \times 0.00056 \times N [s] \quad (51)$$

Regarding the video TXOP limit, we use as packet duration parameter the transmission time value needed for sending the largest possible packet at MAC layer when 11Mbps is considered. Then the equation is:

$$TXOP_Limit = 0.00221 \times N [s] \quad (52)$$

Applying above mentioned algorithm to our simulator, let us first check how good the estimated number of packets in a queue is. In Figure 202 the number of packets¹² in the AP voice queue is marked in colour blue, whereas the estimated average of a number of packets in the AP voice queue, obtained by means of the proposed algorithm, is marked in red.

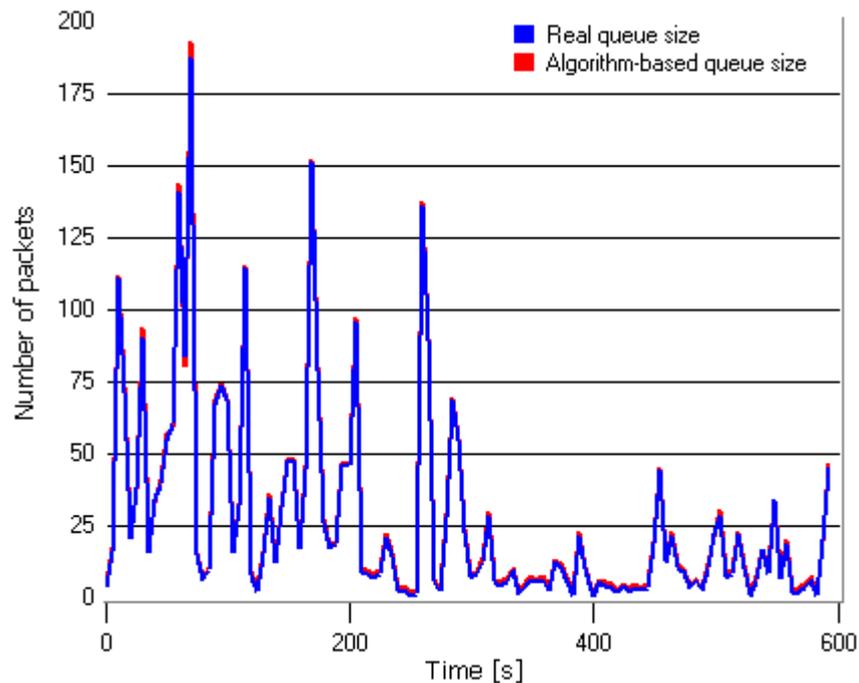


Figure 202 Comparison of the real number of packets in AP voice queue and the number obtained by proposed algorithm

The slight difference between these two graphs is due to the fact that the proposed algorithm only computes the queue size when a packet arrives.. However, this minor difference has not importance because this value is further multiplied by some factors to finally get TXOP limit. The comparison of the system performance with dynamic and optimum TXOP durations is summarised in Table 53 and Figure 203 and Figure 204.

Table 53 System parameters for two approaches for TXOP configuration: optimum and dynamic

		E2E Delay in 95% [s]	AP avg. pkt. num. in queues	Lost pkts	Retrans. attempts per pkt	Avg. TXOP limit [s]
Optimum	Voice	0.0284	1.363	94	0.21820	0.014
	Video	0.0552	0.471	8	0.21095	0.006
	Web	7.3910	48.935	2	0.27758	0
Dynamic	Voice	0.0296	1.499	131	0.22349	0.0034
	Video	0.0576	0.508	3	0.22523	0.0041
	Web	7.1637	52.582	2	0.28685	0

¹² obtained as a time average of a number of packets in the queue when a packet arrives or leaves the queue

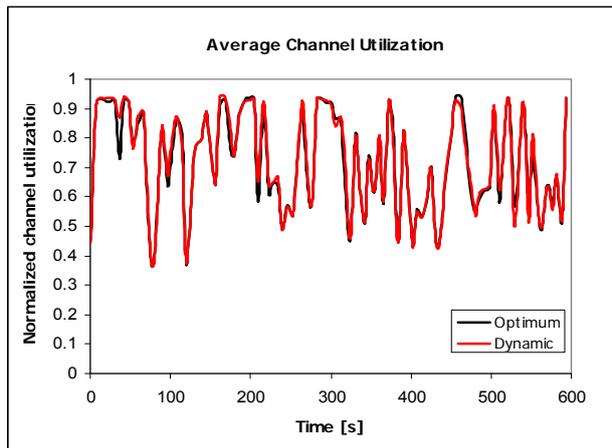


Figure 203 Average channel utilization for optimum and dynamic TXOP configuration

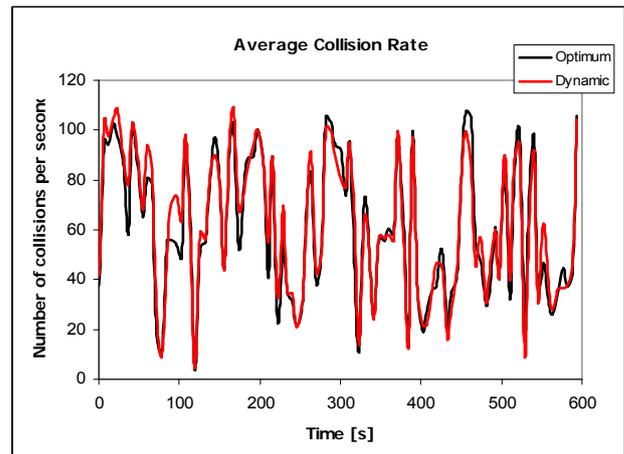


Figure 204 Average collision rate for optimum and dynamic TXOP configuration

The results achieved with optimum and dynamic TXOP limits are quite similar. The dynamic procedure to obtain the TXOP limit value results in slightly inferior parameters values than the obtained for the optimum case. For instance, the channel utilization is 72.9 % for optimum case whereas the dynamic estimated value is 72.3% or in the case of the average value of collision number per second the obtained value is of 62.4 collisions for optimum case whereas the estimated value is 60.5. In addition, there is no significant difference between the link layer delays metric obtained and estimated neither between number of packets in AP queues metric.

Therefore, we can conclude that TXOP limits obtained with the proposed algorithm are close enough to the optimum values and hence a quasi-optimal system performance is reached. However, the great advantage of dynamic configuration of TXOP limit is the independence of the system set-up.

4.5.4.5 Packets fragmentation to enhanced QoS guarantees for high priority traffics

Although the TXOP mechanism allows multiple packet transmission it may provoke packet fragmentation if packet transmission duration is superior to its TXOP limit. Fragmentation of packets, in most cases deteriorates system performance due to the additional packet overhead introduced to the system. Moreover, fragmentation will increase packet end-to-end delay and will provoke higher packets accumulation in queues.

However, applying fragmentation to lower priority traffics (or non real time traffics) may limit their transmission possibilities and in consequence increase transmission probabilities of higher priority traffics. Therefore depending on which traffic class fragmentation is applied, by means of TXOP duration, we could further enhanced traffic prioritisation.

In our scenario web traffics are a limiting factor as they are transmitted at the lowest velocity. As web traffic is not timely concerned, we fragmented each packet to fragments of 0.003 seconds duration. Jointly with the web packets fragmentation, we assume dynamic configuration of the voice and video TXOP limits. The obtained results are compared with the optimum case, and presented in Table 54.

Table 54 Comparison of system parameters between optimum and dynamic with fragmentation TXOP configuration scenario

		E2E Delay in 95% [s]	AP avg. pkt. num. in queue	Lost pkts	Avg. ch. util. [%]	Avg. coll. rate
Optimum	Voice	0.0284	1.363	94	72.3	60.7
	Video	0.0552	0.471	8		
	Web	7.3910	48.935	2		
Fragmentation	Voice	0.0204	0.949	76	74	52.8
	Video	0.0395	0.307	3		
	Web	9.3001	70.639	1		

Table 54 shows that with fragmentation of web packets further improvements of higher priority traffics parameters is possible but with simultaneous deterioration of web traffic conditions. Consequently, the decrease of the link layer delay from 0.0284 seconds to 0.0204 seconds is achieved with the concurrent increase of web link layer delay from 7.39 seconds to 9.30 seconds. The comparison of end-to-end delay at link layer level between three configuration scenarios of TXOP limits: optimum, dynamic and dynamic with fragmentation is presented in Figure 205, where clear advantage of fragmentation is shown. Moreover, fragmentation results in higher channel utilization due to bandwidth overhead introduced by the headers and the acknowledgements of each fragment. In addition, decreases the number of collisions per second due to the fact that generated fragments are short and do not provoke packet accumulation in higher priority queues and hence there is lower number of simultaneous channel access.

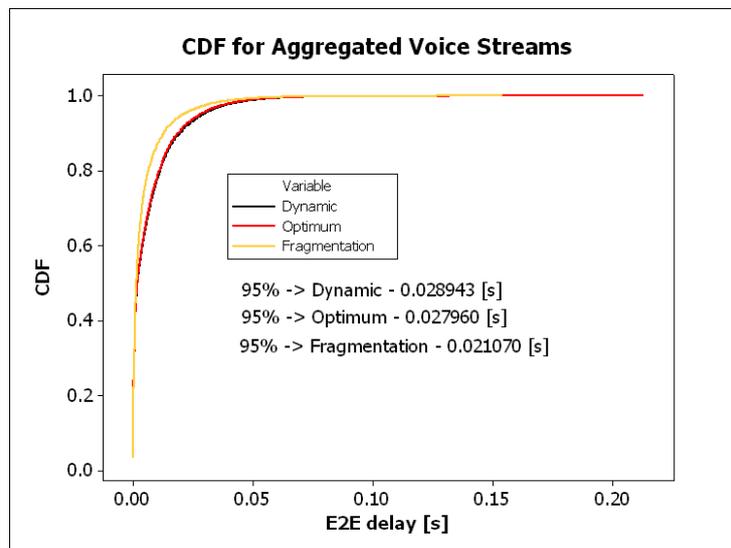


Figure 205 Cumulative distribution function of the end-to-end delay at link layer level of aggregated voice streams

4.5.4.6 Conclusions

In the previous sections, the benefit of using the Transmission Opportunity and fragmentation mechanisms has been evaluated and demonstrated through simulations.. By means of this option destructive influence of downward transmission rate shift can be control at some extend, giving higher priority, through higher TXOP limit, to stations with high QoS requirements. Moreover, simulation results show that there exist optimum TXOP limit values in terms of system performance. However, as these values change with service mix distribution a novel dynamic TXOP configuration algorithm based on the number of packets

in AP queues has been proposed. The developed algorithm provides TXOP limits that are within minimum most advantageous TXOP limits and hence quasi-optimal system performance is reached independently of system service mix.

Finally, by using the TXOP limit to perform fragmentation of low priority packets, further enhancement of prioritization mechanism is obtain.

5 COMMON RRM

5.1 INTRODUCTION: THE CRRM FRAMEWORK

The heterogeneous network concept is intended to propose a flexible and open architecture for a large variety of wireless access technologies, applications and services with different QoS demands, as well as different protocol stacks. Figure 206 shows an example of such heterogeneous networks scenario. It is constituted by several radio access networks (RAN) interfacing a common core network. Radio access networks include cellular networks, e.g. UTRAN (UMTS Terrestrial Radio Access Network) with the two modes FDD (Frequency Division Duplex) and TDD (Time Division Duplex), and GERAN. These networks may in turn be subdivided in different cellular layers (e.g. macro, micro or picocells) depending on the expected coverage area, and also other public non-cellular access networks (e.g. WLAN). The core network infrastructure is typically subdivided in the circuit switched (CS) and packet switched (PS) domains providing access to external networks, e.g. PSTN (Public Switched Telephone Network) or Internet. These external networks can also include other private and public WLANs, from which terminals may also have access to the core network. The scenario assumes the existence of multi-mode terminals, providing connectivity to multiple access networks either in different time instants or even simultaneously.

Wireless access networks differ from each other by air interface technology, cell-size, services, price, access, coverage and ownership. The complementary characteristics that these networks may offer make possible to exploit the trunking gain resulting from the joint consideration of the different networks as a whole, thus leading to a better overall performance than the accumulated performances of the stand-alone systems. However, constraints deriving from non-compatibility among RATs (e.g. offered bit rates, which also vary with cell size even with identical technologies) as well as terminal capabilities (e.g. the existence of a number of single-mode terminals that can only be connected to one specific network) may limit this potential trunking gain.

This challenge calls for the introduction of new RRM algorithms operating from a common perspective that take into account the overall amount of resources offered by the available RATs, and therefore are referred as CRRM (Common Radio Resource Management) algorithms. Furthermore, for a proper support of such algorithms, suitable network architectures and procedures must ensure the desired interworking capabilities between the different technologies.

These new scenarios where different RATs will coexist and will operate in a coordinated way are often referred as beyond 3G systems. The interworking architecture enhanced with CRRM functionality will pave the way for the extension of these heterogeneous networks to include also new 4G radio access technologies.

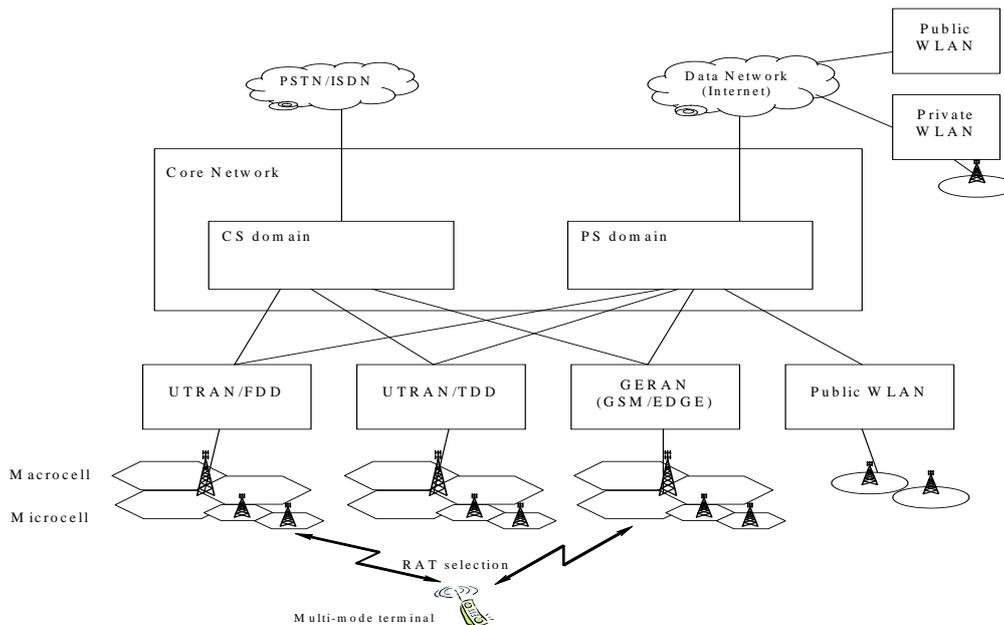


Figure 206 Heterogeneous networks environment

5.1.1 CRRM functional model

Common Radio Resource Management (CRRM) refers to the set of functions that are devoted to ensure an efficient use of the available radio resources in heterogeneous networks scenarios by means of a proper coordination between the different radio access networks. The functional model assumed in 3GPP (Third Generation Partnership Project) for CRRM operation considers the total amount of resources available for an operator divided into radio resource pools. Each radio resource pool consists of the resources available in a set of cells, typically under the control of a RNC (Radio Network Controller) in UTRAN or a BSC (Base Station Controller) in GERAN. Two types of entities are considered for the management of these radio resource pools [121][122], as shown in Figure 207.

- The RRM entity, which carries out the management of the resources in one radio resource pool of a certain radio access network. This functional entity involves different physical entities in the RNS (Radio Network Subsystem) or BSS (Base Station Subsystem) depending on the specific considered functions, although for representation purposes it is usual to assume the RRM entity residing in the RNC or the BSC.
- The CRRM entity, which is involved in the coordinated management of the resource pools under different RRM entities. In this way, decisions on radio resources usage may take into account the resource availability in several RRM entities.

The interactions between RRM and CRRM entities involve mainly two types of functions:

a) Information reporting function

The information reporting function allows the RRM entity to report relevant information to its controlling CRRM. The reporting can be performed periodical or event-triggered, or even at given instant, and it is totally up to CRRM entity's request. The reported information consists in dynamic measurements and static information on cells controlled by a RNC or BSC.

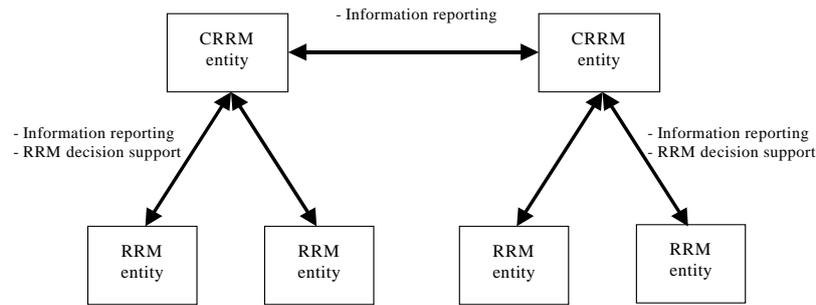


Figure 207 CRRM functional model

b) RRM decision support function

This function describes the way how the CRRM and RRM entities interact for taking decisions. For example, it is possible that the CRRM simply advises the RRM entity, so that the RRM remains as the master of the decisions, and, on the contrary, it is also possible that the CRRM is the master so that its decisions are binding for the RRM entity.

5.1.2 CRRM functionalities

CRRM is designed to coordinately manage resource pools over the heterogeneous air interface in an efficient way. This efficiency depends on how to construct its functionalities. There exist a range of possibilities for the set of functionalities that CRRM entity may undertake, which mainly depend on the following two factors:

1. RRM or CRRM entity is the master to make radio resource management decisions.
2. The degree of interactions between RRM and CRRM entities

In Deliverable [17] and in [123], these possibilities were further explored by considering how to incorporate the CRRM functionalities into the RRM functionalities to support the different procedures. In the following, the main points described in [17][123] are summarised.

The RRM functionalities arising in the context of a single RAT are the admission and congestion control, the horizontal (intra-system) handover, the packet scheduling and the power control. When these functionalities are coordinated between different RATs in a heterogeneous scenario, they can be denoted as “common” (i.e. thus having the common admission control, common congestion control, etc.). In turn, when an heterogeneous scenario is considered, two specific additional functionalities arise, namely the Initial RAT selection (i.e. the functionality devoted to decide to which RAT a given service request should be allocated) and the Vertical (inter-system) handover (i.e. the functionality devoted to decide a seamless RAT switching for an on-going service).

Then, the different possibilities that are envisaged when considering the operation between RRM and CRRM entities are the following ones[17]:

- No CRRM functionalities: In this case, it is considered that, although different RATs operate in a heterogeneous scenario, no coordination among them is carried out and, consequently, no specific functionalities are associated to CRRM level. In this case, the Initial RAT selection and Vertical Handover algorithms are associated with RRM entities, so that the decisions are taken without any knowledge from the radio network conditions in other RATs.

- Initial RAT selection and Vertical Handover: In this situation, as depicted in Figure 208 the RAT selection procedures are associated with the CRRM entity. The local RRM entities provide RRM measurements including the list of candidate cells for the different RATs and cell load measurements, so that the CRRM can take into account the availability of each RAT for the corresponding mobile terminal.

- Common Admission and Congestion control: This approach consists in moving to the CRRM entity those local functionalities that operate on a longer-term basis, like the admission and congestion control algorithms, while keeping in the local RRM entities the functions that operate at the radio frame level or below, like the packet scheduling or the power control.

- Common Packet Scheduling: This approach, shown in Figure 209, provides the highest degree of interaction between CRRM and local RRM by executing joint scheduling algorithms in the CRRM entity. The local RRM functionality would remain to a minimum, limited to the transfer of the adequate messages to CRRM and some specific technology dependent procedures that occur in very short periods of time (e.g. inner loop power control in case of UTRAN, which occurs with periods below 1 ms). This solution would require for CRRM decisions to be taken at a very short time scale (in the order of milliseconds).

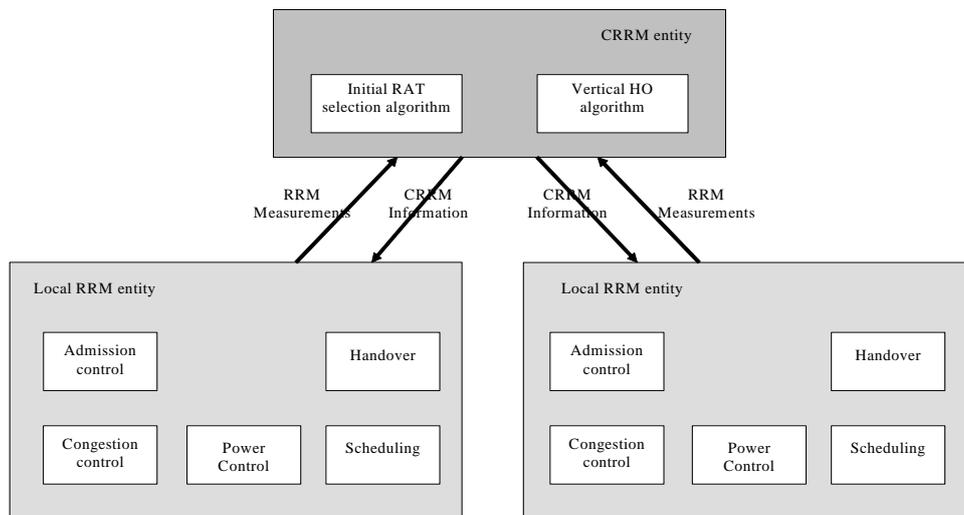


Figure 208 CRRM carrying out initial RAT selection and vertical handover

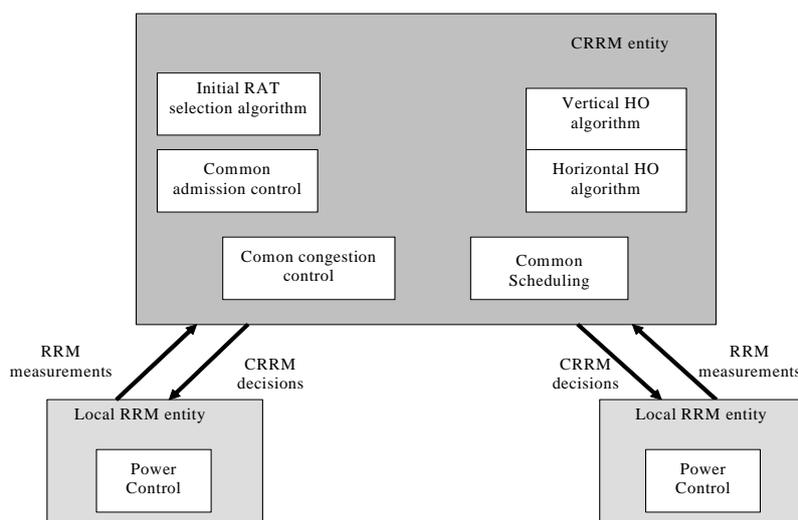


Figure 209 Common scheduling in the CRRM entity

5.1.3 CRRM implementation

The suitable approach for the implementation of CRRM mechanisms clearly depends on the functionalities associated to CRRM, since these will set the requirements in terms of

interactions between CRRM and RRM entities (i.e. information reporting and decision support). Further, such interactions will only be possible provided that the proper interworking capabilities among the different RATs are enabled, as discussed in section 4.1 of [17]. In all approaches it is important to note that the trade-off between the highest possible gains and the additional delay and signalling load must be considered. A more detailed explanation of this topic can be found in [17].

5.1.4 Scope of this chapter

According to the above framework, and taking into account the scope and time-frame of the EVEREST project, the CRRM studies reported here assume the functionality split shown in Figure 208, in which the CRRM takes charge of the RAT selection procedures, including both initial RAT selection and vertical handover, while the RAT-specific RRM algorithms, like admission control, congestion control or packet scheduling are executed locally at the RRM entities. Then, the chapter will provide an analysis of different RAT selection strategies devised according to specific policies. In particular, the study will start with the analysis of the service-based RAT selection policies and will continue with the policies devised according to a load balancing principle that tries to keep the same load level in the considered RATs. Finally, a RAT selection strategy according to path loss measurements will be analysed as well as a strategy based on RAT priority list. The chapter will conclude with the study of CRRM over TCP throughput and with the analysis of the implications of having a certain number of single mode terminals in the scenarios over the CRRM algorithms.

5.2 RRM POLICIES IN HETEROGENEOUS NETWORKS

The policies for RRM can differ when users, networks and services appear in a heterogeneous mix. Operators typically want to offer high QoS for user requesting special service level agreements.

The policy rules for QoS and user preferences must have major impact on the algorithms optimising the system capacity and link throughput. The Common RRM optimisation of capacity over all RATs while keeping the QoS requirements and user preferences is critical and specific service level agreements (SLA) must be met.

The policy rules have been taken into account and incorporated in the RAT selection and the algorithms for load balancing evaluated in the next sub-sections.

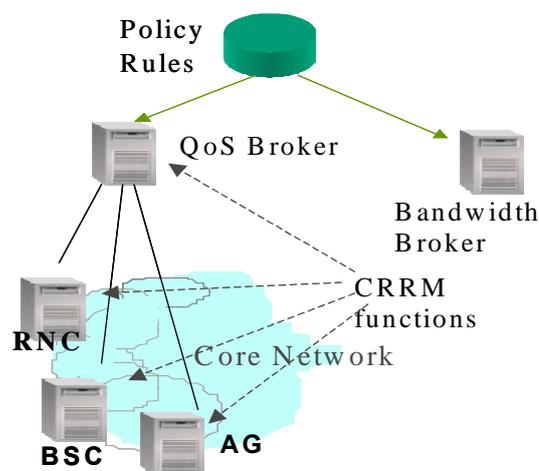


Figure 210 Operator policy rules impact on CRRM functions.

Each of the radio technologies has specific entities for dealing with the radio resource management, so, each one has knowledge of its radio resources availability, but does not know the status of the others. To handle with this, 3GPP proposes a Common Radio Resource Manager that should be aware of the radio resources availability in each technology, deciding to which one the user should be connected to.

Whenever there is more than one radio technology that may attend the services demands, it is needed to decide in which of them the mobile should be linked to. Several factors may influence that decision: the network accessibility and the radio resources availability are the most visible ones, but the operator preferences should also be taken into consideration. Next graphic shows the overall factors that may influence the final PDF decision. As can be seen, besides the network availability and the services demands, the commercial strategies also play a role. Based on all of these factors, the PDF (Point Decision Function) [124] will be able to make the right decision related with each user at each time.

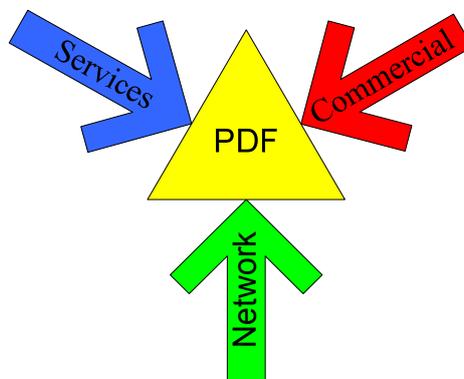


Figure 211 Factors influencing the PDF

By adding a set of rules that goes further the resources availability, operators may put in action their marketing strategies, impacting on the resource management. The problem of taking the commercial issues into the radio management is that these rules are most of the times contradictory between them: to the same service, an operator for one reason may wish to provide one radio technology and, at the same time, due to another commercial motive it may wish to use other technology; therefore it is needed to get an algorithm that solves this operator opposite preferences.

In some situations, like in admission control, it is needed to look for all the RATs and see which one best fits the services demands. When the information sent by the WQB says that there is only one RAT covering the user area then the operator preferences are good for nothing. But when the WQB informs that there are several RATs available, then the PDF must look into the operator preferences. Therefore, it is needed an algorithm to perform the RAT selection when there are more than one RAT available, which supports the desired service.

It may also happen to have several users contending for the same resources as, for example, when all the RATs are at the maximum of their capacities; hence the PDF needs to be able to solve this contention difficulty by using a contention solution procedure. To differentiate the user characteristics operators may develop user profiles giving special network accesses to the "most important" ones. The profile is used whenever it is desired to distinguish between customers benefits. When there are several users contending for the same resources, the operator may wish to provide its resources to the VIP users.

In [1], it was proposed an algorithm that takes into consideration all the operator preferences, even the opposite's ones, and allocates different weights to the operator preferences. It solves the contradictory rules problem, enabling the operator commercial strategies to have a real impact on the radio selection that should be used by the customer to access their services within a heterogeneous network environment. Also in [1], it was developed a new algorithm that solves the contending situations by giving different benefits to the users, allowing special access privileges to the main ones. This means that users are differentiated according to their profiles and in a congestion situation they may or may not access the network resources depending on their privileges.

5.3 SERVICE-BASED RAT SELECTION POLICIES

5.3.1 Introduction

Within the set of radio resource management functions, the initial RAT selection and the vertical or intersystem handover are devoted to decide the appropriate RAT for a given service at session initiation and during the session lifetime, respectively. Therefore, they necessarily involve different radio access technologies and it is appropriate to devise them from a CRRM perspective. In that sense, the algorithm operation might then respond to specific policies taking into account both technical and/or economical aspects (e.g. operator or user preferences).

Not much effort has been devoted up to date in the open literature to deal with the CRRM problem. In [125] the benefits of CRRM in terms of inter-system handover and inter-system network controlled cell reselection are analysed in a heterogeneous UTRAN/GERAN scenario. With respect to the combination of cellular and WLAN technologies, in [126] a methodology based on fuzzy logic and reinforcement learning mechanisms is presented that combines technical and economical issues to provide the specific RAT and bandwidth allocations. Similarly, in [127] a joint scheduling algorithm between UMTS and HIPERLAN is proposed. On the other hand, access selection in heterogeneous networks has been also covered in the literature in a number of papers. To cite only a couple of them, please refer to [128] and [129] for details.

Within the CRRM context, this chapter discusses a framework for developing policy-based initial RAT selection algorithms and provides an insight into some sample policies taking into consideration service-based criteria. Then, the results will be extended to include also radio network-based aspects.

5.3.2 Performance of basic policies

Let assume a heterogeneous scenario in which a set of radio access networks are available, and let define R as the domain of corresponding RATs. For instance, and without lack of generality, R can be given by $\{\text{UTRAN}, \text{GERAN}, \text{WLAN}\}$. A basic initial RAT selection policy can be then defined as a function f that, given a set of different inputs $(\xi_1, \xi_2, \dots, \xi_M)$, e.g. service class, load in each RAN, UE features, mobile speed, etc. provides a suitable RAT to be allocated. Mathematically, a policy p can then be expressed as:

$$p=f(\xi_1, \xi_2, \dots, \xi_M) \in R \quad (54)$$

Some sample examples of basic policies for a scenario with GERAN and UTRAN networks and a mix of voice and interactive users (e.g. www browsing) are defined in the following:

- VG (voice GERAN) policy: This policy has only the service type as input, and allocates voice users into GERAN and other services into UTRAN. This is:

$$p_{VG} = f(\text{service}) = \begin{cases} \text{GERAN, if service = voice} \\ \text{UTRAN, if service = www} \end{cases} \quad (55)$$

- VU (voice UTRAN) policy: This policy acts in the opposite direction as VG and allocates voice users to UTRAN and interactive users to GERAN.

$$p_{VU} = f(\text{service}) = \begin{cases} \text{UTRAN, if service = voice} \\ \text{GERAN, if service = www} \end{cases} \quad (56)$$

5.3.2.1 Simulation Environment

The previous initial RAT selection policies are evaluated in a scenario that considers UTRAN and GERAN access technologies with 7 omnidirectional cells for GERAN and 7 for UTRAN. The cells of both RANs are collocated. The cells of both RATs are collocated with 1km distance between sites. In case of GERAN, it is assumed that the 7 cells represent a cluster so that all the cells operate with different carrier frequencies. The parameters of the UE and the UTRAN and GERAN cells are summarised in Table 55 and Table 56, respectively. It is assumed that all terminals have multi-mode capabilities, i.e. they can be connected either to UTRAN or to GERAN. Three carriers per cell in the 1800 MHz band are assumed in GERAN and a single UTRAN FDD carrier is considered in UTRAN. In this way, the total bandwidth available in the cluster of seven GERAN cells is approximately the same as the bandwidth used by UTRAN. The urban macrocell propagation model in [130] is considered for both systems, with the path loss as a function of the distance d to the base station given by:

$$L_p(\text{dB}) = 128.1 + 37.6 \log(d(\text{km})) \quad (57)$$

An additional log-normal shadowing with $s=10$ dB standard deviation is considered. The mobility model described in [131] is considered with mobile speed 3 km/h and shadowing decorrelation distance 20 m.

Table 55 UTRAN BS and UE parameters

BS parameters	
Cell type	Omnidirectional
Maximum transmitted power	43 dBm
Thermal noise	-104 dBm
Common Control Channels Power	33 dBm
Maximum DL power per user	41 dBm
UE parameters	
Maximum transmitted power	21 dBm
Minimum transmitted power	-44 dBm
Thermal noise	-100 dBm
DL Orthogonality factor	0.4

Table 56 GERAN BS and UE parameters

BS parameters	
Cell type	Omnidirectional
DL transmitted power	43 dBm
Thermal noise	-117 dBm

Number of carriers	3
EGPRS slots	All the slots except the slot 0 of the first carrier are reversible
UE parameters	
Maximum transmitted power	33 dBm
Minimum transmitted power	0 dBm
Thermal noise	-113 dBm
Multislot class	2 UL, 3 DL, 4 UL+DL

In UTRAN, an iterative power control procedure is considered to simulate the inner loop power control aiming at achieving the target (E_b/N_0) that ensures the required Block Error Rate (BLER), thus determining the transmitted power in the uplink and in the downlink directions as well as the measured (E_b/N_0). The relationship between (E_b/N_0) and BLER is obtained from a link layer simulator that takes into account the characterisation of the physical layer for each Radio Access Bearer in terms of e.g. channel coding, spreading and modulation, transmit diversity as well as the channel impulse response [132]. On the other hand, the loss in orthogonality due to multipath in downlink transmissions of a given base station using OVVSF (Orthogonal Variable Spreading Factor) codes is modeled by means of an orthogonality factor equal to 0.4.

With respect to GERAN, a slow power control is simulated in the uplink, so that the transmitted power is changed in steps of 2 dB every measurement period of 0.48s in order to reach a specific sensitivity level. No power control is simulated in the downlink, and all the channels are transmitted with maximum power. The sensitivity values and link layer characterisation is taken from [133].

A mix of voice and interactive users is considered. Voice calls are generated according to a Poisson process with an average call rate of 10 calls/h/user and exponentially distributed call duration with an average of 180 s. In UTRAN, the Radio Access Bearer (RAB) for voice users is the 12.2 kb/s speech defined in [134], considering a dedicated channel (DCH) with spreading factor 64 in the uplink and 128 in the downlink. In turn, in GERAN, voice users are allocated to a TCH-FS (traffic channel full-rate speech), i.e. one time slot in each frame.

Interactive users follow the www browsing model given in [131], with 5 pages per session, an average reading time between pages of 30s, an average of 25 objects (packets) per page, and interarrival packet time 0.125s for the uplink and 0.0228s for the downlink. The average packet size is 366 bytes. This leads to an average bit rate during activity periods of around 24 kb/s in the uplink and 128 kb/s in the downlink. A session rate of 24 sessions/h/user is assumed.

WWW browsing service is provided in UTRAN by means of dedicated channels (DCH) making use of the transport channel type switching procedure (i.e. the DCH is allocated only during activity periods, e.g. page downloads, while during inactivity periods no dedicated resources are allocated). The considered RAB assumes a maximum bit rate of 64 kb/s in the uplink (corresponding to a minimum spreading factor of 16) and 128 kb/s in the downlink (with a spreading factor of 16). The RAB characteristics are given in [134].

In turn, in GERAN, the www service is provided through a PDCH (Packet Data Channel). Depending on the traffic flow generation, the slot availability and the mobile multi-slot capabilities, Temporary Block Flows (TBFs) are allocated in the uplink and/or downlink following the principles given in [136]. Several TBFs belonging to different users can be allocated simultaneously in the same time slot, and a round robin scheduling algorithm is used to decide which TBF is allowed to transmit in each frame.

On the other hand, a link adaptation mechanism operating in periods of 1s is used to select, for each user, the highest modulation and coding scheme (MCS) that ensures the specific sensitivity requirements [133]. The highest modulation scheme considered here is MCS-7, corresponding to a bit rate of 44.8 kb/s per time slot. Then, assuming that the multislot class allows up to 2 uplink slots and 3 downlink slots (see Table 56), the maximum bit rate is 89.6 kb/s in the uplink and 134.4 kb/s in the downlink. Consequently, in terms of maximum bit rate, similar capabilities are considered for both UTRAN and GERAN.

A summary of the main RRM parameters residing at the local RRM entities in both UTRAN and GERAN is given in Table 57 and Table 58. For a detailed definition of the meaning of the different parameters the reader is referred to [135]- [137].

Table 57 UTRAN RRM parameters

UL admission threshold (η_{\max})	1.0
DL admission threshold (P_{\max})	42 dBm
Measurement time	1s
Active Set size	1
Replacement hysteresis	3 dB
Time to trigger handover	0.64 s
BLER target voice	1%
BLER target interactive	10%
Dropping condition	1 dB below target during 20 s

Table 58 GERAN RRM parameters

Measurement period	0.48s
BS_CV_MAX	15
GPRS_MS_TXPWR_MAX_CCH	43 dBm
GPRS_RESELECT_OFFSET	-2 dB
GPRS_RXLEV_ACCESS_MIN	-105 dBm
Maximum number of TBFs per slot	UL: 8, DL:32
Minimum UL received power to trigger handover (L_RXLEV_UL_H)	-100 dBm
Minimum DL received power to trigger handover (L_RXLEV_DL_H)	-100 dBm
Number of consecutive samples below the L_RXLEV_UL_H or L_RXLEV_DL_H to trigger handover (P5)	3
Dropping condition	5 dB below sensitivity during 20 s or 10 consecutive unsuccessful HHO trials

With respect to the admission control procedure in UTRAN, three conditions are checked [28], namely the uplink load factor should be below the threshold η_{\max} , the downlink transmitted power below P_{\max} and there must be available OVFS codes in the base station. The condition of OVFS codes is not checked for interactive users since they will wait for a DCH in case that there are not OVFS codes available. With respect to GERAN, voice users

are accepted provided that there are available time slots, while interactive users are always accepted at session initiation in idle state (i.e. without a TBF established). Voice users have precedence over www users, so that slots occupied by www users are allocated to incoming voice users in case that there are not other free slots. During a handover procedure for interactive users with an established TBF, it is assumed that the handover is not carried out if the TBF cannot be established in the new cell.

In order to stress the influence of the initial RAT selection procedure, without taking into consideration handover issues, no vertical handover strategy is initially considered in these simulations, so that each mobile keeps the selected RAT throughout all the session lifetime.

Table 59 presents the aggregate throughput (in Mbit/s) achieved with the sum of both RATs (GERAN and UTRAN) and with the sum of both services (voice and www) for the two considered service-based policies (VU and VG). Simulations consider a total of 400 voice users in the scenario together with three different interactive load levels, corresponding to 200, 600 and 1000 www users in the scenario. Additionally, the case of VG without Transport Channel Type Switching mechanisms (i.e. an interactive user keeps a dedicated OVSF code even during page reading times) is also presented. In this respect, it is shown that the throughput is greatly reduced in this later case, so that it is advisable for the operator to take full advantage of the transition to RACH/FACH state if DCH are used for interactive users. With respect to VU and VG policies comparison, there are not substantial differences on the overall achieved throughput for the case of 500 m cell radius. Nevertheless, for the 1 Km radius the VG policy achieves somehow better throughput as long as the shorter coverage range for UTRAN is causing some quality problems on voice users (i.e. increase in the block error rate and eventually droppings) if the offered load is high.

Table 59 Aggregate throughput for the different policies

	VU				VG				VG (no TrCH switch)			
	UL		DL		UL		DL		UL		DL	
	0.5 Km	1 Km	0.5 Km	1 Km	0.5 Km	1 Km						
www users	0.5 Km	1 Km	0.5 Km	1 Km	0.5 Km	1 Km						
200	2.18	2.08	2.22	2.17	2.14	2.14	2.20	2.22	2.03	2.01	2.08	2.07
600	3.01	2.88	3.15	3.09	2.96	2.95	3.16	3.15	2.06	2.05	2.11	2.11
1000	3.80	3.64	4.05	3.96	3.77	3.76	4.08	4.08	2.08	2.05	2.14	2.13

Table 60 Average page delay for www users with the different policies

	VU				VG			
	UL		DL		UL		DL	
	0.5 Km	1 Km						
www users	0.5 Km	1 Km						
200	2.91	3.09	0.74	0.76	2.89	2.88	0.76	0.76
600	2.94	3.15	0.77	0.83	2.90	2.90	0.76	0.76
1000	3.03	3.74	0.99	1.26	2.91	2.93	0.76	0.77

Table 60 shows the average page delay for the case that 400 voice users are in the scenario together with a variable number of web browsing users. The delay is presented for both uplink and downlink and for two different cell radii for each of the two service-based policies. It can be observed that VG (i.e. voice users through GERAN while interactive users through UTRAN) tends to provide lower delays. This is because of the higher efficiency for non-real time traffic transmission in UTRAN achieved in the VG case, since web browsing traffic is supported by means of dedicated channels whereas in VU a packet scheduling algorithm must be implemented in GERAN. It is also worth noting that, for 1 Km cell radii, delay increase in VG compared to 500 m cell radii is almost negligible. On the contrary, more noticeable page delay increase is found in VU. This is because in VU (i.e. web supported by GERAN), the link adaptation mechanisms forces to use modulation and coding schemes with lower associated transmission rates, then increasing the delay. We note that in the case of VG (i.e. web supported by UTRAN), the higher coverage radii causes some increase in the BLER beyond the target value of 10%, which causes some moderate delay increase due to increase in packet retransmissions.

5.3.3 Radio network considerations

Another type of basic initial RAT selection policies are those that take into account radio network considerations of each specific RAT. In particular, under the observation that WCDMA capacity is highly degraded by indoor traffic users, as stated in [138], where capacity reductions of up to 80% are observed when half of the users in a scenario are indoor, the following basic policy is considered.

- IN (indoor) policy: In this case the selection would be done taking into account whether a user is indoor or outdoor, so that indoor users would be allocated in GERAN. Then:

$$p_{IN} = f(\text{indoor_user}) = \begin{cases} \text{GERAN, if indoor_user} = \text{true} \\ \text{UTRAN, if indoor_user} = \text{false} \end{cases} \quad (58)$$

It is worth mentioning that it is assumed that for this policy it is possible to know whether a user is located indoor or outdoor, just in order to provide an initial estimation of the potential performance improvement if this information was known. In that sense, in chapter 5.5 a more practical approach to this policy, based on path loss measurements, will be given in scenarios with indoor users.

Figure 212 and Figure 213 consider a scenario where 30% of the users are indoor and the distance between sites is 1 km. In order to see clearly the effects of the IN policy (i.e. allocate indoor users to GERAN) only voice traffic is considered, and the IN policy is compared with a reference random policy (RN) in which users are allocated randomly with equal probability in GERAN and in UTRAN. The results are presented in terms of the block error rate (BLER) in the uplink direction for both UTRAN and GERAN systems. It can be observed that, when the IN policy is applied, the BLER is reduced in UTRAN. On the other hand, in GERAN an increase in the BLER is observed because there are more indoor users than with the RN policy. Nevertheless, the BLER improvement experienced in UTRAN is significantly higher than the degradation in GERAN, which suggests the suitability of using IN policy in the presence of indoor users.

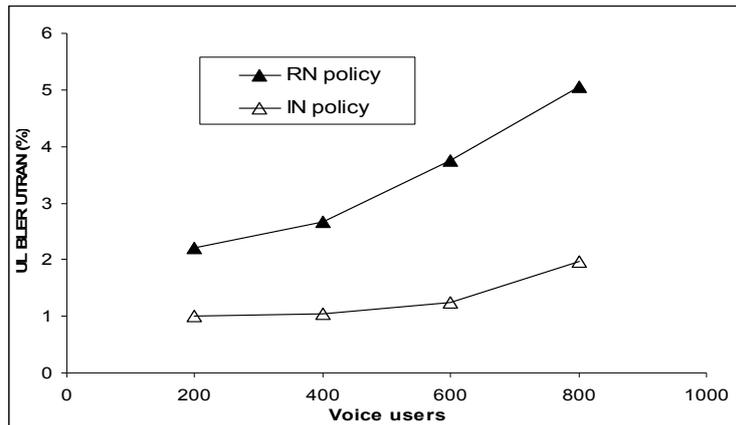


Figure 212. UL BLER in UTRAN for the IN and RN policies.

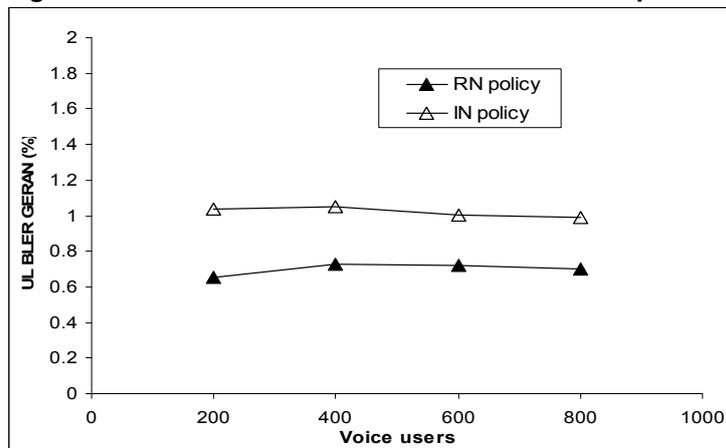


Figure 213. UL BLER in GERAN for the IN and RN policies.

5.3.3.1 Combination of basic policies: n-complex policies

Notice that the strict application of the basic policies described in the previous sub-section would mean that, if there is no capacity available in the selected RAT, the service request is blocked because, otherwise, the policy would be violated. Consequently, these basic policies could lead to blockings even if there is capacity available in other RATs. This undesirable effect can be avoided by defining complex policies in which the output is a prioritised list of RATs, so that if no capacity is available in the first one, the second one would be selected, and so on. In this case, the service would be blocked if there is no capacity in any of the listed RATs. Mathematically, let define an n-complex policy as a function:

$$p=f(\xi_1, \xi_2, \dots, \xi_M) \in R^n \tag{59}$$

leading to a list of n RATs. Notice that the combination of basic policies leads to n-complex policies. In that sense, it is defined the combination of two basic policies $p_i * p_j$ as a list of two RATs, the first one according to p_i (corresponding to the first choice) and the second according to p_j (corresponding to the second choice if there is not capacity in the first RAT). Furthermore, another advantage of n-complex policies is that they are able to combine both service-based aspects with radio-network considerations.

Some examples of 2-complex policies constituted by the previous basic policies are presented in the following:

$$VG * IN=f(\text{service}, \text{indoor_user}) \tag{60}$$

service	indoor_user	$VG*IN \in R^2$
voice	true	GERAN, GERAN
voice	false	GERAN, UTRAN
www	true	UTRAN, GERAN
www	false	UTRAN, UTRAN

As an example of this policy, if the service is voice and the user is outdoor, the first choice will be to allocate it in GERAN (i.e. according to the VG policy). If no capacity is available in GERAN, the second choice will be to allocate it in UTRAN (i.e. according to the IN policy). Notice also that, if the service is www and the user is outdoor, a blocking will occur if there is not capacity in UTRAN because, otherwise, both VG and IN policies would be violated. The same occurs if the service is voice and the user is indoor and there is not capacity in GERAN.

$$IN*VG=f(\text{service}, \text{indoor_user}) \quad (61)$$

service	indoor_user	$IN*VG \in R^2$
voice	true	GERAN, GERAN
voice	false	UTRAN, GERAN
www	true	GERAN, UTRAN
www	false	UTRAN, UTRAN

In this case, the first choice takes into account whether the user is indoor or outdoor and, if no capacity is available in the selected RAT, the second choice considers the service type.

$$VG*VU=f(\text{service}) \quad (62)$$

service	$VG*VU \in R^2$
voice	GERAN, UTRAN
www	UTRAN, GERAN

According to this policy, voice users will first fill the capacity available in GERAN and then they will be directed to UTRAN. In turn, www users will first fill the capacity in UTRAN and then they will be directed to GERAN. In this case, no request is blocked provided that there is capacity available in either UTRAN or GERAN.

In realistic scenarios, where different types of services are used by customers located either indoor or outdoor, the use of basic policies like VG or IN may not sufficiently capture the required features to do a proper initial RAT selection and therefore, n-complex policies should be applied. For such a scenario, Table 61 and Table 62 present the aggregated throughput for different numbers of voice and www users and when there are 10% and 50% of indoor users, respectively. Results are presented for both the uplink and downlink directions. The 2-complex policies $VG*IN$, $IN*VG$ and $VG*VU$ are compared.

Table 61 Total throughput (Mb/s) with 10% indoor users

10% indoor		VG*IN		IN*VG		VG*VU	
Voice users	www users	UL	DL	UL	DL	UL	DL
200	200	1.38	1.44	1.37	1.43	1.37	1.43
	600	2.21	2.40	2.15	2.32	2.18	2.36
	1000	2.99	3.30	2.94	3.23	2.98	3.30
400	200	2.14	2.22	2.30	2.34	2.16	2.23
	600	2.96	3.15	2.94	3.10	2.95	3.14
	1000	3.79	4.10	3.58	3.81	3.78	4.10
600	200	2.59	2.65	3.13	3.11	2.76	2.86
	600	3.41	3.59	3.63	3.69	3.51	3.73
	800	3.80	4.04	3.83	3.93	3.91	4.19

Table 62 Total throughput (Mb/s) with 50% indoor users

50% indoor		VG*IN		IN*VG		VG*VU	
Voice users	www users	UL	DL	UL	DL	UL	DL
200	200	1.37	1.45	1.33	1.39	1.37	1.45
	600	2.14	2.34	2.02	2.21	2.16	2.36
	1000	2.94	3.27	2.71	3.00	2.93	3.26
400	200	2.12	2.19	2.24	2.31	2.14	2.21
	600	2.88	3.09	2.95	3.13	2.93	3.14
	1000	3.68	4.02	3.62	3.92	3.71	4.06
600	200	2.38	2.47	3.21	3.28	2.76	2.86
	600	3.18	3.38	3.78	3.94	3.51	3.74
	800	3.60	3.86	4.06	4.29	3.90	4.19

Up to medium voice loads (i.e. 200 users) no relevant differences between the policies are observed, although in general the performance of IN*VG is somewhat poorer, mainly when the number of www users increases. The reason is that, with IN*VG, there is a higher number of www users that are served through GERAN (i.e. those that are indoor), which provides higher delays and lower www throughput than UTRAN (see Figure 214 and Figure 215). Further, when the ratio of indoor users increases, the number of interactive users allocated in GERAN also increases and, consequently, IN*VG performance is more degraded.

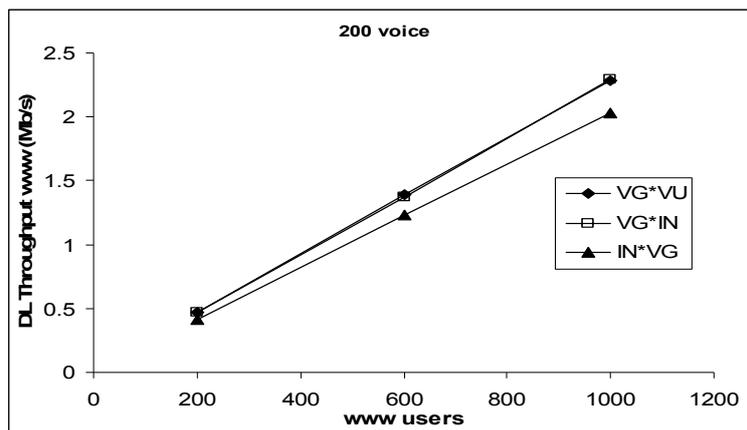


Figure 214 Total DL www throughput with 50% indoor traffic.

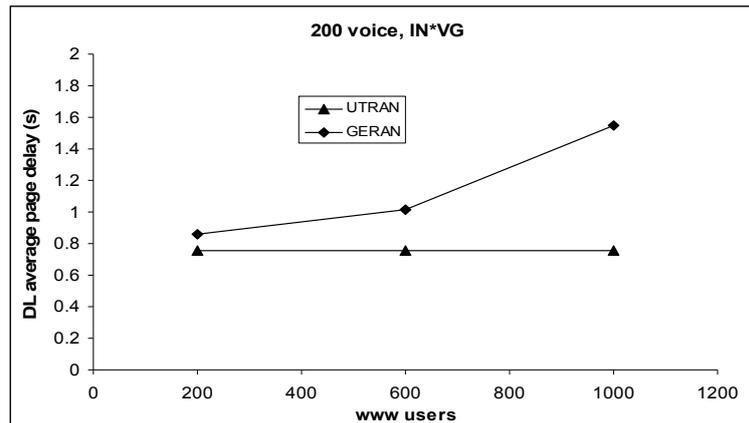


Figure 215 DL average page delay with IN*VG policy and 50% indoor traffic.

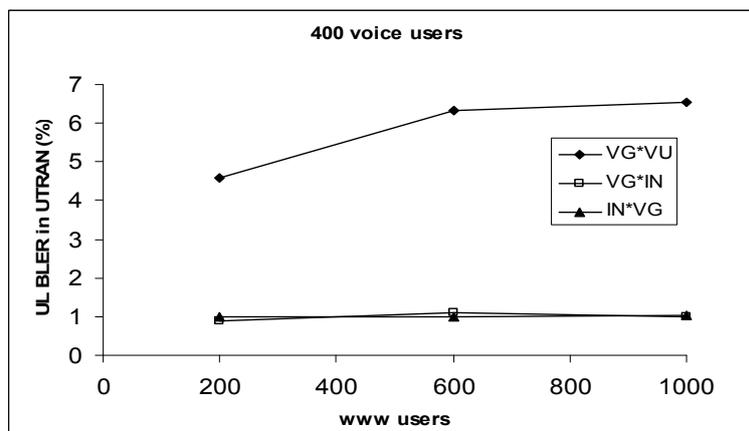


Figure 216 UL BLER in UTRAN for voice users with 50% indoor traffic.

Nevertheless, when the voice load increases (i.e. 400 and 600 voice users), policy IN*VG achieves a higher throughput than the other two policies. The improvement is more significant for low number of www users because of the throughput reduction when www users are served through GERAN. It is important to note that with IN*VG the load is more distributed between both access networks. For example, for 400 voice users and 50% indoor, with IN*VG 50% of the voice traffic goes through GERAN while with VG*IN the ratio is approximately 99%, thus existing a higher occupation in GERAN with VG*IN that can originate some voice droppings. Furthermore, Figure 216 shows the improvement in terms of BLER when the policies that take into account the indoor condition (i.e. IN*VG and VG*IN) are compared with the policy VG*VU.

Finally, with respect to the comparison between VG*VU and VG*IN, notice that, in general, both have similar performance although for high loads VG*VU uses to have a higher throughput because of the low number of blockings (i.e. notice that in VG*VU a blocking only occurs when there is no capacity in neither UTRAN nor GERAN, while with VG*IN this is no longer true).

5.3.4 Vertical Handover

After the initial RAT selection decision, taken at session initiation, vertical handover is the first step to carry out interoperation among access networks in heterogeneous systems. The successful execution of a seamless and fast vertical handover is essential for hiding to the user the underlying enabling infrastructure. Issues related to vertical handover comprise scanning procedures for the terminal to discover available RATs, measurement mechanisms

to capture the status of the air interface in the different RATs, vertical handover triggers (i.e. the events occurring in the heterogeneous network scenario that require the system to consider whether a vertical handover is actually required or not), vertical handover algorithm (i.e. the criteria used to decide whether a vertical handover is to be performed or not) and protocol and architectural aspects to support handover execution.

Vertical handover procedures from one RAT to another may be useful to support a variety of objectives, such as avoiding disconnections due to lack of coverage in the current RAT, blocking due to overload in the current RAT, possible improvement of QoS by changing the RAT, support of user's and operator's preferences in terms of RATs usage or load balancing among RATs. Thus, the vertical handover procedure enables another dimension into the CRRM problem and provides an additional degree of freedom in rearranging traffic, which is eventually exploited by means of the specific vertical handover algorithm. In this respect, there would be many possibilities to follow both in the decision of changing the initially selected RAT (e.g. when a session arrives to a saturated cell in its current RAT and RAT change is not possible due to incompatible services or technology, the vertical handover algorithm may decide to handover other more flexible sessions instead) and in the return policy (i.e. deciding the suitable instant in which sessions that are not in the initially selected RAT must return to it). The trade-off arising here is between flexibility in CRRM and signalling overhead, as it is further discussed in the next sub-section.

The research community has already identified the importance of the vertical handover mechanism in future mobile scenarios and, consequently, a number of contributions have appeared in the recent years [125][139]-[141]. Specifically, in [125] the benefits of Common Radio Resource Management (CRRM) by carrying out load-based vertical handovers in order to balance the load of the different cells and RATs are analysed. Similarly, in [139] the advantages of distributing the load among different networks to increase flexibility and reduce network equipment costs are addressed, and in [140] different policies to overflow sessions that arrive to saturated RANs are discussed. Finally, in [141] different procedures for making measurements of different RATs are discussed as a means to provide support for vertical handover decisions.

5.3.4.1 Loose and Tight Interworking between Vertical and Horizontal Handover

This section introduces two different degrees of interworking between the vertical and the horizontal handover algorithms. They are referred to as loose interworking, denoted here as L-VHO, and tight interworking, denoted here as T-VHO. These terms intend to describe how often the suitability to carry out a vertical handover is considered for each connection. In particular, L-VHO stands for the case that the vertical handover algorithm is executed only when a horizontal handover fails or when a call dropping is about to occur due to bad propagation conditions, so that it is seen simply as a mechanism to avoid call droppings. In this case, after having executed a vertical handover the session will remain in the new RAT as long as the propagation conditions are satisfactory and there are available resources in the subsequent horizontal handovers. In turn, T-VHO stands for the case that the vertical handover decision algorithm is executed at every instant that the horizontal handover algorithm is considered, so that both possibilities are considered prior to taking a decision. In the VHO decision algorithm the RAT selection policy is evaluated for the specific connection, as explained in section V. Figure 217 reflects the interactions between vertical and horizontal handover functions, indicating that horizontal handover algorithm is an inherent part of the RRM entity while vertical handover algorithm belongs to the CRRM entity. Similarly, Figure 218 summarises in the form of a flow diagram the L-VHO and T-VHO alternatives. We note that e.g. a periodic triggering of the vertical handover algorithm would also be possible, though not considered here.

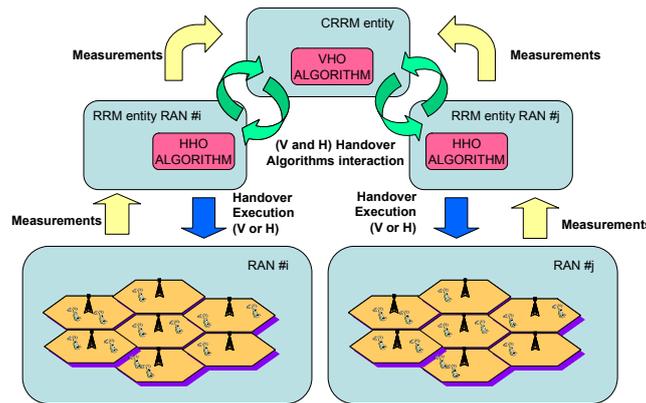


Figure 217 Interactions between VHO and HHO.

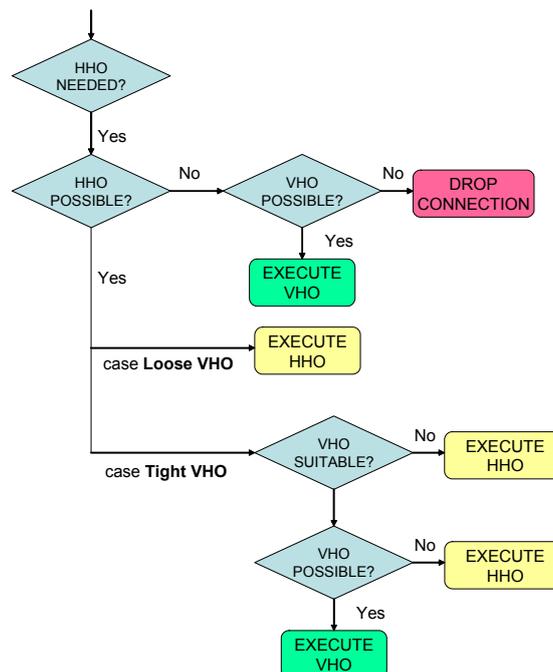


Figure 218 Flow diagram of the loose and tight interworking approaches between horizontal and vertical handover

There are two main aspects that need to be considered when comparing T-VHO with L-VHO. Firstly, CRRM entity implementation: if the CRRM entity is implemented in every existing RNC/BSC, then HHO and VHO are tightly coupled in a natural way and the interaction is simply an internal matter of the RNC/BSC. On the contrary, if the CRRM entity is implemented only in some RNC/BSC [121], then delay in taking decisions plays a role and tends to impact more on T-VHO (see Figure 219) because of the signalling exchange required between the nodes where the CRRM and RRM entities reside. Notice that, depending on how these nodes are interconnected, this signalling exchange may involve other network elements. In turn, if the CRRM entity is implemented in a separate node of the network (see Figure 220), then delay in taking decisions poses further constraints impacting even more on T-VHO, because in this case a signalling exchange with the CRRM entity is always required before a HHO. Secondly, the achieved performance will be a result of the RAT selection policy, as described in the next section. Then, as long as T-VHO and L-VHO imply different rates at which a vertical handover is considered, there will also be different chances to be able to follow the established policy (e.g. a service policy here) and, consequently, different performances can be expected. Consequently, a trade-off may arise.

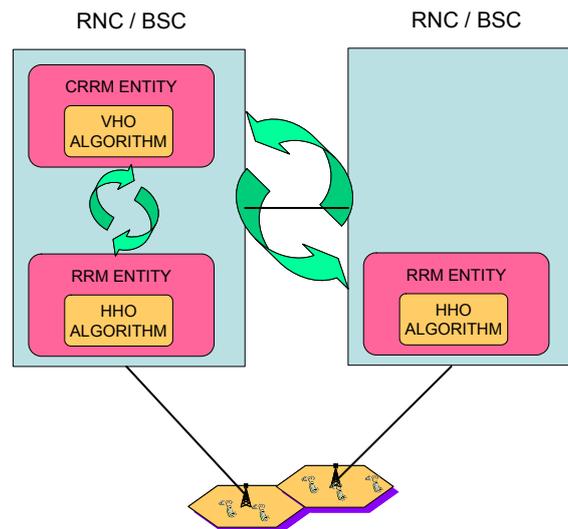


Figure 219 Interactions between horizontal and vertical handover when CRRM entities reside in existing RNC/BSC nodes

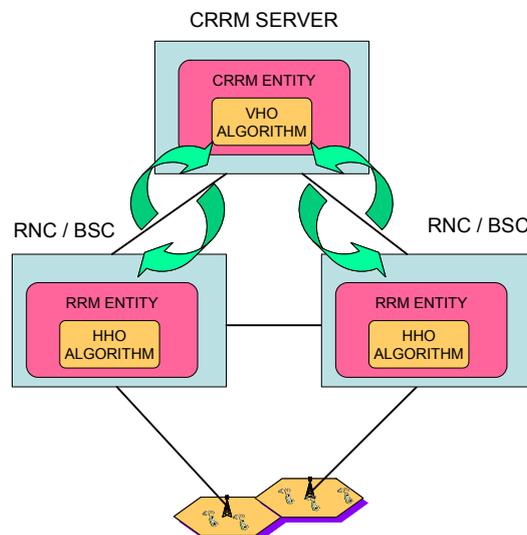


Figure 220 Interactions between horizontal and vertical handover when CRRM entity resides in an external node

With respect to the RAT selection criterion used either at session initiation or in the VHO decision algorithm, the 2-complex service-based policy VG*VU described in sub-section 5.3.3.1 is considered here, although the L-VHO and T-VHO approaches could be used also with other policies taking into account e.g. load balancing principles, radio propagation conditions, etc. Specifically, in the VG*VU policy voice traffic is directed to GERAN if capacity is available while interactive traffic is directed to UTRAN. If capacity is not available in the preferred RAT (i.e. if the admission control in this RAT rejects the connection), the other RAT is selected instead. If no capacity is neither available in the alternative RAT, the call is blocked. The rationale for this service-based policy relays in the better ability of UTRAN to handle interactive users by allocating them on dedicated channels than in case of GERAN, where interactive traffic shares the available packet data channels.

Notice that, when considering this service-based policy in combination with the T-VHO approach (see Figure 218), voice users allocated in UTRAN will try to execute a VHO to GERAN each time a HHO in UTRAN is decided while interactive users allocated in GERAN will try to execute a VHO to UTRAN on each GERAN HHO.

The considered interworking approaches have been evaluated by means of system level simulations with the same simulation conditions described in sub-section 5.3.2. The two considered possibilities, T-VHO and L-VHO are considered and, for comparison purposes, the situation in which no vertical handover is available (i.e. the users are kept in the RAT selected at the beginning of each session) is simulated.

The performance comparison for the different strategies is made according to the following performance metrics:

- Total aggregated throughput: It corresponds to the total number of successfully transmitted bits in both RATs from both www and voice users divided by the simulation time.
- Block Error Rate of voice users: It corresponds to the total number of erroneous blocks (i.e. GERAN radio blocks or UTRAN transport blocks) with respect to the total number of transmitted blocks.
- Average www packet delay: It corresponds to the average time between the generation of a www packet and its complete transmission through the air interface, including retransmissions.
- Dropping probability: A session can be dropped due to bad channel conditions during a certain period of time, in any of the two RATs, as indicated in Table 57 and Table 58. This will normally occur when the user enters a new cell and is not allowed to handover because of overload. The dropping probability is then measured as the fraction between the number of dropped sessions with respect to the total number of sessions.
- Average number of vertical handovers per call or session, either from UTRAN to GERAN or from GERAN to UTRAN.

In the following, the performance obtained with the two vertical handover approaches under different conditions regarding load, traffic mix and scenario parameters is analysed. In particular, for a better discrimination of the different effects, the analysis starts with the improvement that is achieved by means of the inclusion of the vertical handover procedure when compared to the case without vertical handover. Afterwards, the loose and the tight interworking approaches are compared. Finally, the effect of the cell radius and the propagation conditions over the performance of the two approaches is analysed.

5.3.4.1.1 Flexibility provided by vertical handovers

Vertical handover clearly extends the degrees of freedom in the management of the radio resources, so that it is a suitable mechanism to avoid call droppings that would occur within a single RAT due to a mobile arriving to a blocked cell. This is reflected in Figure 221, where the dropping probability for voice users is plotted against an increase in the number of web users with 400 voice users in the scenario, showing that when no vertical handover is available a large percentage of calls get dropped. It should be mentioned that for this voice traffic load of 400 voice users the average occupation in GERAN is around 96%, meaning that it is very likely that a user reaches a blocked cell. As a result of these call droppings, the total aggregated throughput is reduced, as reflected in Figure 222. In turn, when the vertical handover strategy is considered, either with the loose or the tight approach, the dropping is reduced and the total throughput increases, because voice users that reach a blocked GERAN cell are transferred to UTRAN. It is worth mentioning that the throughput results are presented for the uplink but similar trends are observed in the downlink direction as well.

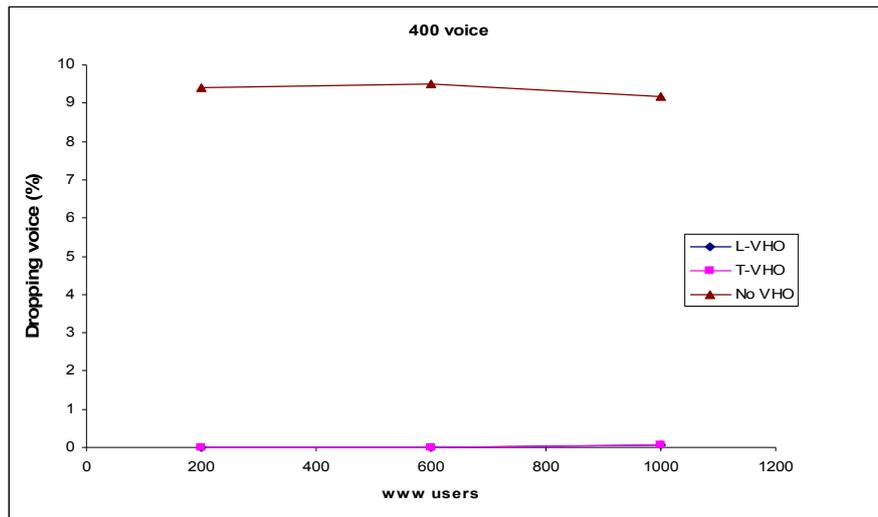


Figure 221 Voice dropping probability

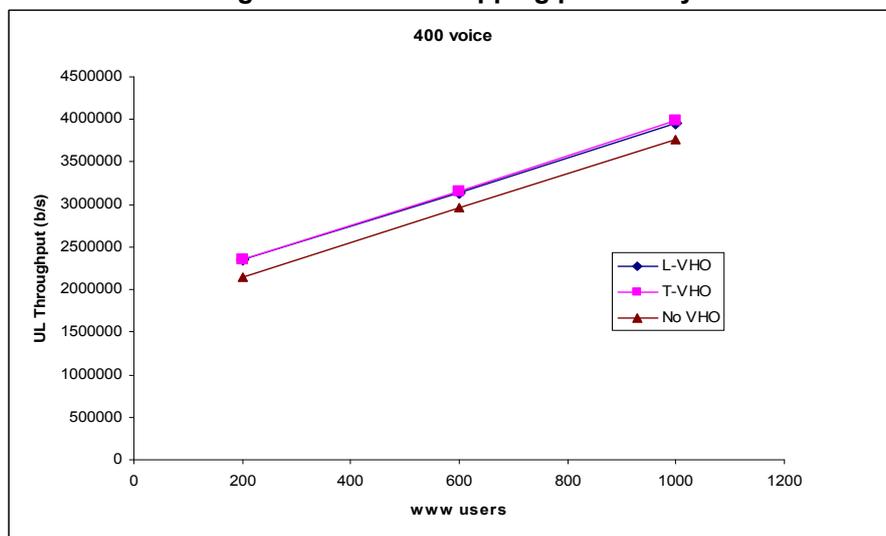


Figure 222 Total uplink aggregated throughput

5.3.4.1.2 Tight versus loose interworking approaches

When comparing T-VHO with L-VHO, the first aspect to consider is how the traffic distribution changes in the different RATs. In particular, when 400 voice users are present in the scenario, Figure 223 shows the percentage of voice traffic served through UTRAN as a function of the number of www users. In this case, even when no vertical handover is available, a certain portion of the voice traffic is served through UTRAN, corresponding to the new voice calls that are originated in a blocked GERAN cell (for this voice traffic load, the probability that this occurs is found to be around 0.5%). In turn, when vertical handover is used, a certain number of voice users are also transferred from GERAN to UTRAN when they reach blocked cells during horizontal handovers, thus increasing the fraction of voice traffic served through UTRAN. Notice that with the tight approach there are less voice users in UTRAN than with the loose approach, because in the latter users that are handed over to UTRAN will tend to remain there, while with the tight approach a voice user in UTRAN will try to return to GERAN in the first UTRAN horizontal handover. In turn, Figure 224 reflects the same statistic but for a much higher load level of 800 voice users. In this case, even without vertical handover there is a high portion of voice traffic served through UTRAN, because the blocking probability in GERAN for new originated voice calls is found to be around 30%. In this situation, the differences between the tight and the loose approach are smaller because

even with the loose approach voice users served through UTRAN have few chances to return to GERAN, which is blocked most of the time (in this case, the average GERAN occupation is around 99%).

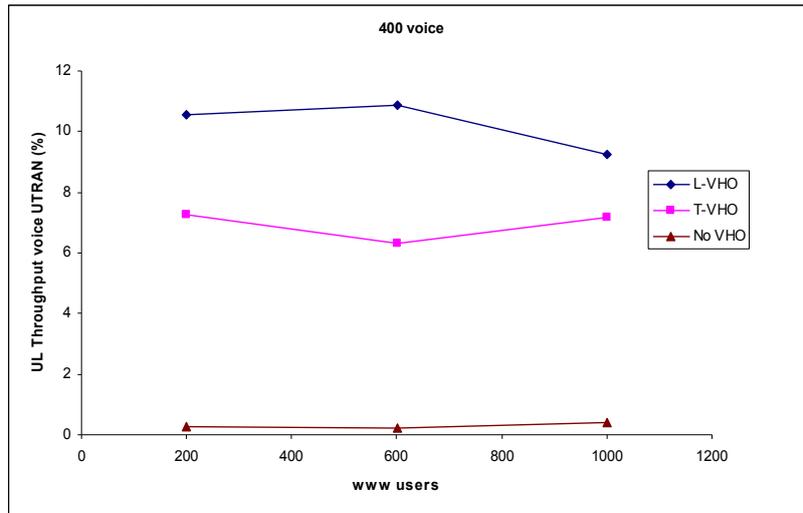


Figure 223 Percentage of voice throughput served through UTRAN for the 400 voice users case

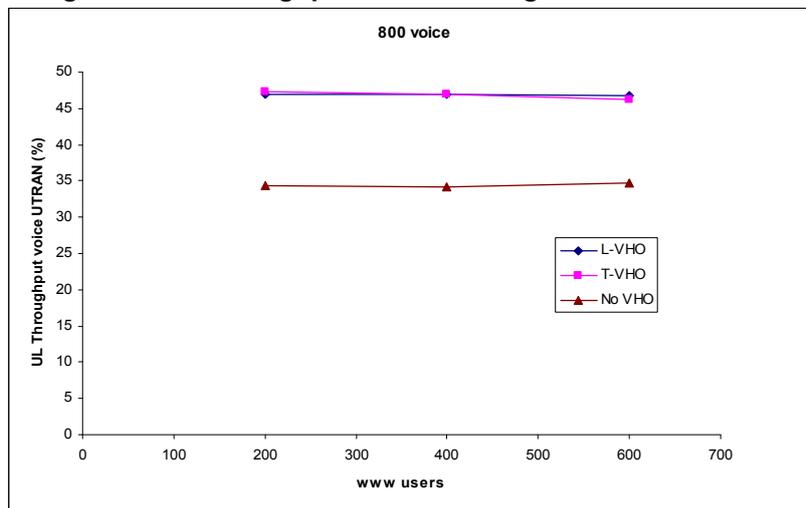


Figure 224 Percentage of voice throughput served through UTRAN for the 800 voice users case

The 800 voice users case corresponds to a situation where not only GERAN is most of the time saturated but even UTRAN perceives a high load, as reflected in Figure 225, which shows the uplink cell load factor for the central UTRAN cell. Notice that, when no vertical handover is used, due to the lower number of voice calls that are served through UTRAN, the load is also smaller. In turn, with the two vertical handover approaches, the uplink load factor level reaches values close to 1, which causes that several www users cannot be admitted in UTRAN and therefore are directed to GERAN at session initiation. This is reflected in Figure 226, where the percentage of www throughput served through UTRAN is presented. Clearly, when no vertical handover is used, almost all the traffic is served through UTRAN. In turn, with the tight approach, also the amount of traffic served through UTRAN is more than 98%, because www users that have initiated session in GERAN will tend to return to UTRAN. On the contrary, with the loose approach www users in GERAN will tend to remain there, thus reducing the amount of www traffic that exists in UTRAN.

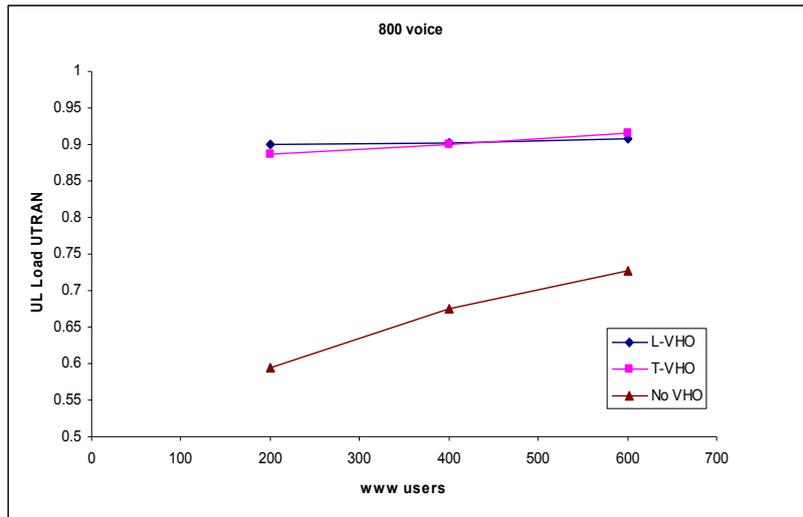


Figure 225 Average uplink load factor in UTRAN

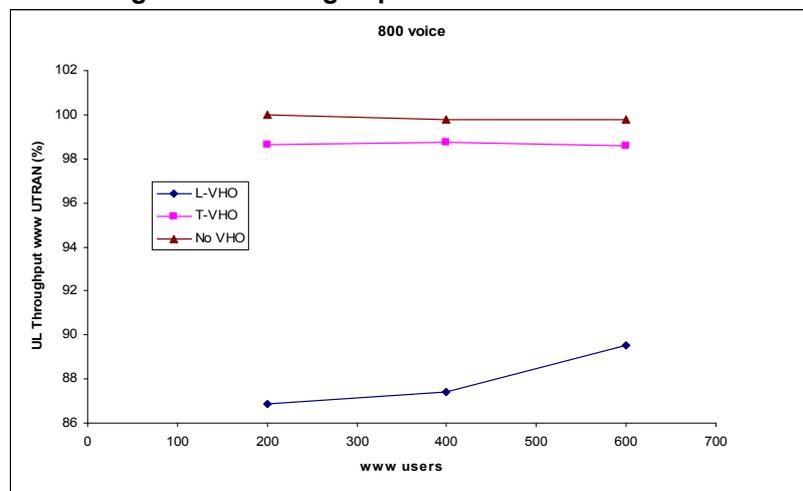


Figure 226 Percentage of www throughput served through UTRAN for the 800 voice users case

As reflected in Figure 223 to Figure 226, the tight vertical handover approach facilitates the fulfilment of the service policy described in section V, serving the www traffic as much as possible through UTRAN. This has an important impact in terms of weighted average delay for www traffic, as reflected in Figure 227 for the 400 voice users case and to a much greater extent in Figure 228 for the 800 voice users case. The differences here arise from the fact that the www delay in GERAN is significantly much higher than the delay in UTRAN, because the www users in GERAN share the few slots remaining after having allocated all the voice users. On the contrary, UTRAN has more room to allocate dedicated channels to www users and therefore the average delay is highly reduced. Then, in terms of the weighted delay that considers both UTRAN and GERAN contributions, the strategies that serve most of the www traffic through UTRAN (i.e. T-VHO and no vertical handover) achieve a lower delay than L-VHO, in which several www users are served through GERAN. Furthermore, when increasing the voice traffic load up to 800 voice users, the differences between L-VHO and T-VHO become higher, because www users in GERAN must share an even lower number of time slots. For comparison purposes, Figure 229 shows the weighted delay in the downlink direction, revealing that similar trends are observed.

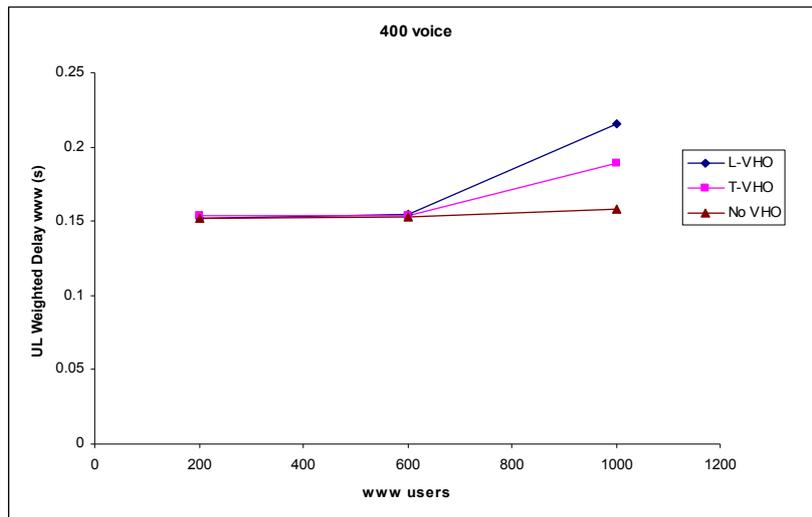


Figure 227 Uplink weighted delay for www traffic with 400 voice users

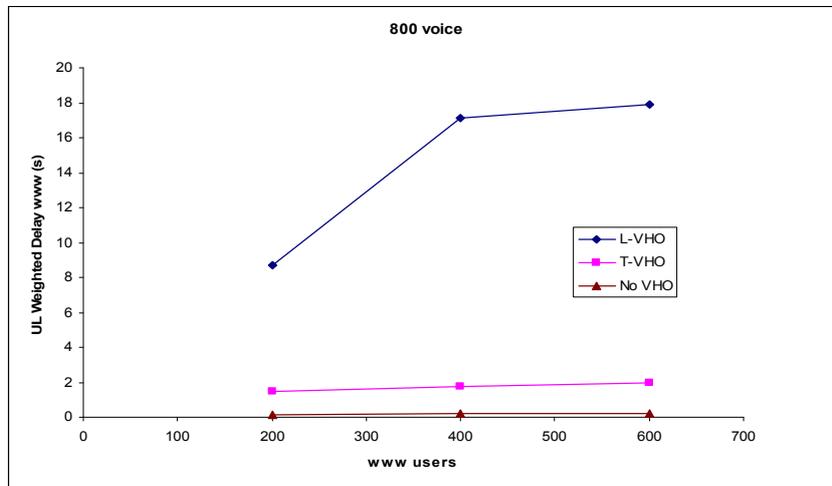


Figure 228 Uplink weighted delay for www traffic with 800 voice users

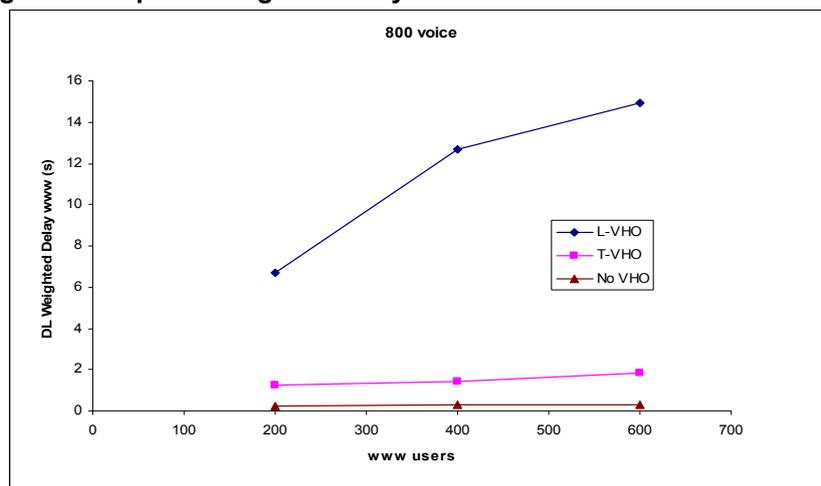


Figure 229 Downlink weighted delay for www traffic with 800 voice users

On the other hand, from the point of view of signalling, T-VHO implies much more frequent vertical handover procedures, then increasing signalling load. This is shown in Figure 230, which plots the average number of vertical handovers from UTRAN to GERAN per voice call. Clearly, T-VHO provides more chances to follow the service policy established for RAT selection, reflected in a higher number of vertical handovers to GERAN than in the loose case, in which the voice users that enter UTRAN tend to remain there until their call ends.

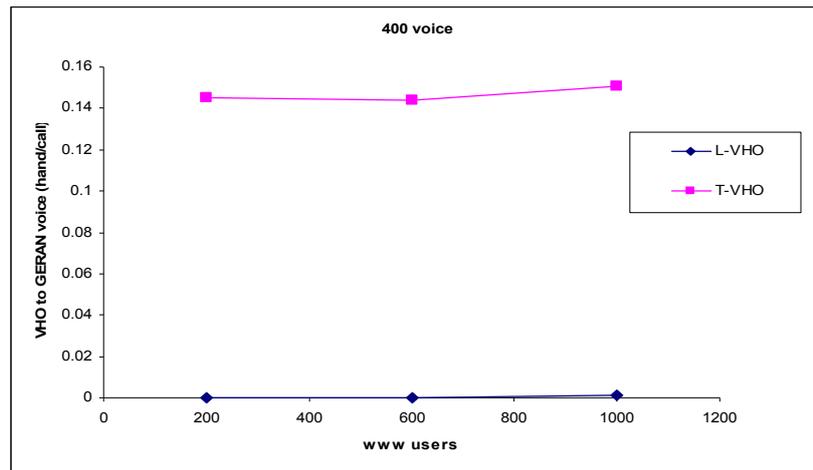


Figure 230 Average number of vertical handovers from UTRAN to GERAN per voice call

5.3.4.1.3 Influence of the cell radius

The influence of cell radius on the achieved performance in a heterogeneous scenario varies depending on the considered RAT. For example, it is discussed in [142] that UTRAN exhibits higher sensitivity to cell radius extension than GERAN and performance is degraded faster in the former RAT for larger cell radii, because of the interference-limited nature of the WCDMA access existing in UTRAN. As an illustration, Figure 231 exhibit the uplink Block Error Rate (BLER) for voice users connected to UTRAN and to GERAN in the T-VHO case with 800 voice users in the scenario, for radius $R=1\text{km}$ and $R=500\text{m}$, respectively. It can be observed that a high degradation in the BLER appears when increasing the cell radius in the UTRAN case, while the degradation is much smaller in GERAN. As a result of this larger degradation in UTRAN, the average weighted delay for www users experiences an increase with the cell radius for both T-VHO and L-VHO approaches due to a higher number of retransmissions, as depicted in Figure 232 for the case with 400 voice users in the scenario. In any case, the differences between T-VHO and L-VHO are similar for the two radii.

With respect to the signalling, Figure 233 presents the average number of vertical handovers from GERAN to UTRAN for voice users with the two approaches and for the two considered radii $R=500\text{m}$ and $R=1\text{km}$. It can be observed how the increase of the cell radius reduces the rate of vertical handovers, exhibiting a higher reduction in the tight approach than in the loose case. The reason is that for higher cell radius the different sessions remain in the same cell during a longer time, thus reducing the rate of horizontal handovers. Consequently, since in the tight approach each horizontal handover triggers the execution of a possible vertical handover, reducing the number of horizontal handovers also reduces the vertical handover rate.

It is worth mentioning that the reduction in the vertical handover rate leads to a worst ability to follow the RAT selection policy (i.e. serve voice users through GERAN and www users through UTRAN), as it is reflected in Figure 234, which presents the average number of www users in UTRAN for the two radii and the two handover approaches. On the one hand, it is observed that the T-VHO serves a higher number of www users through UTRAN than L-VHO and, on the other hand, for $R=1\text{ km}$ the number of www users in UTRAN is smaller than for $R=500\text{ m}$ for both T-VHO and L-VHO, because a www user that enters in GERAN will remain there for a longer time before a vertical handover decision is taken.

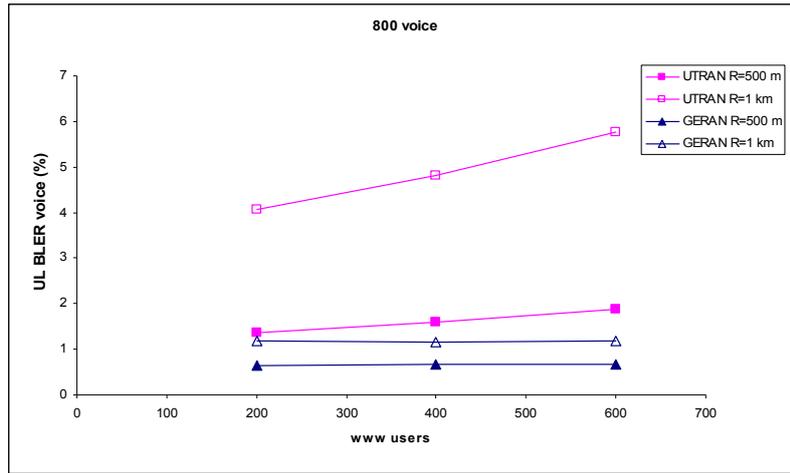


Figure 231 Uplink Block Error Rate of voice users in UTRAN and GERAN with R=1km and R=500m for the T-VHO approach

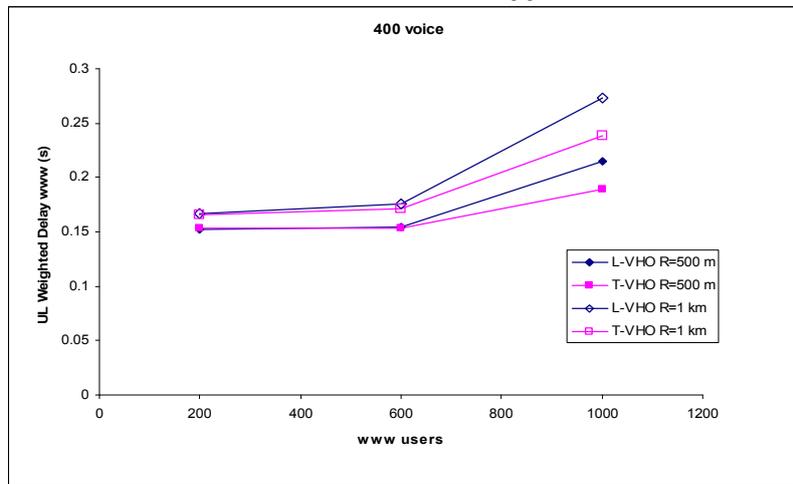


Figure 232 UL weighted delay for web traffic in the cases of R=1 km and R=500 m

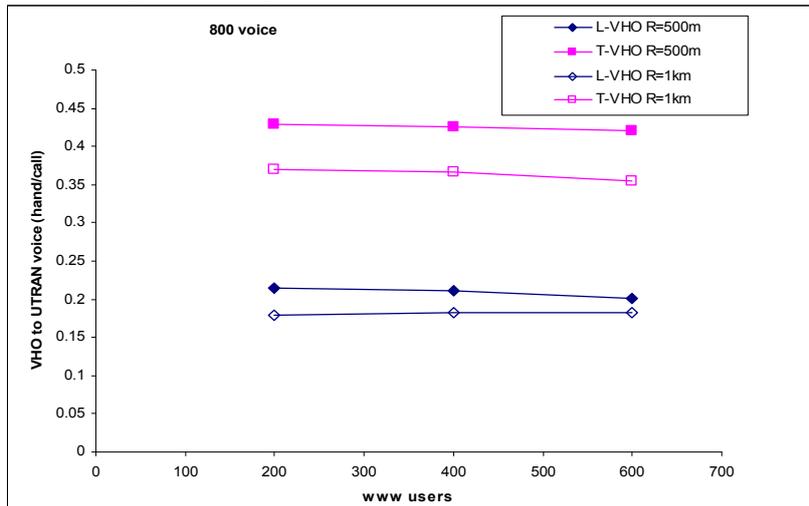


Figure 233 Average number of vertical handovers to UTRAN per voice call for R=500m and R=1km

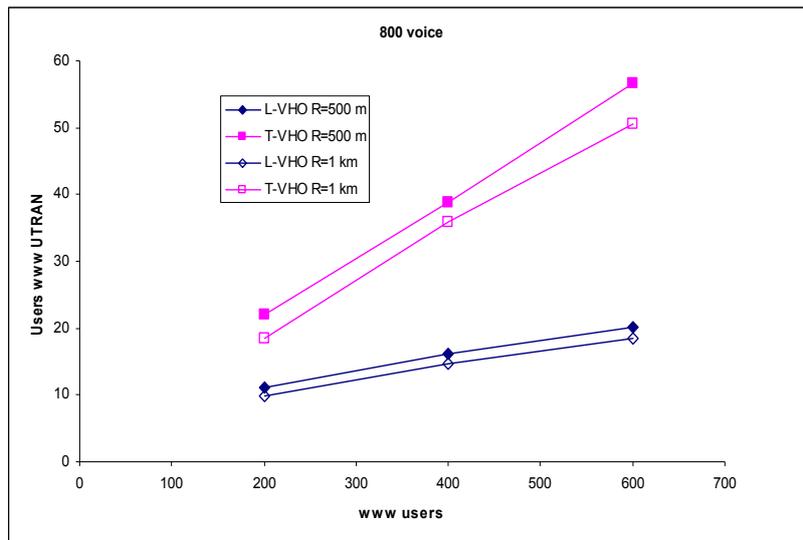


Figure 234 Average number of www users in UTRAN for T-VHO and L-VHO with R=500m and R=1km

5.3.4.1.4 Influence of the propagation conditions

A similar effect to the variation of the cell radius appears when the propagation conditions, and particularly the deviation of the shadowing s , are changed. Specifically, from the point of view of signalling, Figure 235 plots the average number of vertical handovers per call with $s=3\text{dB}$ and $s=10\text{dB}$ for cell radius $R=500\text{m}$. If the tight interworking approach T-VHO is used, in which the possibility of a vertical handover is considered always before a horizontal handover, higher shadowing deviations originate a higher number of vertical handovers, because in this case the number of horizontal handovers increases. In contrast, with the L-VHO approach, the vertical handovers are not directly related to horizontal handovers but to the occurrence of droppings, which are more sensitive to the traffic load or the cell radius than to the shadowing deviation. As a result of that, not very significant differences are observed for the L-VHO case in terms of the number of vertical handovers.

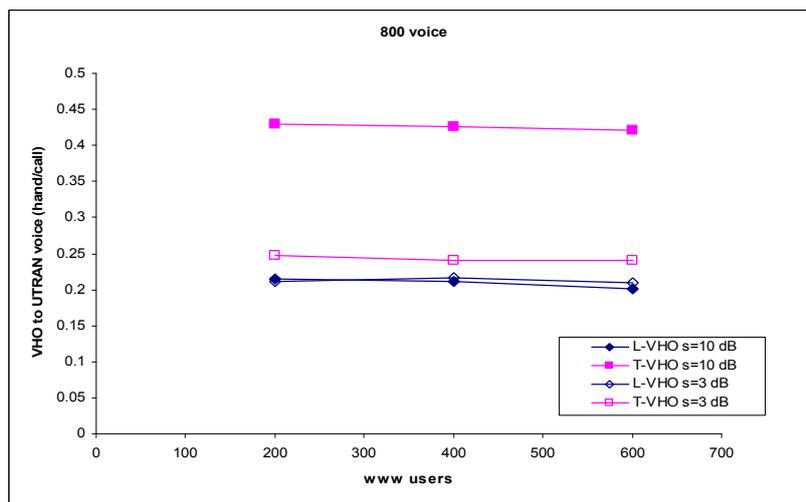


Figure 235 Average number of vertical handovers from GERAN to UTRAN per call for voice users with shadowing deviation 3 dB and 10 dB.

As a result of the above, the number of www users served through UTRAN experiences a reduction with the T-VHO approach for the case $s=3\text{dB}$, as depicted in Figure 236, because www users that enter in GERAN remain there for a longer time. Also in this case the number

of voice users in UTRAN is higher for $s=3\text{dB}$, as reflected in Figure 237, because voice users that enter in UTRAN remain there longer before returning to GERAN.

Consequently, a reduction in the shadowing deviation turns into a worst ability to follow the service policy for the T-VHO approach, which finally leads to having smaller relative differences in terms of performance between L-VHO and T-VHO, as illustrated in Figure 238 for the average weighted www delay.

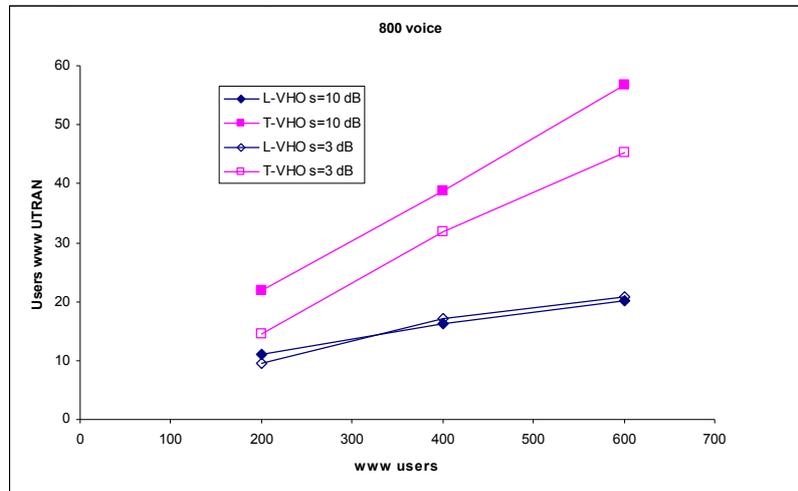


Figure 236 Average number of www users in UTRAN for $R=500\text{ m}$ with shadowing deviations 3 dB and 10 dB.

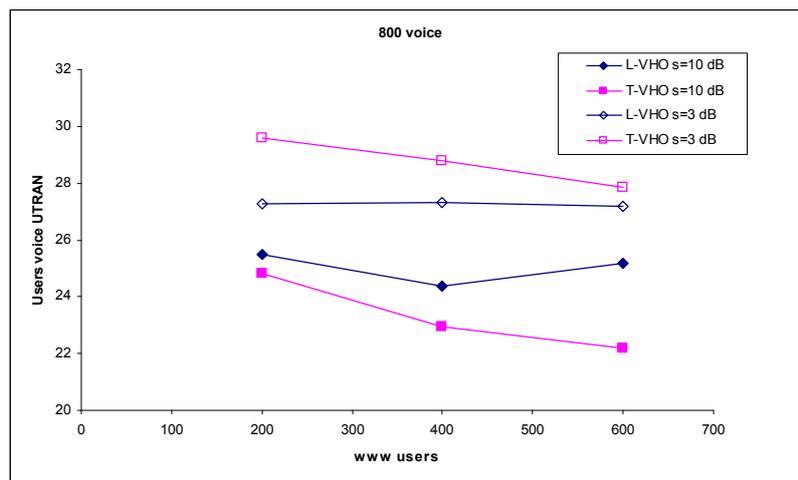


Figure 237 Average number of voice users in UTRAN for $R=500\text{ m}$ with shadowing deviations 3 dB and 10 dB.

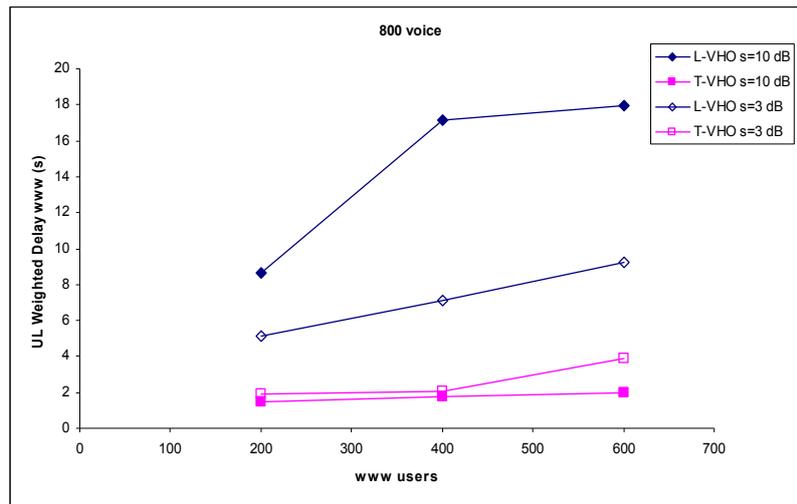


Figure 238 UL weighted delay for www traffic in the case of R=500 m with shadowing deviations 3 dB and 10 dB

5.3.5 Conclusions

This section has focused on the development of RAT selection algorithms for CRRM operation in heterogeneous networks. A general policy-based framework for the specification of such algorithms has been defined and different policies considering the service type as well as the fact that the users may be indoor or outdoor have been evaluated through simulations. It has been obtained that, in outdoor scenarios, VG basic policy turns into a higher throughput than VU. In turn, in scenarios with a mix of indoor and outdoor users with different services, the performance of IN*VG policy improves when the voice load increases, the www load decreases and there is a high fraction of indoor users. On the contrary, for low voice loads and high www loads VG*IN achieves a better throughput. This suggests that the suitable configuration of the RRM and CRRM entities according to specific policies depends on the existing traffic conditions and therefore it may be modified at e.g. different periods of the day.

With respect to vertical handover, the interworking between horizontal and vertical handovers has been studied, with two considered approaches, namely the tight approach T-VHO, in which the vertical handover algorithm is executed at every time that a horizontal handover algorithm should be carried out, so that both possibilities are considered prior to taking a decision, and the loose approach L-VHO, in which the vertical handover algorithm is executed only when a horizontal handover fails or when a call dropping is about to occur due to bad propagation conditions. When compared to the case without vertical handover, it has been observed that both approaches allow a reduction in the number of dropped calls, thanks to the flexibility to transfer sessions between RATs.

With respect to the comparison between T-VHO and L-VHO, it has been shown that the traffic distribution among the considered RATs can be quite different with the two approaches. In that sense, the tight approach allows a better fulfilment of the initial RAT selection policy, due to having more chances to execute a vertical handover than in the loose case. Consequently, and for a service-based RAT selection policy, it has been observed that the tight approach offers a better performance in terms of lower delay for interactive www users than the loose approach because most of the www traffic is served through UTRAN.

On the other hand, a higher number of vertical handover procedures are also required with the tight approach, which increases the signalling overhead.

The impact of the cell radius over the performance of the considered approaches has also been analysed, under the rationale that UTRAN suffers a higher degradation when increasing the cell radius. Under this situation, it has been observed that, on the one hand, the performance in terms of delay is degraded with the two approaches, and, on the other hand, higher cell radius lead to a reduction in the number of vertical handover procedures.

Finally, the influence of the shadowing deviation has been also studied, revealing that for large deviations there is a more significant relative improvement of the tight approach with respect to the loose one, thanks to the higher number of vertical handovers that are executed in the first case, which allow a better fulfilment of the service-based policy.

5.4 LOAD BALANCING - BASED RAT SELECTION

5.4.1 Introduction

Load balancing (LB) is a possible guiding principle for resource allocation in which the RAT selection policy will distribute the load among all resources as evenly as possible. However, in some situations a load balancing policy may not be desirable, at least not as a unique policy to be applied. Indeed, at one stage, an operator may be more interested in allocating users according to a service policy, as explained in section 5.3, e.g. because it increases the revenue, rather than performing a load balancing assignment. Taking this into account, in the following sections we intend to evaluate the performance of a load balancing RAT selection policy, as an alternative to the service-based policies.

The term load balancing appears in the literature in a wide variety of contexts but profusely in the area of distributed computing where, e.g., jobs or tasks are to be assigned to a set of processors [143]. In the context of wireless access networks, load balancing may refer to the allocation of users requesting a given service to a certain cell, carrier frequency, radio access technology, etc. This allocation may be at a call/session establishment, i.e. initial RAT selection, or within an ongoing call (i.e. during vertical handover). Note that, in most cases, this assignment of mobile terminals imposes a more complex set of constraints than the case of assigning tasks to processors due to inherent properties of the wireless link, such as time-variant channel conditions, limited assignment of terminals to cells, RATs, frequencies, etc.

Load balancing algorithms have been considered to improve the performance among cells in single-RAT wireless cellular networks [144]. In this particular case, the algorithm operates when the coverage areas of different base stations overlap. Thus, whenever a mobile station can attach to more than one base station, the new call can be directed to the base station with greatest number of available channels, i.e. the least loaded base station.

For multi-RAT wireless access networks the allocation problem is extended in a way that resources may be assigned in different RATs. Literature has covered this topic in the last years in a scarce number of papers, with special focus on the effects of load balancing in inter-RAT handover procedures.

In particular, in [145], the effect of tuning the load-based handover (HO) thresholds depending on the load of inter-system/inter-layer/inter-frequency cells is studied. In order to avoid unnecessary HOs and HO signalling, a minimum load threshold ensures that no load balancing activities are carried out below that value. However, to reduce the HO attempts and HO failure rates, adjustable thresholds using neighbour load information are suggested and evaluated.

In [146], a force-based load balancing approach is proposed for initial RAT selection and vertical HO decision making. So-called *forces* model the diametric aspects of gain and cost of a decision. This decision is based on the load in the target and the source cell, the QoS difference between the radio links, the time elapsed from the last HO and the HO overhead.

Nevertheless, abovementioned references either compare results obtained in the combined UMTS/GSM systems with the disjoint systems and observe the so-called *trunking gain*, [146], or just consider a single load balancing approach with changes on the algorithm parameters [145]. Therefore, the suitability of applying load balancing techniques in a multi-RAT scenario as opposed to applying other techniques in the same scenario is not addressed.

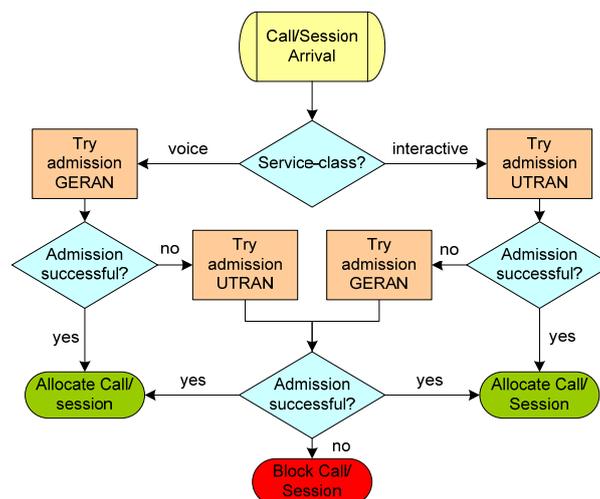
In subsequent sections we deal with the initial RAT selection problem for new incoming users requesting service in either of the available RATs. In order to retain the effect of load balancing in initial RAT selection, in the first place, vertical HO (VHO) is not considered. Later on, the effects of VHO will be considered jointly with the initial RAT selection policies.

5.4.2 Initial RAT selection

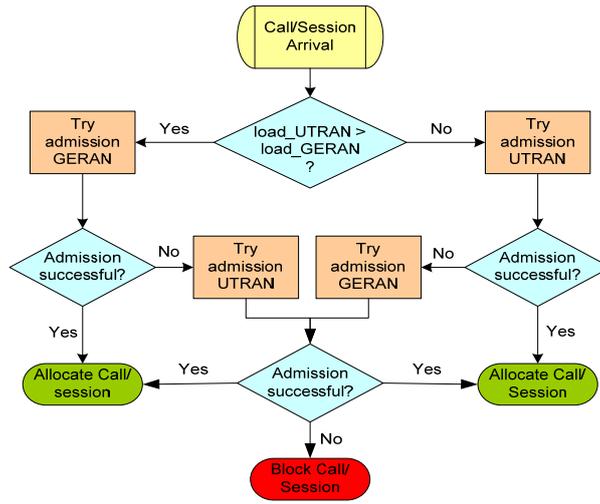
A basic RAT selection policy can be defined as a function that selects a suitable RAT according to some input parameters, in our case, the service class and the load in each RAN. The performance of the RAT selection policies are evaluated considering UTRAN and GERAN access technologies supporting a mixing of voice and interactive (www) users. Let the following policies be considered hereon:

- Service-based policy (termed VG*VU): This corresponds to the 2-complex policy VG*VU defined in 5.3. At first, this policy attempts to assign voice users to GERAN and www users to UTRAN. If no resources are available in GERAN, voice users try admission in UTRAN. Similarly, rejected www users in UTRAN will attempt admission in GERAN. If no resources are available in any of the RATs, the user gets blocked.
- Load balancing policy (termed LB): Upon call/session arrival, this policy adaptively selects the RAT with the minimum load metric, as described in the following, provided that there are available resources in this RAT. Otherwise, the user gets blocked.

Figure 239 shows the flow-charts for the two abovementioned RAT selection policies.



(a)



(b)

Figure 239 Flow-charts of (a) VGVU policy and (b) LB policy

An influential run-time parameter in a load balancing decision-making procedure is the *load metric*. In our study the following metrics are considered: In UTRAN, the well-known load factor expressions [147] are used in their window-averaged form, here defined as

$$\eta_T(i) = \frac{\sum_{j=1}^T \eta(i-j)}{T} \tag{63}$$

where T is the window size for averaging, which is given in number of UTRAN frames. The uplink load factor in the i th frame is estimated as

$$\eta_{UL} = 1 - \frac{P_N}{I_{total}} \tag{64}$$

with P_N the background thermal noise and I_{total} the total received wideband power. The downlink load factor in the i th frame is

$$\eta_{DL} = \frac{P_{total}}{P_{max}} \tag{65}$$

where P_{total} is the total downlink transmission power and P_{max} is the maximum Node-B transmission power.

As for GERAN, a useful way to measure the data load is to measure the average amount of time slots (TSL) utilized by GSM/EDGE services [133]. A window-averaged timeslot utilization factor is defined

$$TSL_{utilisation,T} = \frac{\sum_{j=1}^T TSL_{utilisation}(i-j)}{T} \tag{66}$$

with the timeslot utilization factor in the i th frame being

$$TSL_{utilization} = \frac{\text{Used TSL in previous frame}}{\text{Available TSL for GSM + EGPRS}} \quad (67)$$

where T is the window size for averaging given in number of EGPRS frames. The above expressions are particularized for both the uplink and downlink.

In general, upon call/session arrival, the cell selection procedure selects the base station with best received signal strength, in the case of GERAN, and best E_c/I_0 in the case of UTRAN. When using the load balancing algorithm, the network selects two target base stations, one for each supported RAT. These base stations are chosen to be the ones with best signal strength and best E_c/I_0 for GERAN and UTRAN respectively. For the selected base stations, load metrics are measured and users allocated according to the defined policy. From here on, and for brevity purposes, the term *load* accompanied by the corresponding RAT name will be used when referring to the load factor in UTRAN and the timeslot utilization factor in GERAN.

5.4.2.1 Performance evaluation

The LB RAT selection strategy has been evaluated and compared against the VG*VU service-based policy by means of system level simulations in the same scenario described in sub-section 5.3.2.1. As for the load balancing parameters, load metrics are considered in the uplink direction, i.e. those described by equations (66) and (67) considering the uplink timeslot utilization factor in the last case. Load averaging windows (T) are chosen to be of length 10 seconds.

Given that policy VG*VU allocates users according to the demanded service-type, we can foresee that the traffic mix will impact the performance of this policy. Therefore, in order to evaluate the suitability of policy LB, we consider two representative service mixes, SM1 and SM2, which are chosen so that different stress conditions are noted in GERAN when policy VGVU is applied. In SM1 the number of interactive users is fixed while voice users increase. On the contrary, in SM2 the number of voice users is fixed while the number of interactive users increases.

For the sake of brevity, results are shown in the uplink direction although the same trend was observed in the downlink.

5.4.2.1.1 Service Mix 1 (SM1)

Figure 240 shows the average cell load of the central base station in both RATs for SM1 when policies VGVU and LB are used. Note that for VGVU policy, voice users are directed to GERAN while not fully-loaded; otherwise, requests are transferred to UTRAN. Load balancing policy LB behaves as expected, maintaining cell load levels in both RATs at approximately the same level.

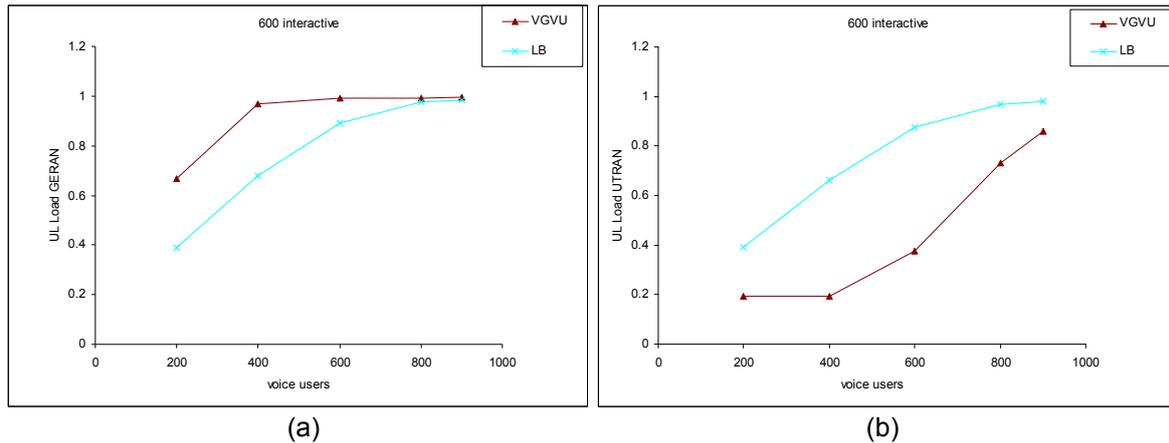


Figure 240 Average UL load in (a) GERAN and (b) UTRAN for policies VGVU and LB with SM1.

Figure 241 illustrates the total aggregated throughput for SM1 when using policies VGVU and LB. Results show an improvement of total aggregated throughput with policy LB. Due to VGVU policy, users in GERAN bear higher load conditions (see Figure 240), which in turn causes dropping to increase. Therefore, throughput contribution of these users, mostly voice users allocated by VGVU, diminishes. Recall that no inter-RAT handovers are considered in this case study.

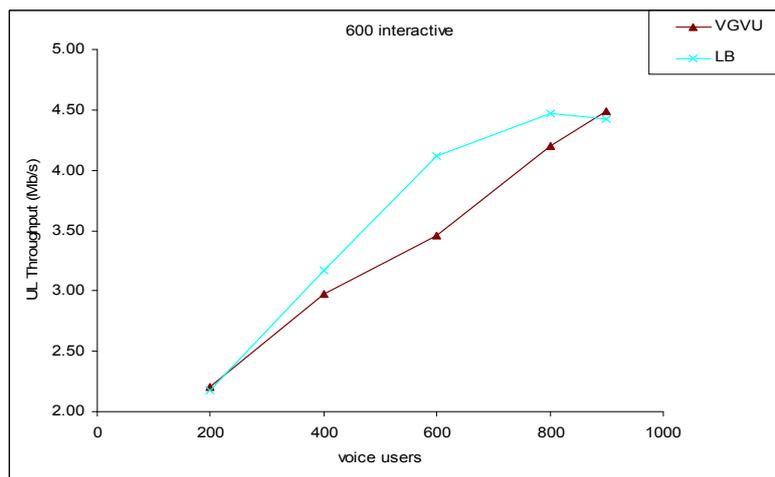


Figure 241 Total UL aggregated throughput with SM1.

Figure 242 shows the average packet delay of interactive users for both policies. Results indicate that interactive users being allocated with policy VGVU undergo lower average packet delays, which benefit the perceived QoS of those users.

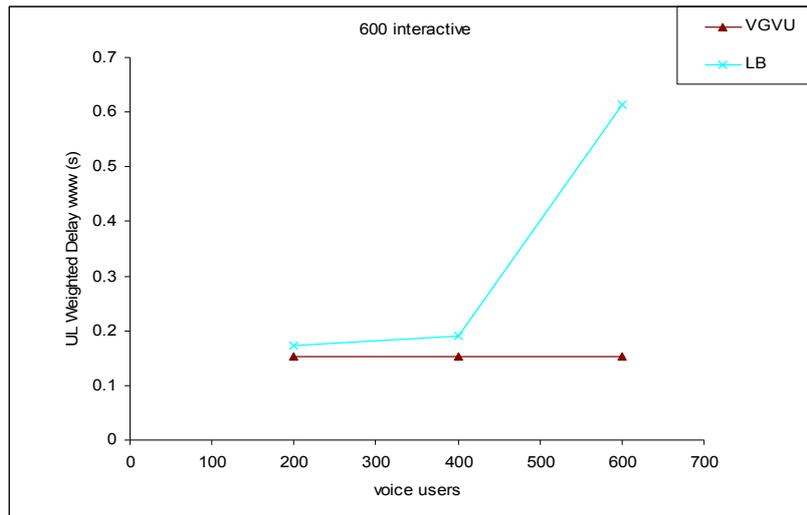


Figure 242 Average packet delay for interactive users with SM1.

It has been shown that, for SM1, load balancing policy tradeoffs the overall performance of the system, in terms of total aggregated throughput, with the performance reduction of interactive users. Interactive users that are forced to GERAN by means of load balancing procedures may exhibit degradation in terms of average delay packet delay.

5.4.2.1.2 Service Mix 2 (SM2)

The average cell load for SM2 is depicted for the abovementioned policies in Figure 243. In this case, the stress is not set on GERAN, which can manage its share of 200 users. As for UTRAN, a moderate load increase is observed, although easily handled. On the other hand, policy LB exhibits the expected behaviour in terms of similar load levels in both RATs.

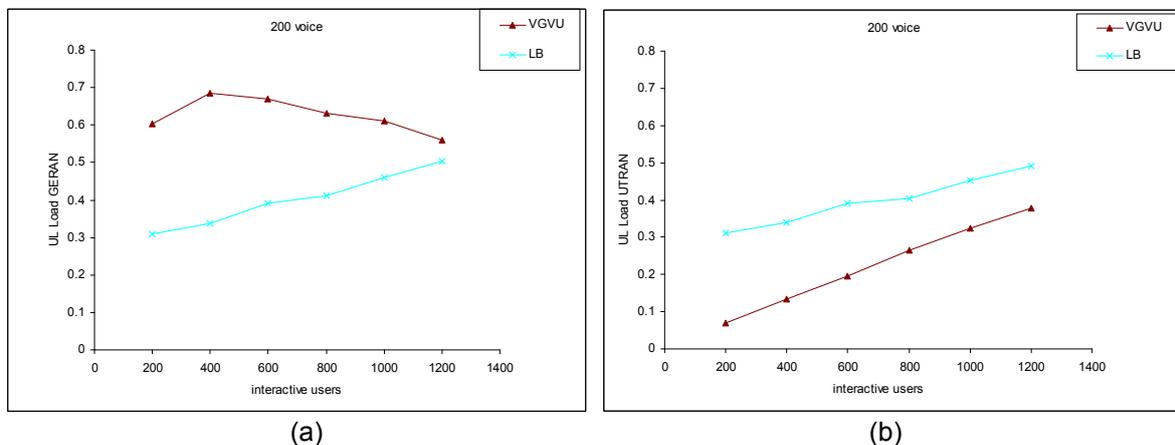


Figure 243 Average UL load in (a) GERAN and (b) UTRAN for policies VGUV and LB with SM2.

Figure 244 shows the total uplink aggregated throughput for policies VGUV and LB. For this service mixing, policy LB does not show a visible improvement with respect to the service class policy VGUV. Average load curves (Figure 243) indicate that load levels are kept low, compared to SM1, and therefore no severe dropping occurs.

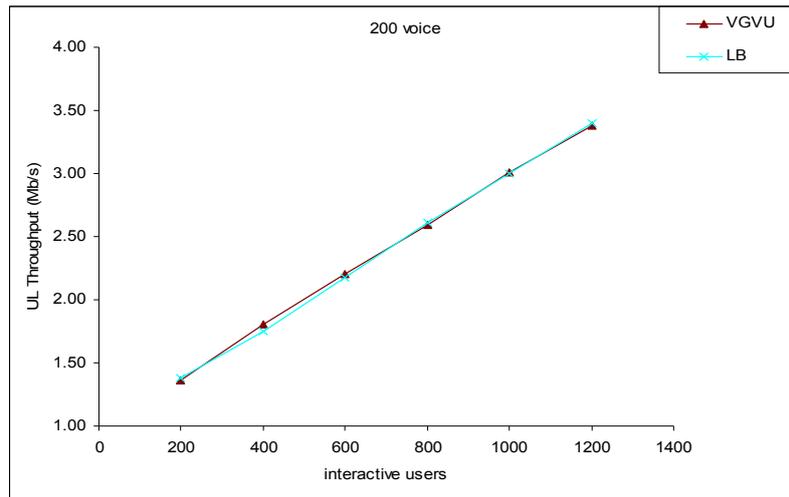


Figure 244 Total UL aggregated throughput with SM2.

Finally, the average weighted packet delay exhibited by interactive users in both RATs is depicted in Figure 245. Similar to the case of SM1, a degradation of delay performance is noted by forcing load balancing among RATs when interactive users are actually best served in UTRAN. Note that the degradation in terms of packet delay is less severe than for SM1, in part because GERAN can now manage interactive users better.

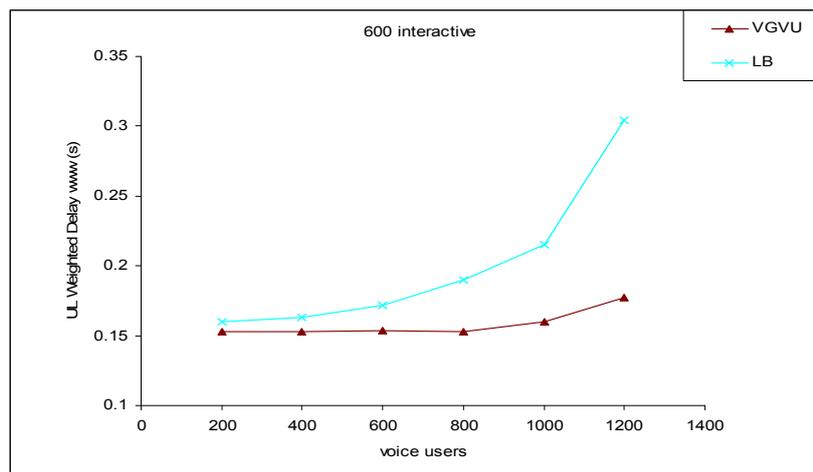


Figure 245 Average weighted packet delay for interactive users with SM2.

5.4.3 Scenario including vertical handover

This section presents some results comparing the performance of policies VGVU and LB in a GERAN/UTRAN heterogeneous network scenario with enabled VHO capabilities. In particular, for VHO we consider the approach denoted as tight-VHO (T-VHO). This term refers to the coupling between horizontal handovers (HHO) and VHO, as described in sub-section 5.3.4.

5.4.3.1 Load and service-class distribution

Figure 246 shows the uplink average cell load in GERAN (Figure 246.a) and UTRAN (Figure 246.b). For policy VGVU, once GERAN is fully loaded (i.e. for 400 voice users), UTRAN load starts to ramp up due to the redirected voice users that cannot be served in GERAN. On the other hand, load balancing policy behaves as expected, maintaining loads in both RATs at the same level.

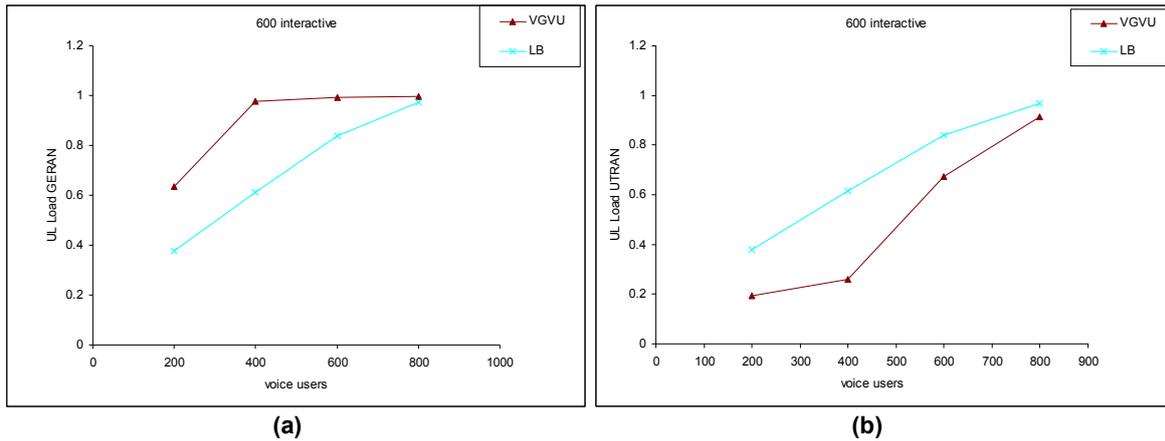


Figure 246 UL average cell load in (a) GERAN and (b) UTRAN for policies VGUV and LB considering VHO.

In order to compare the loads in both RATs, a new KPI (Key Performance Indicator) is considered, denoted as *Load Balancing Factor* (LBF). This function acts as a fairness quantifier and it was proposed in [148]. In particular, the LBF for a set of n different RATs is given by:

$$LBF = \frac{\left(\sum_{i=1}^n \mathcal{L}_i\right)^2}{n \cdot \left(\sum_{i=1}^n \mathcal{L}_i^2\right)} \quad (68)$$

with \mathcal{L}_i the average load in RAT i , and $1 \leq i \leq n$ with n the total number of available RATs.

The LBF is bounded between 0 and 1. A perfect load distribution (i.e. equal loads in each RAT) yields to a LBF equal to 1. On the other hand, a total unbalanced situation yields to a LBF equal to $1/n$.

In Figure 247 the LBF is plotted for both RAT selection policies. As expected, LB policy, exhibits a perfect load balancing among both RATs. Regarding the VGUV policy, once GERAN is full, redirected voice users to UTRAN causes the loads in both RATs to balance.

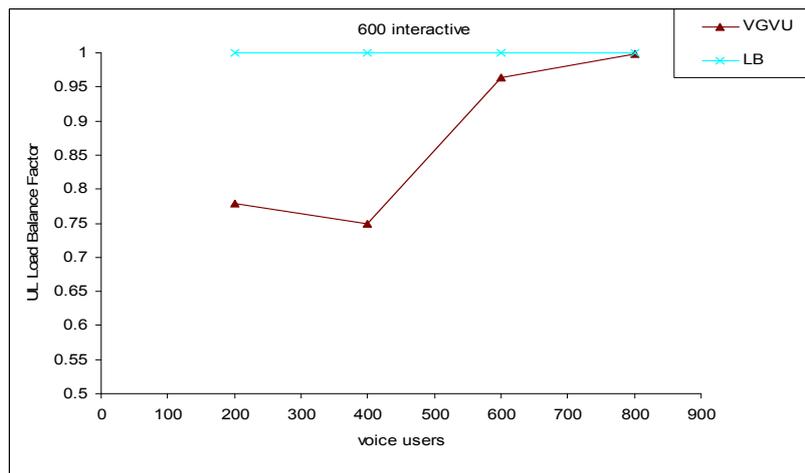


Figure 247 UL Load Balance Factor for RAT selection policies VGUV and LB considering VHO.

Figure 248 shows the average number of allocated users per cell of each service class in each RAT. A similar pattern to what was shown for the load levels in each RAT (Figure 246) is observed in the average number of voice users allocated in GERAN and UTRAN (Figure

248.a and Figure 248.c respectively). Keep in mind that voice users are driving the major load changes in our study, therefore this situation should not be unexpected.

We also note that in order to keep both RATs balanced, a higher number of voice users are needed in UTRAN than in GERAN, i.e. the load consumption of voice users in GERAN is higher than in UTRAN. Interactive users, on the other hand, when using policy VGUV are mostly allocated in UTRAN. However, when the number of voice users is very high (e.g. 800) the VGUV may allocate interactive users in GERAN because no resources are left in UTRAN. In this situation, with high load in both systems, VGUV tends to behave as load balancing scheme as we saw in Figure 247.

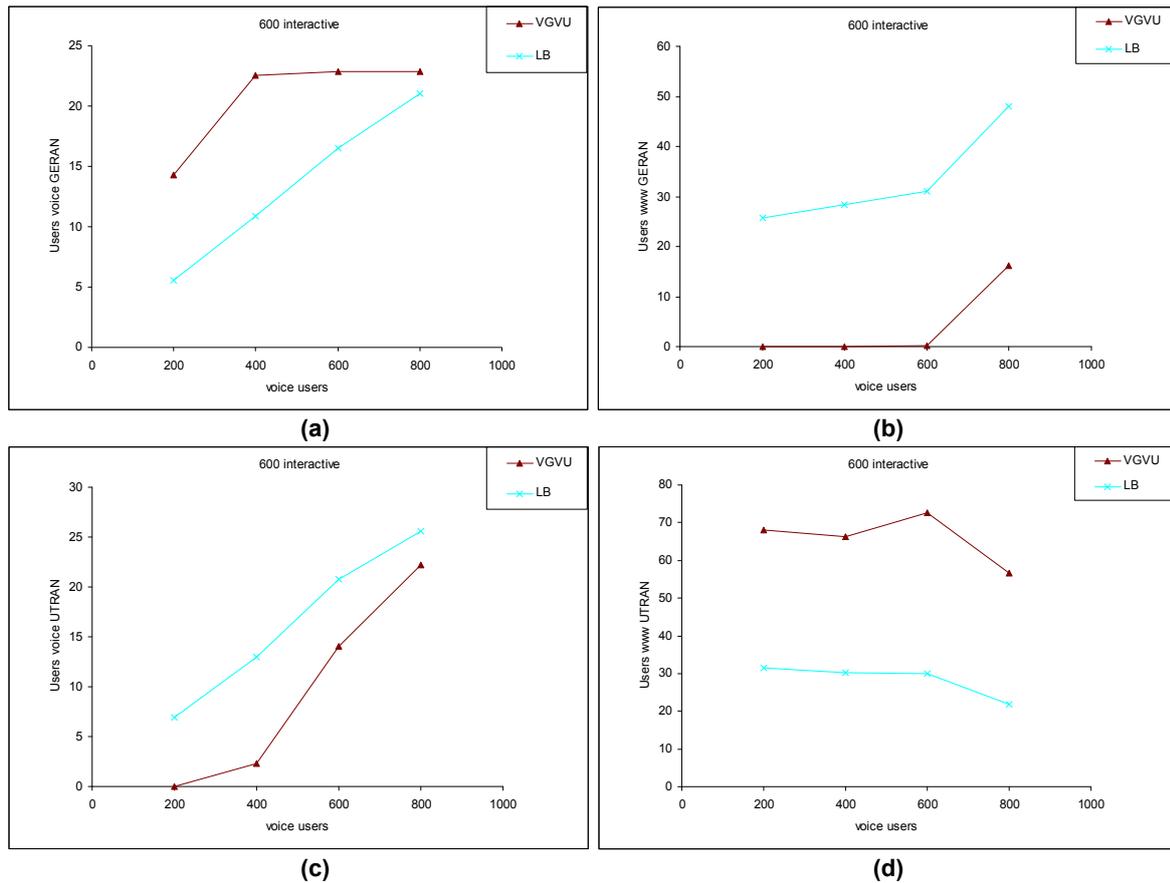


Figure 248 Average number of (a) voice users in GERAN (b) interactive users in GERAN (c) voice users in UTRAN (d) interactive users in UTRAN for RAT selection policies VGUV and LB considering VHO.

5.4.3.2 Vertical Handover Rates

Figure 249 shows the handover rates (i.e. the average number of successful vertical handovers per call or session) for voice and interactive users migrating to and from GERAN and UTRAN. For policy VGUV, increasing the number of voice users forces these users to be redirected by means of a VHO to UTRAN (see Figure 249.a).

Notice in Figure 249.a, how for 800 voice users, the rate of VHO is the same as for 600. It happens that when both systems are fully loaded, VHO will only take place whenever a call leaves the system.

Figure 249.b shows the VHO rate of voice users to GERAN. A similar trend in the rate of VHO is noted to the one observed in Figure 249.a. As mentioned in sub-section 5.3.4, tight VHO allows a better fulfilment of the indications given by the RAT selection policy, in this

case VGVU, so voice users allocated in UTRAN will try to return to GERAN. Interactive users, on the other side, experience lower VHO rates in both directions for the case of policy VGVU, meaning that these users are better served in UTRAN. In the case that UTRAN cannot serve interactive users (because it is at full load) interactive users will be redirected to GERAN. However, the considered VHO approach will tend to move the interactive users back to UTRAN, which can be seen in the increasing VHO rate for VGVU policy (Figure 249.c).

Regarding the LB policy, voice VHO towards UTRAN will happen whenever the load in UTRAN is less than in GERAN. This is likely to happen, due to the fact that users contribute to a higher load in GERAN than in UTRAN. The rate of voice VHO to GERAN will also occur as long as the load is low. If the system is highly loaded, resources in GERAN are easily occupied, thus many voice users will remain in UTRAN. This is noted in the decreasing slope of VHO rates for voice users to GERAN (Figure 249.b).

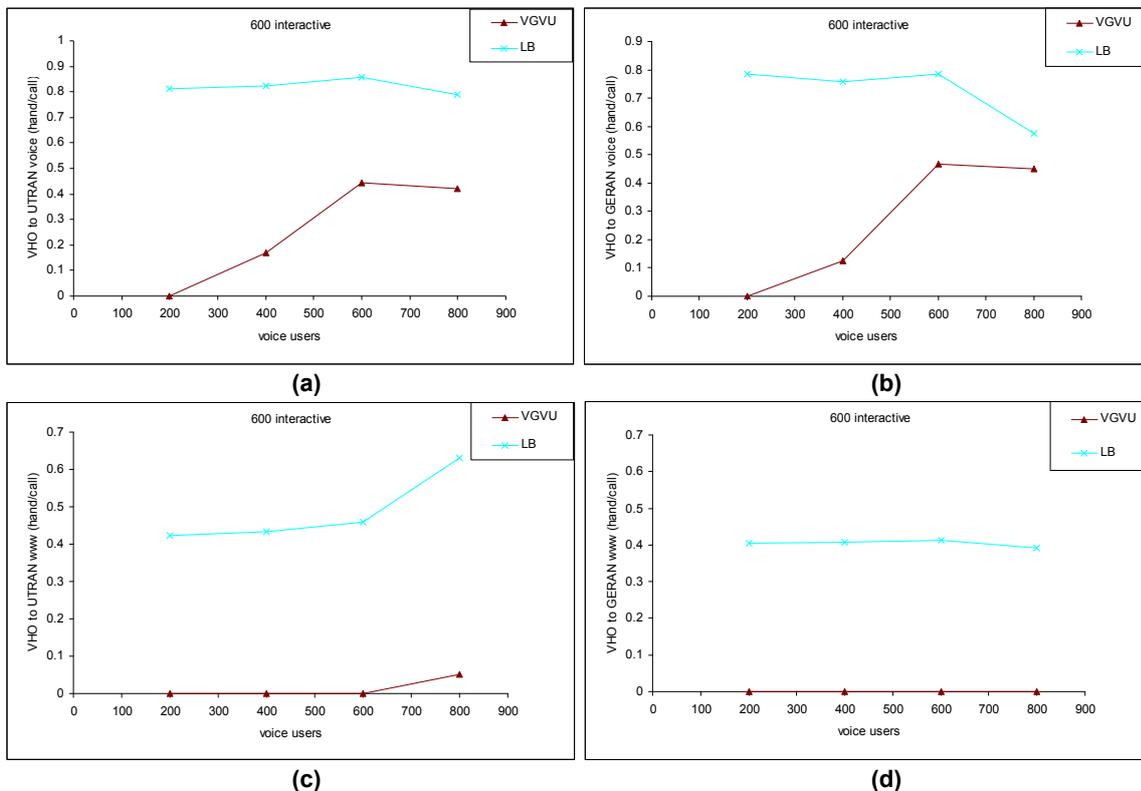


Figure 249 VHO Rates (a) voice users to UTRAN (b) voice users to GERAN (c) interactive users to UTRAN (d) interactive users to GERAN for RAT selection policies VGVU and LB considering VHO.

5.4.3.3 Performance evaluation of voice users

The performance evaluation for voice users being served in this multi-access scenario is done by means of two metrics: the Block Error Rate (BLER) of voice users in UTRAN and the total dropping probability of voice users.

Figure 250 shows the BLER of voice users measured in the uplink in UTRAN. An acceptable BLER is defined to be at most 1% for voice users. Up to 600 voice users, both policies approximately achieve the desired BLER level. For 800 voice users however, it can be seen that the VGVU policy experiences a lower BLER than the LB policy.

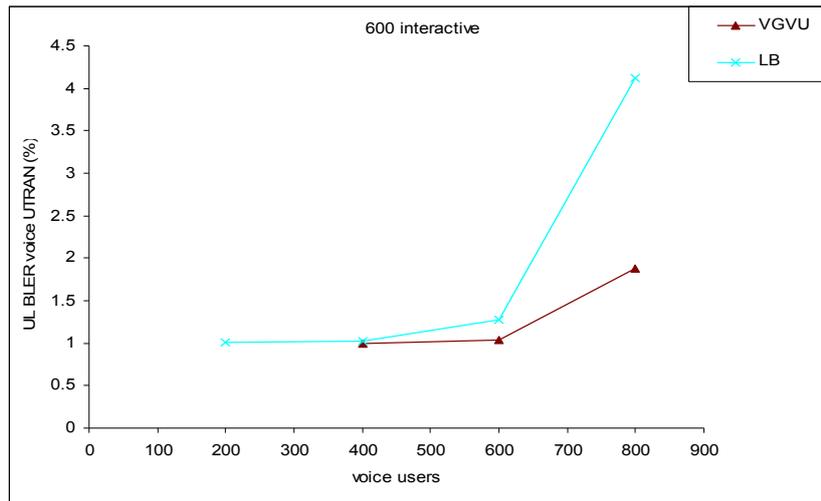


Figure 250 UL BLER (%) for voice users in UTRAN considering policies VGVU and LB for VHO.

Figure 251 shows the voice call dropping probabilities (in %) for the selection policies under study. Up to 600 voice users, dropping values are kept sufficiently low. For 800 voice users however, policy VGVU reveals higher dropping values than policy LB. The higher dropping rates experienced by VGVU policy may be explained bearing in mind the load distribution in GERAN induced by policies VGVU and LB (Figure 252). In particular, for VGVU, the load is at its maximum value most of the time which implies a lack of flexibility in order to accommodate handover users being redirected to GERAN. Therefore, VGVU may incur in more potential dropping situations than in the case of LB policy appliance which, as can be seen, presents some fluctuation in its load values and could provide resources to incoming handover users if necessary.

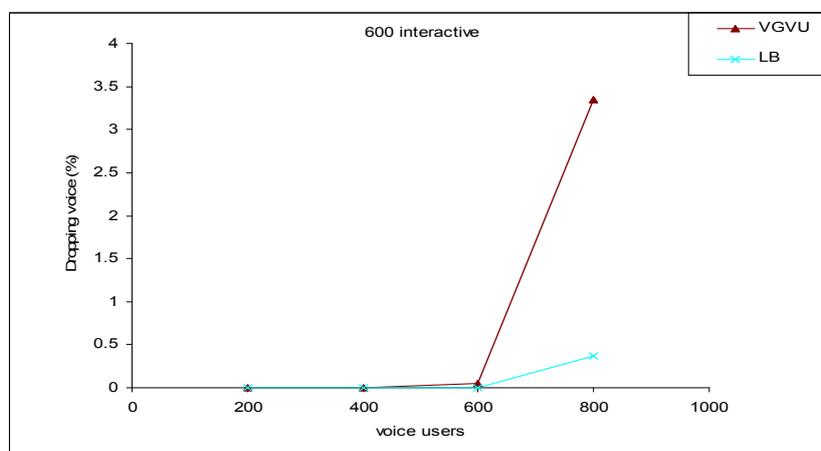


Figure 251 Dropping Probability (%) for voice users considering policies VGVU and LB with VHO.

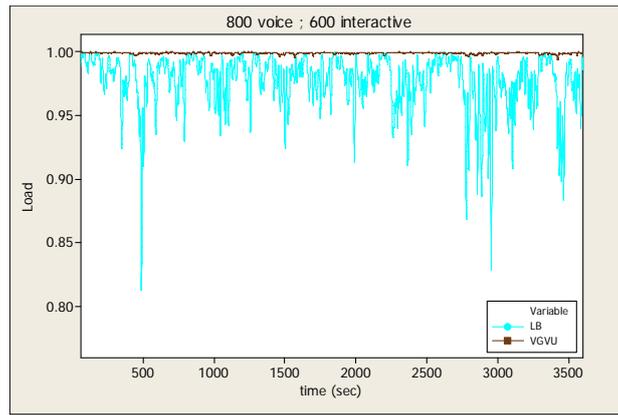


Figure 252 Load in GERAN vs. time considering policies VGVU and LB with VHO.

5.4.3.4 Performance evaluation of interactive users

Figure 253 shows the average downlink packet delay exhibited by interactive traffic. Results show that VGVU policies reveal lower delays than with the appliance of policy LB, especially when increasing voice load. Bear in mind that once UTRAN is fully loaded, interactive users will be redirected to GERAN. This will in turn cause interactive users to wait until a TBF is assigned to them, which will dramatically increase the delay of such users.

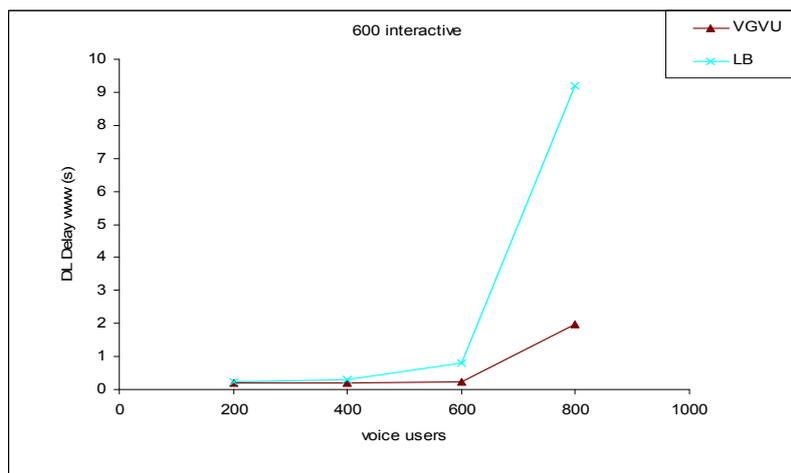


Figure 253 DL Average Packet delay (s) for interactive users considering policies VGVU and LB with VHO.

5.4.3.5 Throughput performance

Figure 254 shows the total aggregated throughput for policies VGVU and LB. Results indicate that no RAT selection policy combined with the given VHO strategy performs better than any other one. The flexibility provided by VHO is capable enough to redirect users to the most appropriate RAT depending on the current network situation, e.g. load, BLER, etc. Therefore, no significant improvement is introduced by trying to load balance the system.

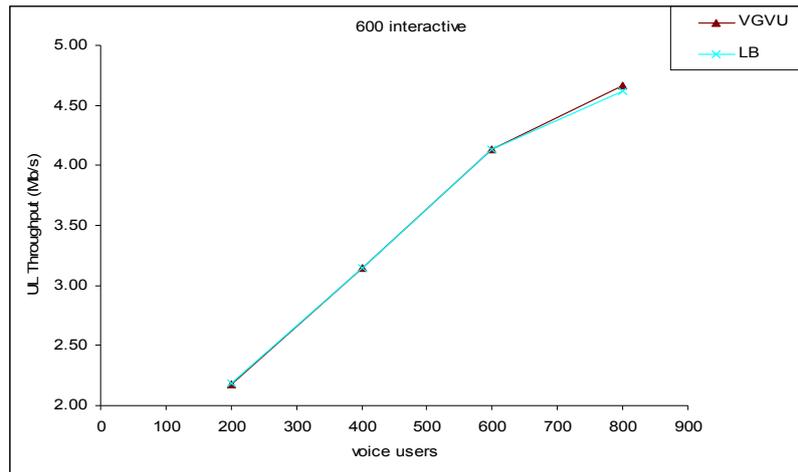


Figure 254 Total UL aggregated throughput (Mb/s) for policies VGUV and LB considering VHO.

5.4.3.6 Admission probability

Admission probabilities of voice users and interactive users are depicted in Figure 255.a and Figure 255.b respectively. Higher voice admission values are achieved by LB as opposed to VGUV during high load situations. Certainly, as already seen, load conditions in GERAN caused by VGUV policy appliance are more stringent than in the case of LB, thus more voice users get blocked in the case of VGUV. Interactive users, on the other hand, are always admitted. However, they will remain in a buffer until they can get served, thus impacting on their average delay.

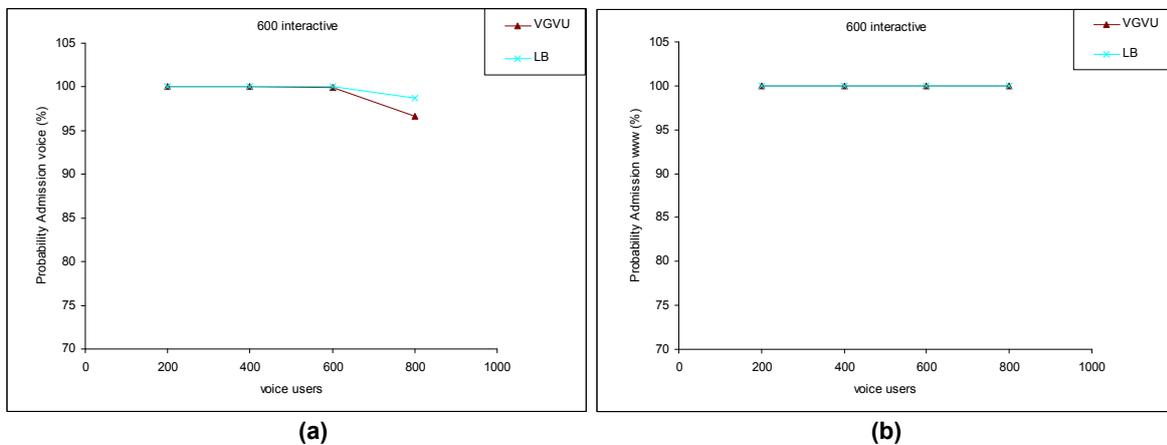


Figure 255 Admission probability (%) for (a) voice users and (b) interactive users considering policies VGUV and LB with VHO.

5.4.4 Conclusions

This sub-section has analysed the performance of load balancing principles in the RAT selection procedure compared against a service-based policy. With respect to the initial RAT selection without including VHO, results revealed a tight dependency between the suitability of load balancing RAT selection and service-class mixing. It has been shown that even though the overall throughput may increase with load balancing policies, this at the expense of interactive traffic performance. Nevertheless, other service type mixings showed no type of throughput improvement at all. In turn, the introduction of VHO capabilities allows higher flexibility in the allocation of multi-service users in a multi-access scenario. We have compared two initial RAT selection policies along with a tight approach for VHO procedures. Results indicate that no remarkable improvement is noted on the total aggregate throughput

when using the LB policy as opposed to the VGVU policy. Moreover, with LB, interactive users undergo higher average packet delays which impact the user's perceived QoS. However, we have seen that load balancing procedures may improve the call dropping probability due to a more flexible allocation of users onto both RATs, which is also a key performance indicator to consider.

5.5 PATH LOSS - BASED RAT SELECTION

5.5.1 Introduction

In general, cellular wireless systems become interference-limited and, consequently, any engineering technique devoted to either reduce interference or to improve the robustness of the system to bear interference will readily increase network capacity and operator's revenue. In this context, this section intends to exploit the different sensitivity that diverse RATs may exhibit to interference so that a smart CRRM follows. In particular, in FDMA/TDMA-based access systems (e.g. GSM/GPRS) there is no intra-cell interference. In turn, inter-cell interference is caused by a single user in every co-channel cell. In contrast, in CDMA-based systems (e.g. UMTS) the intra-cell interference is caused by every single user transmitting in the cell. Furthermore, inter-cell interference is also originated by all simultaneous users in all neighbouring cells, since a complete frequency reuse is considered. Consequently, CDMA systems are much more sensitive to multi-user interference than FDMA/TDMA ones.

The underlying idea of the CRRM approach developed here is to take advantage of the coverage overlap that several RANs using different access technologies may provide in a certain service area in order to improve the overall interference pattern generated in the scenario for the CDMA-based systems and, consequently, to improve the capacity of the overall heterogeneous scenario. This can be achieved by controlling the effective cell radius of CDMA-based systems (i.e. a controlled cell-breathing effect) through appropriate RAT selection approaches that take into account the measured path loss. In this way, the interference level in CDMA-based RATs is reduced while at the same time the target coverage area is assured by means of the cooperation of the FDMA/TDMA-based RATs.

The above concept, denoted here as Network-Controlled Cell-Breathing (NCCB) can be effectively complemented with load balancing considerations. That is, the control of the CDMA effective radius has also a straight impact on the way how the load is distributed among the RATs, i.e. if the CDMA radius is too small, a higher number of users will exist in the FDMA/TDMA while if the CDMA radius is too high, the opposite situation may occur, thus leading to load unbalance situations that may limit the flexibility of the CRRM approach. Then, the performance achieved with a NCCB strategy can be improved if load balancing principles are also applied into the RAT selection criteria. In this sense, this section advances state-of-the art developments in the CRRM field by exploiting the concepts of network-controlled cell-breathing and load balancing through CRRM strategies, proposing specific algorithms and evaluating them with detailed system level simulations in order to proof the concepts.

The proposed controlled cell-breathing strategy is illustrated in Figure 256 for a situation where CDMA and FDMA/TDMA cells are co-sited. R_T denotes the planned cell radius in FDMA/TDMA and R_C denotes the variable effective cell radius in CDMA. Notice that, for a given service, the FDMA/TDMA cells ensure coverage in the whole area. In turn, by an appropriate control of the effective cell radius R_C (e.g. in the figure by changing from R_{C2} to R_{C1}) in CDMA cells through CRRM strategies, the required transmitted power levels and the inter-cell interference will be reduced, thus improving the capacity for the considered service in the CDMA RAT. Notice that, depending on the existing load conditions as well as the

robustness of the specific services to interference, the CDMA radius could be eventually set equal to R_T . Similarly, Figure 257 illustrates the situation in which no co-siting exists between FDMA/TDMA and CDMA cells, reflecting that the proposed concept would also be applicable in this case.

In practice, due to the shadowing effects, the cell radius is controlled by setting the maximum propagation loss that can be allowed for a given RAT. Taking this into account, this section proposes a CRRM strategy that allocates users to RATs according to their propagation losses. Then, users with low propagation loss will be allocated to the CDMA cells and users with high propagation loss will be allocated to the FDMA/TDMA cells. It will be shown that, by setting a suitable maximum path loss threshold, the CDMA-based RAT achieves the desired network-controlled cell-breathing effect and the corresponding increase in performance.

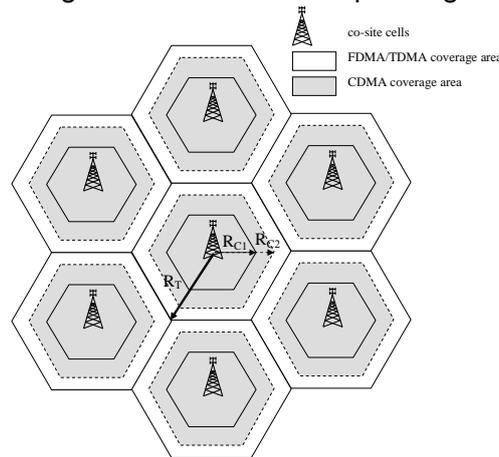


Figure 256 Network controlled cell-breathing when CDMA and FDMA/TDMA cells are co-sited

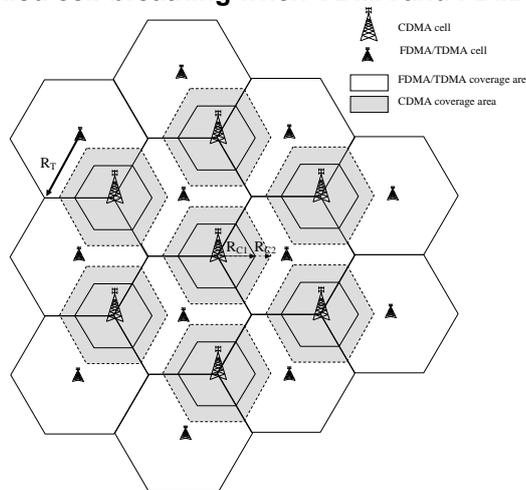


Figure 257 Network controlled cell-breathing without co-siting between CDMA and FDMA/TDMA cells

At this stage, it is worth mentioning that, in order to introduce the potentials of the network-controlled cell-breathing concept introduced here, the feasibility of a suitable mapping between equivalent bearers in the considered RATs will be assumed (e.g. a voice service supported with a given bearer in RAT#1 can be supported with similar QoS with a given bearer in RAT#2). The extension of the proposed strategy to multi-service scenarios with some services not having equivalent bearers (e.g. a 384 kb/s bearer supporting an interactive service in WCDMA does not have the counterpart in 2G TDMA-based system) is considered out of the scope of this analysis.

5.5.2 Preliminary theoretical evaluation

The scenario we studied is illustrated in Figure 256 for a situation where CDMA and FDMA/TDMA Base Stations (BS) are co-sited. R_T denotes the planned cell radius in FDMA/TDMA and R_C denotes the effective cell radius in CDMA. With CRRM functionalities placed in this heterogeneous network, the traffic can be freely located between the two systems through the vertical handover (VHO) and initial RAT selection procedures. Thus the resource pool co-ordinated by CRRM in one cell is given as

$$C_T = C_{CDMA} + C_{F/T} \quad (69)$$

where C_T is the total number of channels offered by the two systems, C_{CDMA} is the number of channel that can be offered by the CDMA system and $C_{F/T}$ is the number of channels offered by the FDMA/TDMA system.

If both CDMA and FDMA/TDMA systems offer the same coverage (i.e. $R_T=R_C$), the total the number of channel is C_T for all the users in the cell coverage and then the blocking rate of the cell is given as,

$$P_b = \frac{A^{C_T}}{C_T!} \frac{1}{\sum_{i=0}^{C_T} \frac{A^i}{i!}} \quad (70)$$

where $A=\lambda/\mu$ is the traffic load, λ is the arrival rate and μ is the depart rate.

Figure 258 shows the blocking rate as a function of traffic load A with different values of C_T . Obviously, the blocking rate increases as traffic load increases and, a bigger value of C_T leads to a better blocking rate performance. As mentioned in the previous section, the interference in the FDMA/TDMA system is mainly from distanced co-channel users whilst the CDMA systems are subject to multi-user interference from both intra-cell and inter-cell. So the coverage of a FDMA/TDMA is rather static comparing with CDMA-based cellular system for a given service. The whole area is fully covered by FDMA/TDMA system through system planning. Thus in the FDMA/TDMA system, $C_{F/T}$ is given as the number of time slots for the traffic, which is fixed according to the number of frequency channels allocated in this cell. However C_{CDMA} , which is defined as the number of simultaneous links that can be supported in the cell for a given service with a certain quality requirements, is actually a coverage dependent value. In particular, Table 63 shows the downlink value of C_{CDMA} against the effective cell radius R_C in a scenario with a voice service. C_{CDMA} is defined as the number of simultaneous voice links that can be supported while keeping an outage below 5% (i.e. the probability that the E_b/N_0 is below the target). It is easy to find that with a smaller coverage, a larger C_{CDMA} is expected. Notice that roughly 20-30% capacity gain can be obtained for CDMA systems by reducing 100 meters in the radius

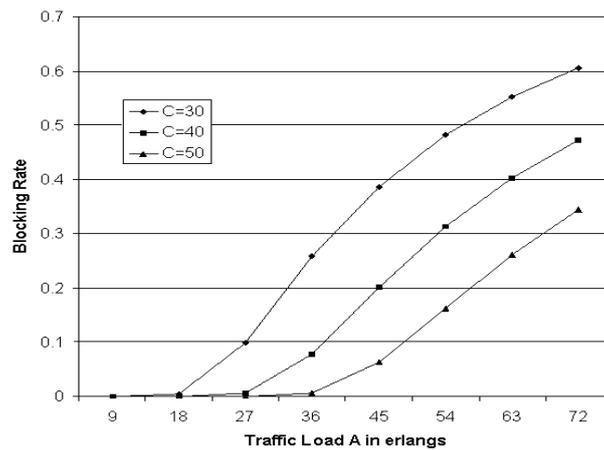


Figure 258 Blocking probability for different values of C_T

Table 63 CDMA capacity as a function of the coverage radius

Radius in m	1000	900	800	700	600	500
C_{cdma}	30	38	45	55	69	91
Gain (%)	0	27	18	22	25	31

In this proposed system, with a fixed value of $C_{F/T}$, based on (69) and (70), one possibility for improving the blocking performance is to increase the value of C_{CDMA} . Assuming that the positions of the base stations are not changed, C_{CDMA} can be increased by a proper CRRM strategy that controls the effective CDMA cell radius. In this strategy, for a given service such as voice, the FDMA/TDMA cells ensure coverage in the whole area. For the CDMA system, we can increase C_{CDMA} value by controlling R_c , e.g. when we reduce the R_c from 1 km to 0.9km, we have an increase of 8 in C_{CDMA} (see Table 63). However since if we reduce the CDMA BS coverage, a certain amount of traffic has to be handed over from CDMA to FDMA/TDMA system, this would affect the blocking probability of the FDMA/TDMA system. Consequently, the benefit achieved by the proposed strategy will depend on the following factors:

1. the number of channels is gained in the area covered by both CDMA and FD/TDMA systems
2. the number of available of FD/TDMA channels in the area not covered by the CDMA system
3. the traffic distribution over the two areas

In the following, we study the coverage-based CRRM control based on the above three factors in order to assess the feasibility if coverage-based CRRM is able to improve the system performance.

In the proposed scheme, the total FDMA/TDMA cell is divided into two areas: IN area and OUT area as shown in Figure 256. The IN area is covered by both FDMA/TDMA and CDMA BSs, and the OUT area is only covered by the FDMA/TDMA BS. To guarantee the continuous coverage for the whole area, R_T in this case is kept at constant by FDMA/TDMA system. And the IN area radius (R_c) is controlled by the CRRM based on the three factors we mentioned above to achieve a better blocking rate.

With this division, to guarantee a certain performance for the calls in the OUT area, a number of FDMA/TDMA channels, denoted as N_2 , are allocated for the OUT area calls. Then the number of channels for IN area is given as

$$N_1 = C_T - N_2 \quad (71)$$

To clearly show how the coverage based CRRM works, we introduce another two performance measures, IN area blocking rate P_1 and OUT area blocking rate P_2 , which are given as

$$P_1 = \frac{\frac{A_1^{N_1}}{N_1!}}{\sum_{i=1}^{N_1} \frac{A_1^i}{i!}} \quad (72)$$

$$\text{and } P_2 = \frac{\frac{A_2^{N_2}}{N_2!}}{\sum_{i=1}^{N_2} \frac{A_2^i}{i!}} \quad (73)$$

where A_1 and A_2 are the traffic load in IN area and OUT area respectively. If the CDMA radius is given as R_C , then we define the coverage probability of p_1 as the ratios of the IN area over the whole cell as the following,

$$p_1 = \frac{A_1}{A_1 + A_2} \quad (74)$$

Then, the blocking rate for a the whole cell is given as

$$P = \frac{\frac{A_1^{N_1}}{N_1!}}{\sum_{i=1}^{N_1} \frac{A_1^i}{i!}} \cdot p_1 + \frac{\frac{A_2^{N_2}}{N_2!}}{\sum_{i=1}^{N_2} \frac{A_2^i}{i!}} \cdot (1 - p_1) \quad (75)$$

A_1 and A_2 are geographical related to R_C . If we assume that the traffic are uniformly distributed, we can have the following

$$p_1 = \frac{\pi R_c^2}{\pi R_T^2}$$

then $A_1 = \frac{\lambda p_1}{\mu}$, $A_2 = \frac{\lambda(1 - p_1)}{\mu}$. (76)

Now we have a close look at how the coverage-based CRRM works with the above model. With a given traffic load in the whole FDMA/TDMA cell, the IN area radius is R_{C_1} as shown in Figure 256, and blocking rate in both IN and OUT areas are the same with N_1 and N_2 as shown by (72) and (73). If we reduced the coverage of CDMA by reducing R_C in Figure 256, in the IN area, it will leads a decrease in traffic load (ΔA) and an increase ((N_1) in N_1 , caused by the increase in CCDMA) thus an improvement in the blocking rate can be obtained In the mean time, it will increase the traffic load ((A) in the OUT area. In that case, if we do not change the number of reserve channel (N_2), the blocking rate will increase for the OUT area. If let P_1 , (and P_2 , (denote the blocking rates for IN and OUT area after the radius has been reduced, so we have the following conditions,

$$P_{1,\Delta} > P_1 \text{ and } P_{2,\Delta} < P_2 \quad (77)$$

With the above conditions, it is possible to make an improvement in the general blocking rate as shown in (75). More interesting, we would like to see if it is possible to bring improvement for the IN area calls without scarifying call quality in OUT area. To achieve that, as shown in Figure 258, a possible way is to increase the number of channels (N_2) allocated to OUT area to cope with the increase in the traffic load. If this increase (N_2) is big enough the blocking rate in OUT area will also be kept the same or even be improved. In this circumstance, the way to increase N_2 in the OUT area is to move more FDMA/TDMA channels from IN area to OUT area if there are some available. So if we have enough available FDMA/TDMA channels in IN area and the increase in CCDMA is also big enough, it is possible to improve both IN and OUT area performance by moving the available FDMA/TDMA channels to OUT area. Now let $N_{1,T}$ and $N_{2,T}$ denote the number channels in IN and OUT area respectively with RC 2, which the following conditions

$$P_{1,\Delta} = \frac{(A_1 - \Delta A)^{N_{1,T}}}{\sum_{i=1}^{N_{1,T}} \frac{(A_1 - \Delta A)^i}{i!}} = P_1 \text{ and } P_{2,\Delta} = \frac{(A_2 + \Delta A)^{N_{2,T}}}{\sum_{i=1}^{N_{2,T}} \frac{(A_2 + \Delta A)^i}{i!}} = P_2 \quad (78)$$

Then the feasible conditions to achieve the improvement for both IN and OUT area by reducing CDMA coverage from R_{C1} to R_{C2} are expressed as following

$$N_{1,T} - N_1 \geq \Delta N_1 - \Delta N_{2,T} \text{ and } C_{F/T} \geq N_2 + \Delta N_{2,T} \quad (79)$$

where $\Delta N_{2,T} = N_{2,T} - N_2$

The first condition in (79) ensures that there is a sufficient increase in C_{CDMA} by reducing the radius and the second make sure that there are sufficient FDMA/TDMA channels available to move from IN area to OUT area. So, in the coverage-based CRRM, two controllable parameters we can use to achieve this feasible condition i.e. the controllable radius R_C and N_2 . In the following section, we give some numerical examples to study the effects of controllable radius R_C and N_2 on the system performance in terms blocking rate and to see if we can achieve the above feasible conditions presented in (79).

Figure 259 gives the blocking probability as the function of traffic load. In this evaluation, the FDMA/TDMA cells contain 3 carriers, corresponding to a total of 23 available channels. The number of channels allocated to the OUT area is fixed at 15. With reducing CDMA coverage to $0.9 R_T$, the blocking rate performance is improved. If we consider 0.05 as acceptable blocking rate for users, the capacity is almost improved by 12% (when both systems offer the same coverage, the traffic that can be supported is 45 erlangs, and with $R_C/R_T=0.9$, it is 50.4 erlangs). This means it is feasible to improve the system performance by controlling the coverage. However, if the radius is reduced to $0.8 R_T$, the performance is even worse than that with $R_C/R_T=1$. The performance in the blocking rate is confirmed with our discussion in the previous section, i.e. as the radius of CDMA system decreasing and with a fixed N_2 values, the performance in the IN area enjoys an improvement favored by reduced traffic load and increased capacity, however users in the OUT are face a different situation (increased traffic load and the same capacity), which eventually leads a higher blocking rate. If the blocking rate improvement in the IN area is not big enough to compensate the degradation in the OUT area, the whole performance will be brought down. In this situation

as mentioned in the previous section, we need to bring more FDMA/TDMA channels to the OUT area or increases the CDMA coverage.

If we apply the feasible conditions presented in (79), we have Table 64, which also gives a close look at blocking rate (general blocking rate, IN and OUT area blocking rate) at the point with traffic load of 45 erlangs from Figure 259. When the CDMA radius is reduced from R_T to $0.9 R_T$, based on (79), we find out that, in order to get a better performance or at least not worse, the minimum channel required for IN and OUT area are 44 ($N_{1,T}$) and 14($N_{2,T}$). That means totally 58 channels are needed. From Table 63, we can find that as the CDMA radius reduces from R_T to $0.9 R_T$, C_{CDMA} increases from 30 to 38, the total number of channels increase from 53 ($30+23$) to 61 ($38+23$). This meets the first condition in (79), which means that the increase in C_{CDMA} is sufficient. With the current system configuration ($N_2=15$, and $C_{F/T}=23$), the number of channels in the OUT area (N_2) is greater than its minimum requirement of 14 and also the number of channel in the IN area is 46 that is also greater than minimum requirement of 44. So in this case, both IN and OUT area blocking rates are improved as shown in Table 64. If reduce the R_C/R_T value from 1 to 0.8, in total we need 55 ($32+23$) channel to improve the system performance, and the minimum requirement for N_2 is 23. From Table 64, with the configuration ($N_2=15$, and $C_{F/T}=23$) we can find that, the increase in C_{CDMA} is sufficient (15 more channels than that with $R_C/R_T=1$), however since $N_{2,T}$ (minimum requirement for N_2 in the OUT area is 23 which is greater than current N_2 , the system performance can only be improved by adding at least 8 more FDMA/TDMA channels to OUT area or increase CDMA coverage e.g. increase to $0.9 R_T$ and back into the feasible region.

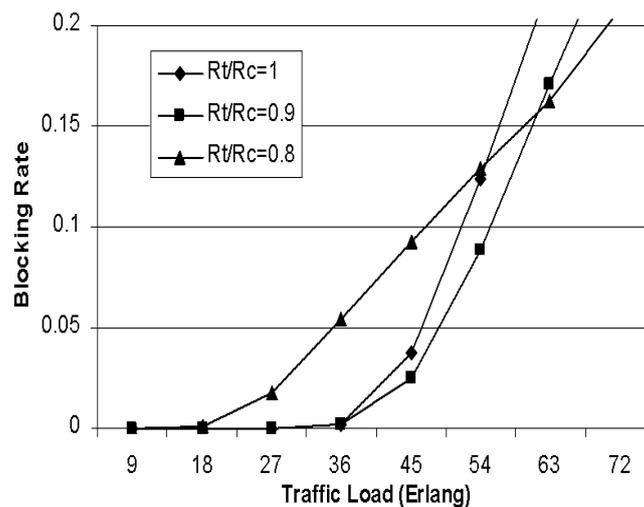


Figure 259 Blocking rate with different coverage ($N_2 = 15$)

Table 64 Feasibility assessment

R_C/R_T	1	0.9	0.8
P_1	0.038	0.025	3e-5
P_2	0.038	0.025	0.26
P	0.038	0.025	0.093
Feasible Conditions			
$N_{1,T}$	--	44*	32*
$N_{2,T}$	--	14*	23*
$C_{F/T}$	23	23	23
$\Delta N1$	--	8	15
$N2$	--	15	15
Feasible (Y/N)	--	Y	N

*Because of Discrete Erlang formula, the value is the closest value

5.5.3 Evaluation in a dynamic scenario

As a result from the above preliminary theoretical evaluation, it can be observed that a certain improvement in terms of capacity can be achieved with the control of the effective cell radius in the CDMA cells by means of proper CRRM strategies. Nevertheless, this gain is subject to the proper configuration of the IN and OUT areas. In order to gain more insight into this configuration and the achieved benefits, in this sub-section, the proposed CRRM strategy is evaluated with dynamic simulations, which allow extending the above results in a more realistic environment including specific initial RAT selection and vertical handover strategies. The evaluation starts with the case in which a single voice service is considered and it is then extended to the case with a mix of voice and www users.

5.5.3.1 Initial RAT selection and vertical handover algorithms

Figure 260 illustrates the flow diagram of the initial RAT selection part in the proposed NCCB strategy. The decision is taken according to the path loss measurements in the best CDMA cell, provided by the terminal in the establishment phase. The path loss is computed by measuring the received downlink power from a common control channel whose transmitted power is broadcast by the network. Measurements are averaged in periods of T seconds. In case that the resulting path loss is higher than a given threshold PL_{th} , the selected RAT will be FDMA/TDMA, while if the path loss is below the threshold the selected RAT will be CDMA. In case that there is no capacity available for the new session in the selected RAT (i.e. the admission control is not passed), the other RAT will be selected instead. Finally, if no capacity is neither available in the other RAT, the session will be blocked.

The corresponding vertical handover decision procedure is shown in Figure 261. The idea behind this procedure is to keep the high path loss users connected to FDMA/TDMA and the low path loss users to CDMA depending on how the propagation conditions vary along the session lifetime. Nevertheless, and in order to avoid undesired ping-pong effects leading to continuous RAT changes for users with path loss close to the threshold PL_{th} , an hysteresis margin Δ (dB) is introduced together with a number of consecutive samples that each condition must be fulfilled before the VHO decision is triggered. This number is M_{up} when the condition is the path loss being above the threshold and M_{down} for the path loss below the threshold. On the other hand, it should be mentioned that the VHO is only executed in case that there is capacity available in the target RAT (i.e. if the admission control is passed). The parameters of the algorithm considered in the evaluation are given in Table 65.

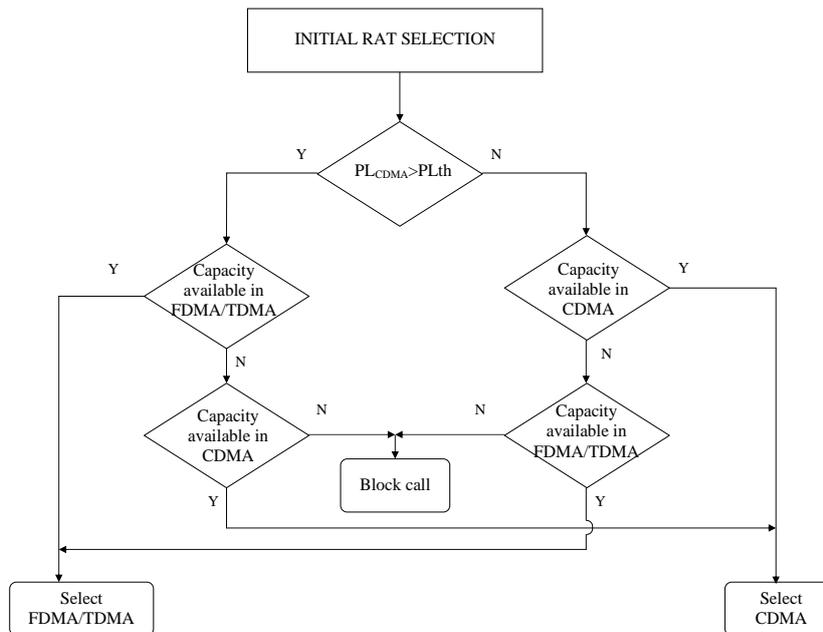


Figure 260 Initial RAT selection in the NCCB strategy

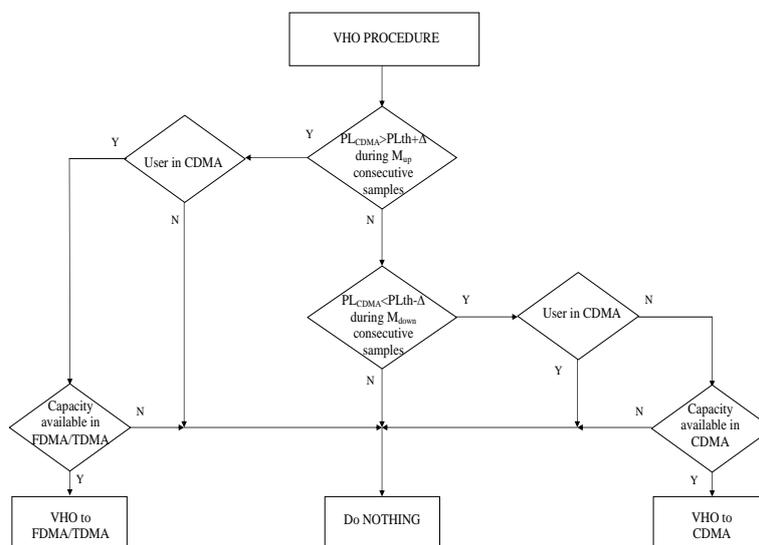


Figure 261 VHO process in the NCCB strategy

Table 65 NCCB Algorithm parameters

Measurement interval (T)	1s
Hysteresis margin (Δ)	1 dB
M _{up}	3
M _{down}	3
PL _{th}	120 dB

5.5.3.2 Evaluation in a single service scenario

5.5.3.2.1 Comparison against load balancing

The results in this section are obtained in the same simulation environment considered in 5.3.2.1 with cell radius 1km with the value of PL_{th}=120 dB. This value corresponds to the 60-th percentile of the path loss distribution, which means that around 60% of the users will perceive a path loss below 120 dB and therefore will be normally allocated in UTRAN while

40% of the users will be allocated in GERAN. The setting of this parameter will be further discussed in the next sub-section.

As a reference, the NCCB algorithm has been compared with a CRRM strategy based on load balancing, which aims at keeping the same load in both CDMA and FDMA/TDMA, as explained in section 5.4. To this end, the initial RAT selection process consists in allocating the user to the RAT with the lowest load level. For CDMA, the load is measured as the uplink load factor, given by the ratio between the intercell and intracell received power with respect to the total received power including background noise. In turn, in FDMA/TDMA the load is measured as the fraction between the occupied slots with respect to the total number of existing slots. Both load measurements are averaged in periods of 10s to smooth load fluctuations and are obtained from the base stations having the lowest path loss among those of each RAT. Furthermore, the vertical handover algorithm is also based on load balancing. To this end, and whenever a horizontal handover is required in the current RAT, the suitability of executing a vertical handover instead is evaluated, so that the mobile is again served by the lowest loaded RAT.

Figure 262 and Figure 263 present the Block Error Rate (BLER) performance for uplink and downlink, respectively, obtained with the NCCB and the LB strategies in the two considered RATs. It can be clearly noticed that the proposed NCCB strategy achieves an important reduction in the BLER in both links for the CDMA-based RAT (i.e. UTRAN) with negligible increase in the BLER of the FDMA/TDMA-based RAT (i.e. GERAN). This reveals the effectiveness of the network-controlled cell-breathing concept, able to reduce the interference and therefore to improve the performance in CDMA.

The interference reduction in CDMA turns into a capacity increase, because, more users can transmit simultaneously, while on the other hand, transmissions are more effective since the BLER is lower. This is observed in Figure 264, which presents the comparison in terms of the total aggregated throughput in the scenario (i.e. including UTRAN and GERAN). The benefits of the proposed strategy can be clearly observed with throughput increases of around 13% in the uplink and 24% in the downlink. Notice in Figure 265, which presents the uplink throughput in each RAT, that the capacity gain comes from UTRAN, while the throughput in GERAN remains similar to the one obtained with the LB strategy.

Similarly, Figure 266 illustrates the improvements obtained in terms of dropping and blocking probabilities, respectively, showing the better efficiency of the NCCB strategy to allocate the available resources in the two RATs among the different users.

Finally, in order to complete the study with some considerations regarding the required signalling, Figure 267 plots the ratio of vertical handovers per second with the NCCB and the LB strategies. Notice that LB requires less vertical handover executions because in this case the decision to make a VHO is only checked when a horizontal handover has been decided in the current RAT (a part from the cases when a VHO is triggered because a dropping is about to occur), while in NCCB the VHO is triggered when crossing the limits imposed by PLth.

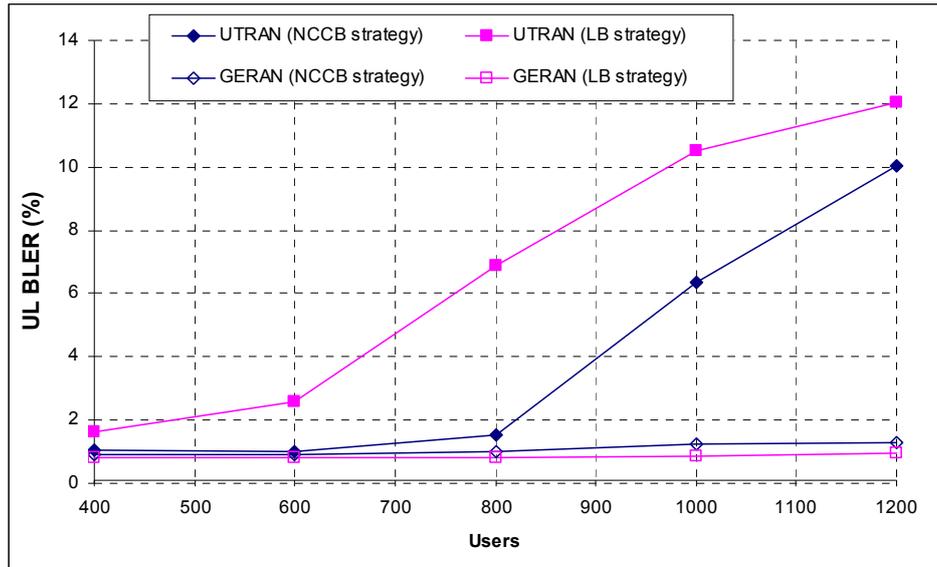


Figure 262 Uplink BLER in UTRAN and GERAN

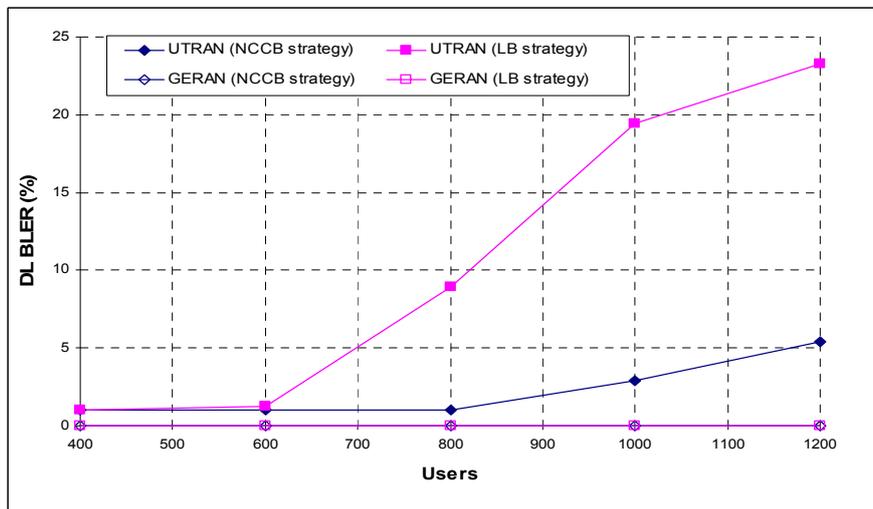


Figure 263 Downlink BLER in UTRAN and GERAN

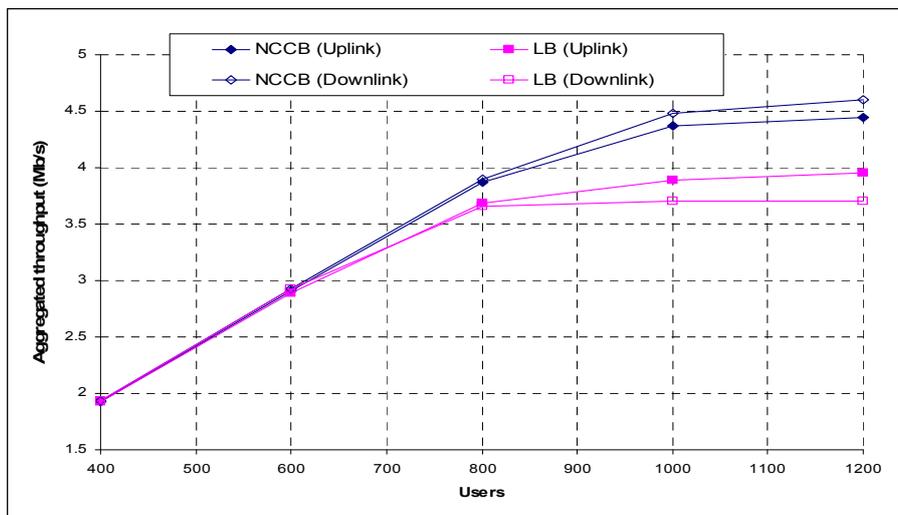


Figure 264 Total throughput in uplink and downlink

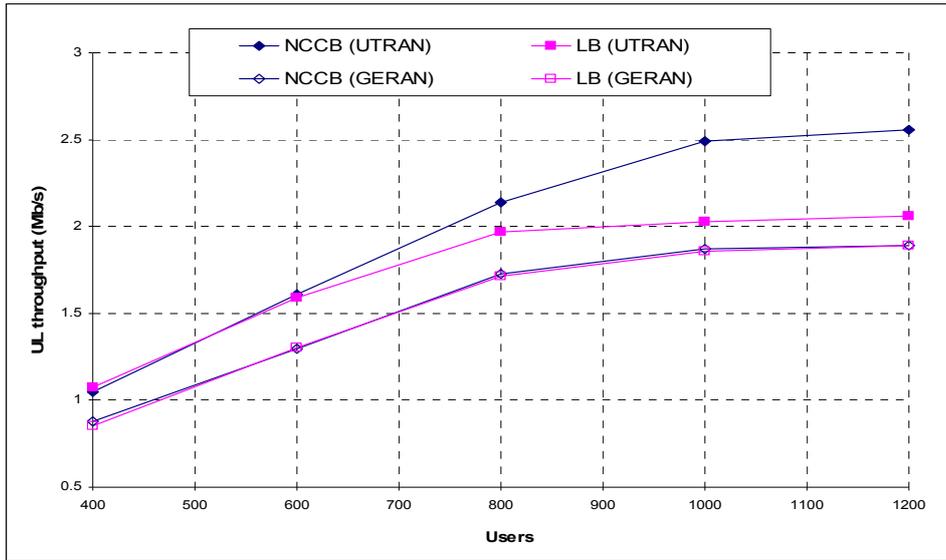


Figure 265 Uplink throughput per RAT

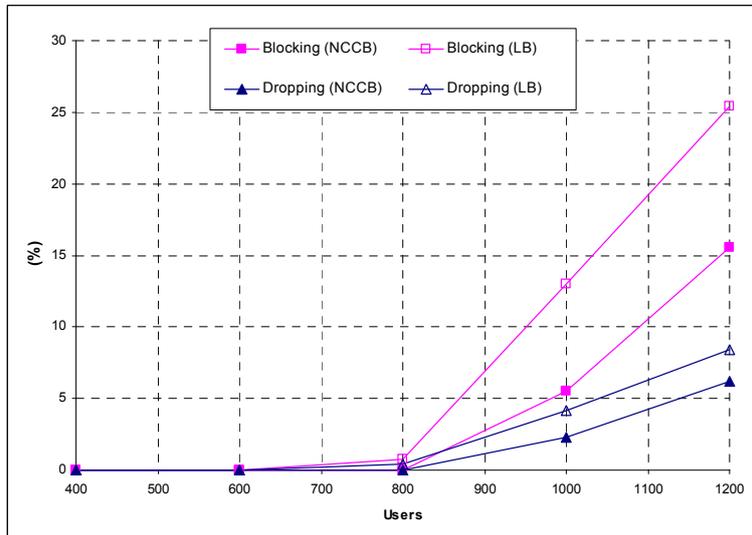


Figure 266 Blocking and dropping probabilities

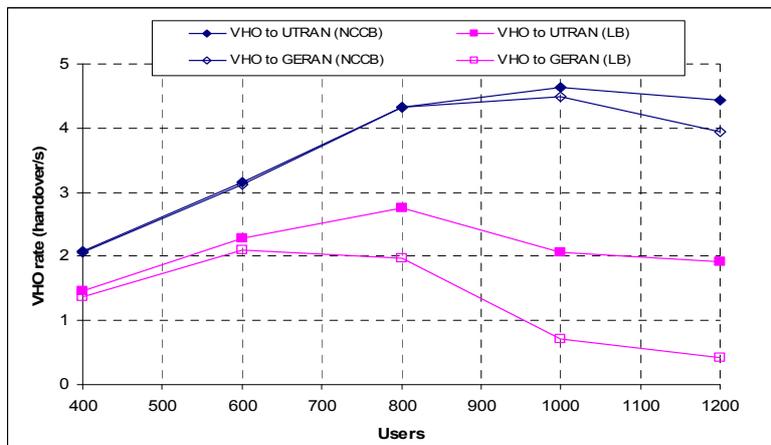


Figure 267 Vertical Handover rate

5.5.3.2.2 Setting of PLth

Clearly, one of the critical parameters to be set in the NCCB algorithm is the path loss threshold, PLth. The setting of this parameter impacts over two different effects. On the one hand, it allows controlling the effective cell radius (i.e. the cell breathing) of the CDMA-based RATs, and, therefore, low PLth values will tend to reduce the CDMA interference thus improving the performance of users connected to these RATs. On the other hand, it also controls the traffic distribution between the considered RATs, in the sense that low PLth values will tend to increase the number of users allocated to GERAN while high values will tend to reduce these number of users and to allocate more users in UTRAN. Consequently, the setting of the path loss threshold PLth results from the trade-off between how much the CDMA interference can be reduced while avoiding an excessive load unbalance.

To illustrate these effects, the path loss statistical distribution in the scenario has been firstly obtained, as shown in Figure 268. Then, three different representatives values of PLth have been selected, $PLth = \{115\text{dB}, 120\text{dB}, 125\text{dB}\}$, corresponding, approximately to the 40-th, 60-th and 80-th percentiles of the path loss distribution, respectively. As a reference, notice that setting PLth to the x-th percentile means that x% of the users will perceive a path loss below PLth and, therefore, they would be allocated in UTRAN, whereas (100-x)% of the users would be allocated in GERAN, provided that the load level is low enough to avoid blocking in one RAT. In practice, for relatively heavy load conditions causing overflow in one RAT as well as due to the system dynamics like e.g. hysteresis margins, the traffic splitting between RATs may exhibit a different distribution.

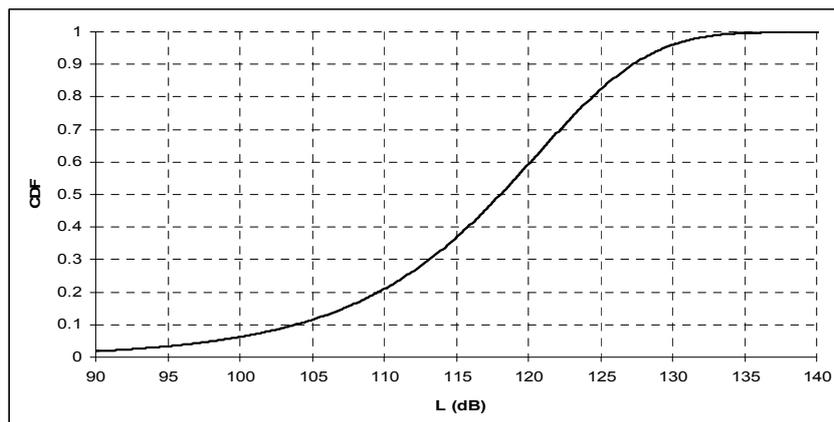


Figure 268 Cumulative Distribution Function (CDF) of the total path loss

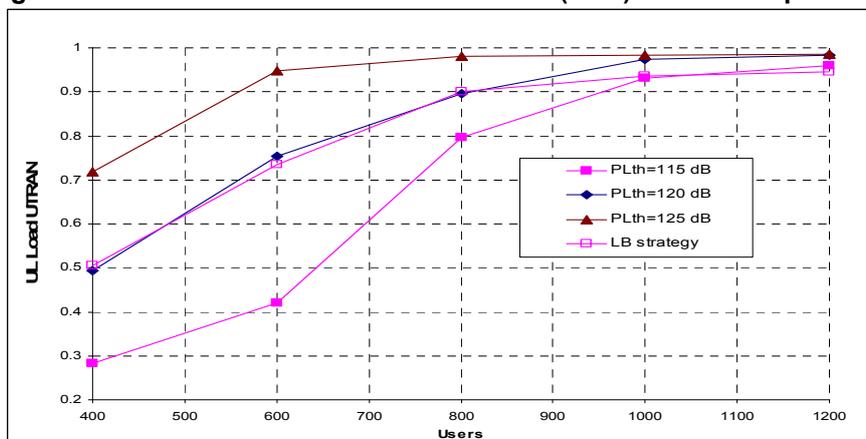


Figure 269 Uplink load in UTRAN

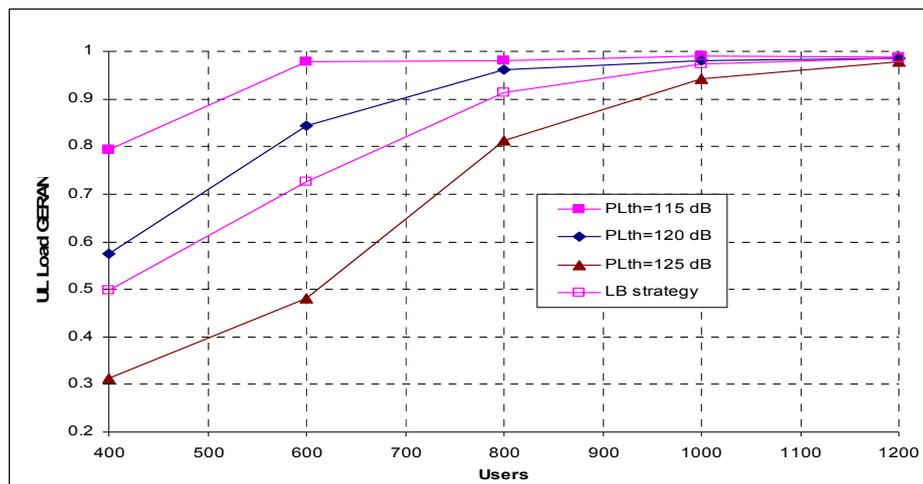


Figure 270 Uplink load in GERAN

From the point of view of load distribution in the two RATs, Figure 269 and Figure 270 plot the average uplink load in UTRAN and GERAN, respectively, for the NCCB strategy with the different PLth values and for the LB strategy. Clearly, the case PLth=120 dB achieves the better load balancing between both RATs, while for PLth=115 dB there is a higher load in GERAN and for PLth=125 dB the load is higher in UTRAN. Then, NCCB algorithm with PLth=120 dB achieves a load distribution very similar to the LB case, so that with this setting of the parameter load balancing considerations are also included in the NCCB algorithm.

From the performance point of view, Figure 271 plots the uplink BLER in UTRAN for different PLth values as well as for the LB strategy. Downlink results exhibit a similar behaviour and are not presented for the sake of brevity. Clearly, the value PLth=115 dB achieves the highest BLER reduction thanks to the smaller effective cell radius of UTRAN when compared to the other cases. Also the value PLth=120 dB achieves a performance closer to the 115 dB case, and much better than for the load balancing case, in which some high path loss users may be allocated in UTRAN at the cell edge, thus increasing the intercell interference and requiring higher power levels. In turn, for the case PLth=125 dB, there is a high degradation, even worse than with the LB strategy. The reason is two-fold: on the one hand, the higher effective UTRAN cell radius resulting in this case provides a lower intercell interference reduction. In addition, the traffic distribution is such that a high number of users are allocated in UTRAN, thus increasing the load existing in this RAT (see Figure 269), resulting in a higher intracell interference. Notice that, although the load values in both UTRAN and GERAN were similar for NCCB with PLth=120 dB and LB, there is a high difference in the performance in terms of BLER, revealing that for a proper CRRM not only the load balancing concept is important but also the way how the traffic is distributed among RATs may have a high impact. This conclusion can also be observed for the downlink direction in Figure 272, which plots the distribution of the total downlink transmitted power in UTRAN for the different cases. Clearly, the highest power is required for PLth=125 dB, in which most of the users are in UTRAN, while the lowest power is obtained with PLth=115 dB.

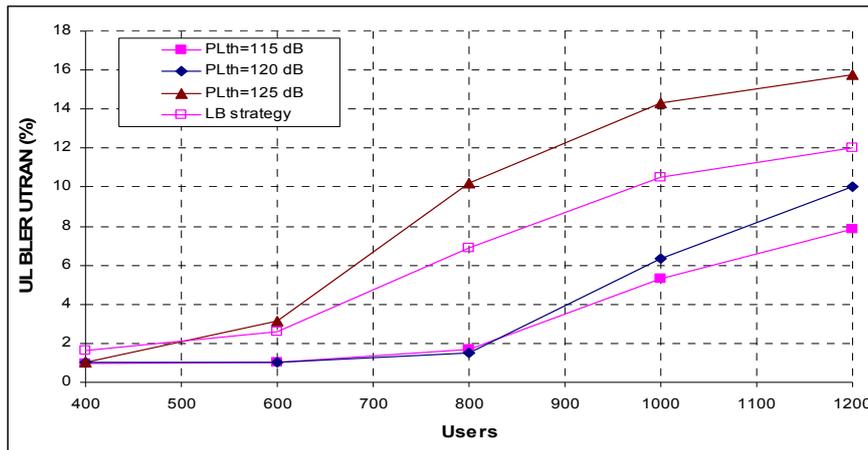


Figure 271 Uplink BLER in UTRAN

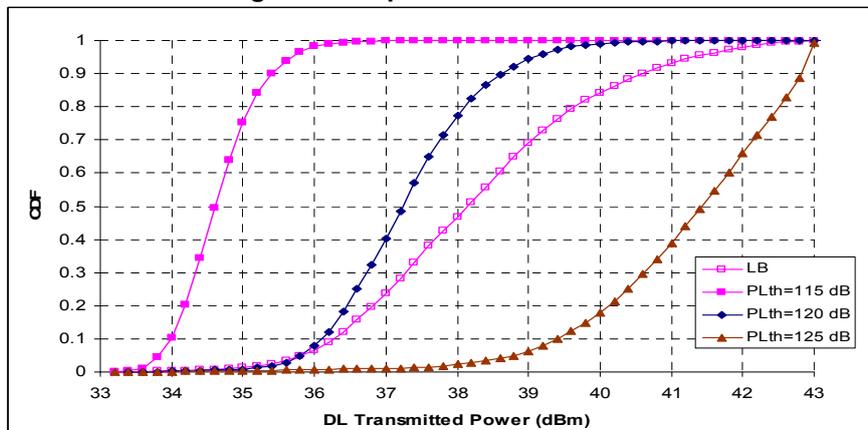


Figure 272 CDF of the downlink transmitted power in UTRAN with 600 users

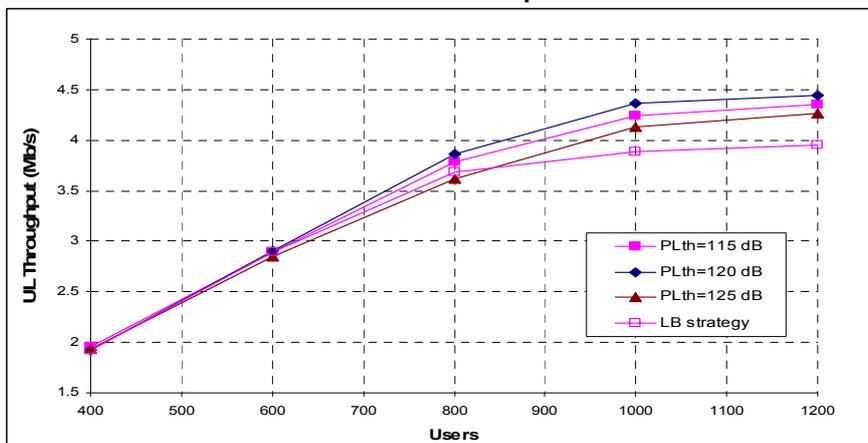


Figure 273 Uplink throughput for different values of PLth

The total aggregated throughput (i.e. including UTRAN and GERAN) is depicted in Figure 273. The highest throughput is provided by PLth=120dB, revealing to be the most suitable solution. Compared to a pure LB, the achieved gain can be up to about 12% for heavy load conditions. The origin of the gain comes from the fact that NCCB with PLth=120dB also achieves load balancing between RATs through a more intelligent and efficient user distribution, reducing the overall interference in the system. Compared to the other settings, i.e. PLth=115dB or 125 dB, the gain comes from the benefits of the better load balancing obtained with PLth=120dB.

5.5.3.2.3 Impact of indoor traffic

This section analyses the performance of the proposed NCCB CRRM strategy in a scenario including indoor traffic. Notice that the underlying concepts behind this strategy are similar to those that led to the definition of the Indoor policy presented in section 5.3, and consisting in allocating indoor users in GERAN and outdoor users in UTRAN under the rationale that UTRAN users suffer a higher degradation when they are indoor. As a matter of fact, the characteristic that makes the indoor condition to be undesirable for UTRAN is the high path loss suffered by indoor users, so by controlling through NCCB the maximum path loss (i.e. the effective cell radius) of UTRAN users, performance can be improved like with the Indoor policy. In addition to that, NCCB constitutes a practical way of achieving similar effects than the indoor policy, without requiring an explicit knowledge about whether a user is indoor or outdoor, simply by analysing path loss measurements instead.

In the following, the NCCB strategy performance is analysed in a scenario with cell radius 500m, including only voice service and with 30% of users located in indoor positions. The in-building penetration losses are 20 dB, which leads to the path loss distribution shown in Figure 274 according to the model presented in [149]. Then, the NCCB strategy with path loss thresholds $PL_{th}=110$ dB (50-th percentile) and $PL_{th}=125$ dB (80-th percentile) will be compared against the indoor policy and the load balancing criterion.

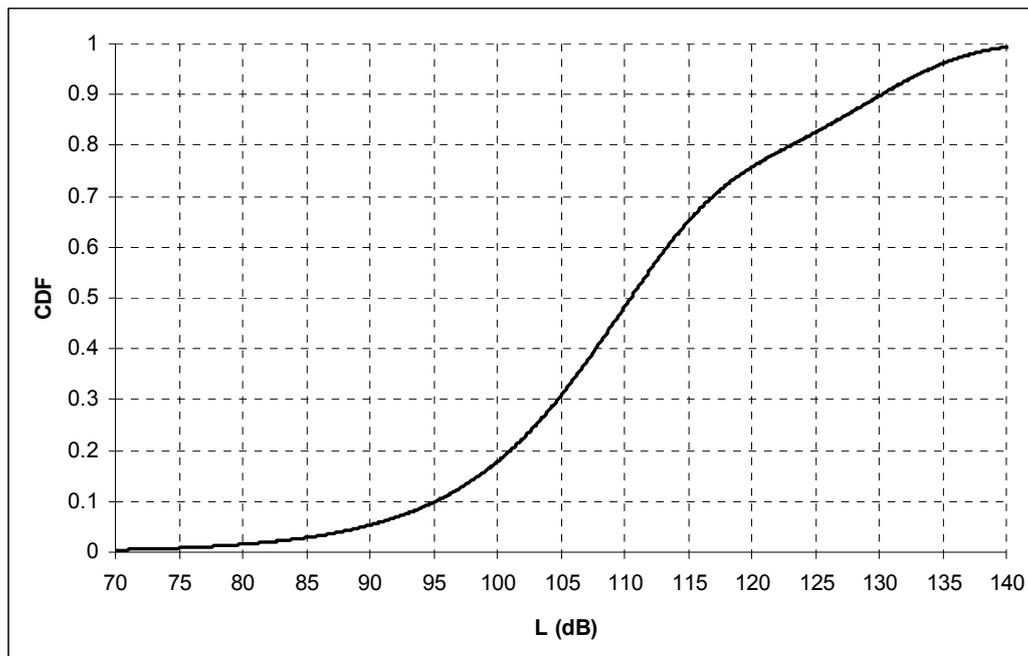


Figure 274 Path loss distribution with indoor traffic

Figure 275 and Figure 276 illustrate the performance in terms of BLER for the considered strategies in uplink and downlink, respectively. Clearly, the load balancing strategy offers the worst performance for low loads, because it may allocate some high path loss users (i.e. indoor) to UTRAN, which may turn into power limitations in the uplink direction. This situation is avoided by both the Indoor and the NCCB strategies. For high loads, a proper selection of PL_{th} around the 50-th percentile turns into a better user distribution between UTRAN and GERAN, achieving even a better performance than with the indoor policy. The reason is that with the indoor policy some low path loss users (i.e. indoor located close to the base station) may be allocated to GERAN unnecessarily, while other outdoor high path loss users (i.e. outdoor located far from the base station) may be allocated in UTRAN, while the NCCB avoids such a situation. It can be also observed that if the value of PL_{th} is not properly set (e.g. $PL_{th}=125$ dB) the load in UTRAN may be excessive and therefore the performance is

degraded for high loads. Figure 277 and Figure 278 illustrate similar results in terms of total aggregated throughput for uplink and downlink, respectively, revealing that the best performance is achieved with NCCB and PLth=110 dB.

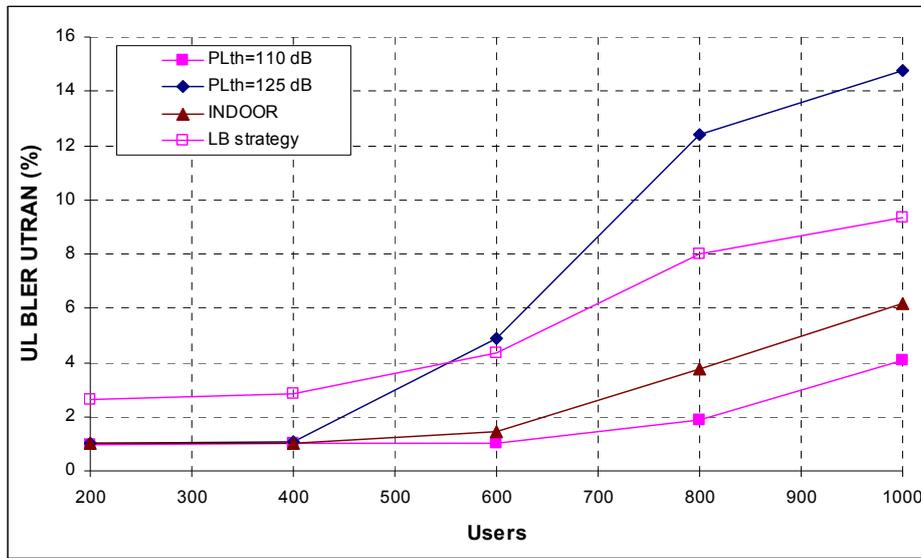


Figure 275 Uplink BLER in UTRAN

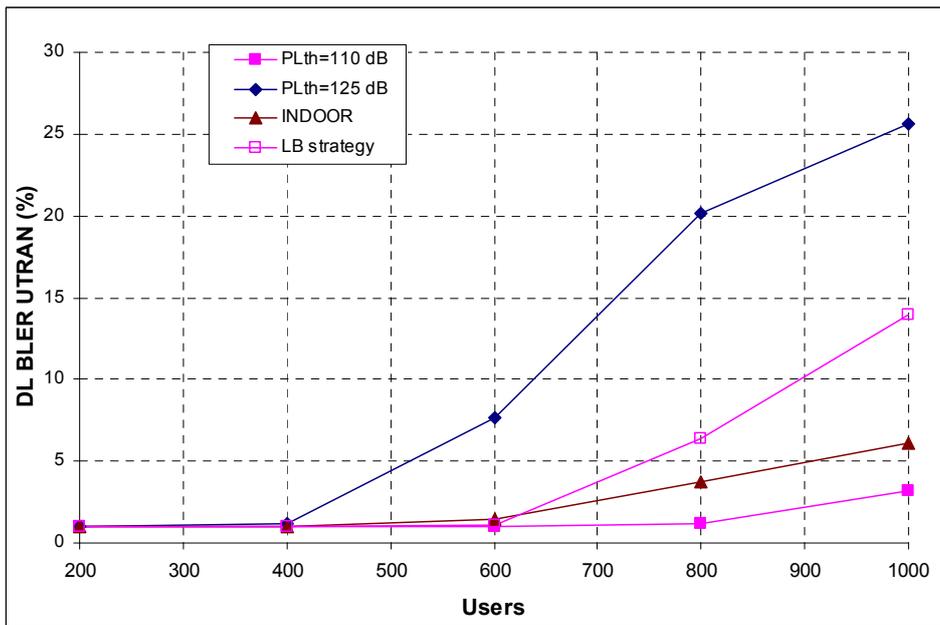


Figure 276 Downlink BLER in UTRAN

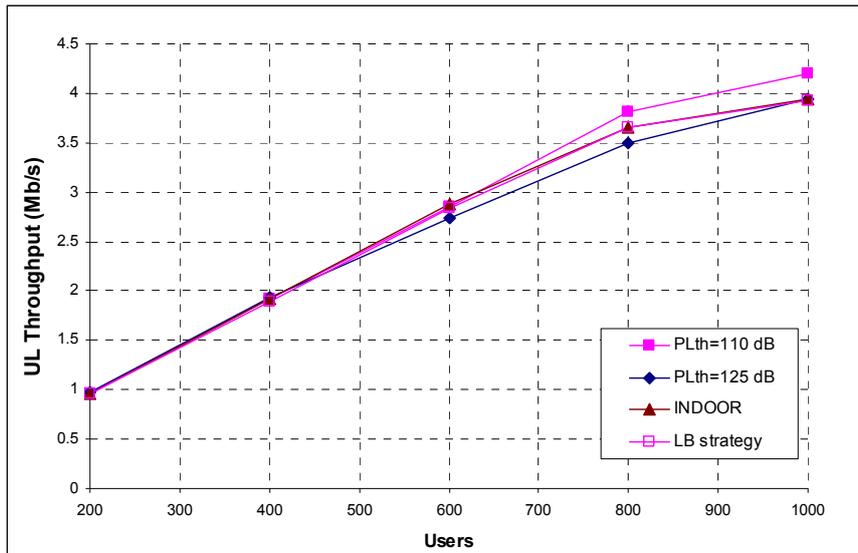


Figure 277 Uplink aggregated throughput

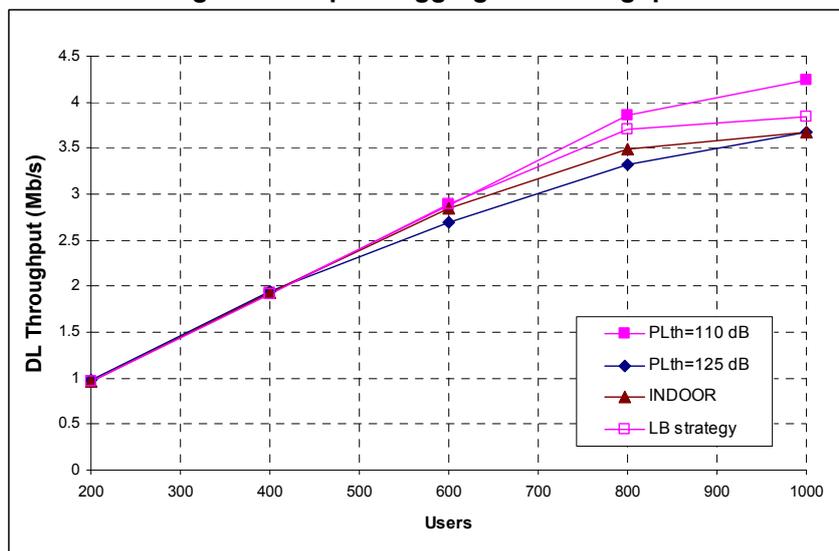


Figure 278 Downlink aggregated throughput

5.5.3.3 Multi-service scenario

The previous analysis has revealed that, in a scenario with only voice users, the NCCB CRRM strategy can provide significant capacity improvements thanks to a proper distribution of the users in the CDMA and FDMA/TDMA RATs according to the measured path loss. In the following, the NCCB strategy is extended to a scenario where also www users exist. In this case, it should be considered that the allocation of www users to GERAN turned into some delay degradation for high loads, as observed in sections 5.3 and 5.4, mainly if there are a lot of voice users in GERAN. Consequently, the performance of the NCCB strategy should be compared against a service based policy like VG*VU explained in section 5.3. Furthermore, different possibilities arise depending on whether the NCCB is only executed for voice users or also for www users.

5.5.3.3.1 Comparison with service-based policies

This study is carried out in the same scenario with cell radius 1 km explained in section 5.3.2.1. For NCCB, PLth is set to 120 dB (60-th percentile). The following strategies are compared:

1) NCCB strategy: In this case, the CRRM strategy is based only on Path Loss, without taking into account service information. Consequently, the same RAT selection condition is applied to both voice and www users, as explained in sub-section 5.5.3.1. Also the T-VHO strategy is considered for vertical handover. Figure 279 shows the distribution of users between the two RATs for two different traffic loads (400voice+200www and 800voice+200www). For low traffic loads and the voice traffic distribution is around 60% in UTRAN (according to the percentile of PLth). For www, the distribution changes a bit for high traffic loads, mainly because the www sessions in GERAN tend to be longer than in UTRAN (because of the higher delay in GERAN for www users), which affects the user distribution in the two RATs.

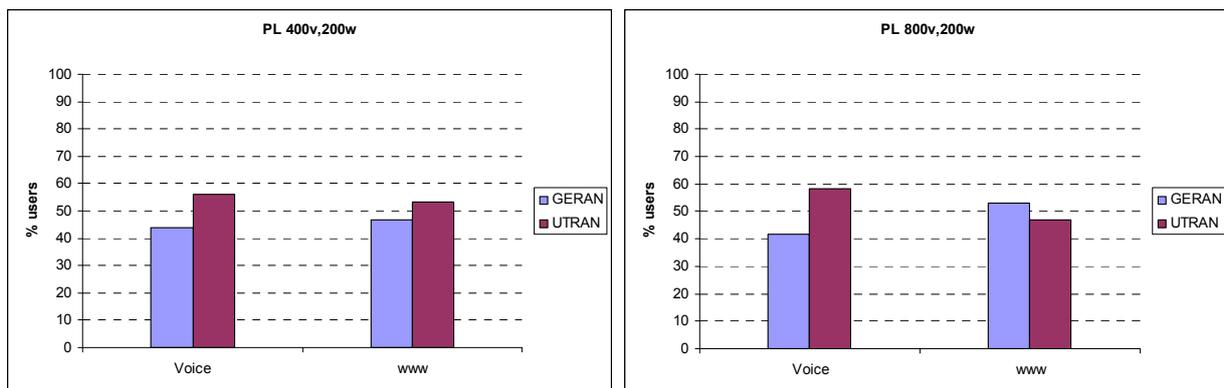


Figure 279 Traffic distribution between UTRAN and GERAN for two different traffic loads with the NCCB strategy

2) NCCB_voice strategy: In this case, the NCCB strategy is applied only for Voice users, while WWW users follow the VG*VU service-based policy (i.e. they are allocated in UTRAN provided that there is enough capacity). The T-VHO approach is also considered for vertical handover. Figure 280 illustrates the traffic distribution for the same traffic loads than Figure 279. Clearly in this case, the traffic distribution reveals that www is served through UTRAN while voice traffic is more equally distributed between UTRAN and GERAN depending on path loss.

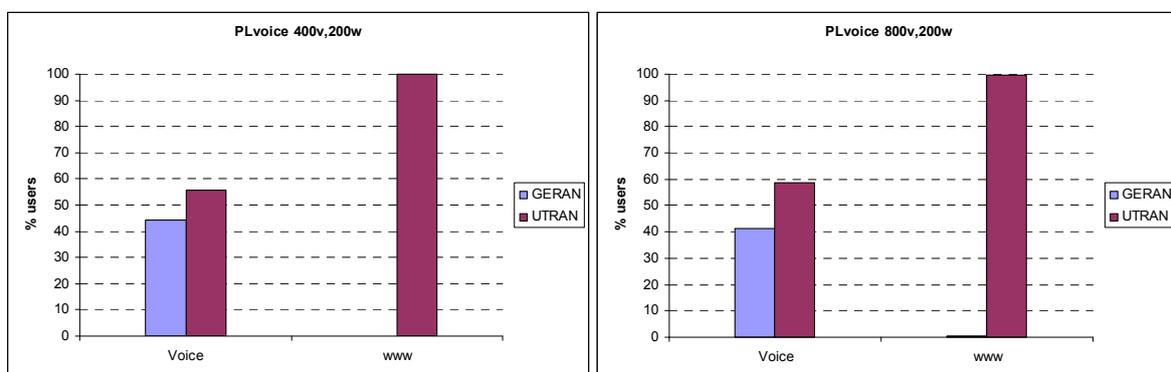


Figure 280 Traffic distribution between UTRAN and GERAN for two different traffic loads with the NCCB_voice strategy

3) VG*VU strategy: In this case, Voice users are served through GERAN and www users through UTRAN, provided that there is enough capacity available in the selected RAT. The T-VHO approach is applied.

Figure 281 shows the traffic split between the two RATs, and reveals that for low loads the VG*VU policy is clearly followed while for high loads half of the voice traffic is directed to UTRAN.

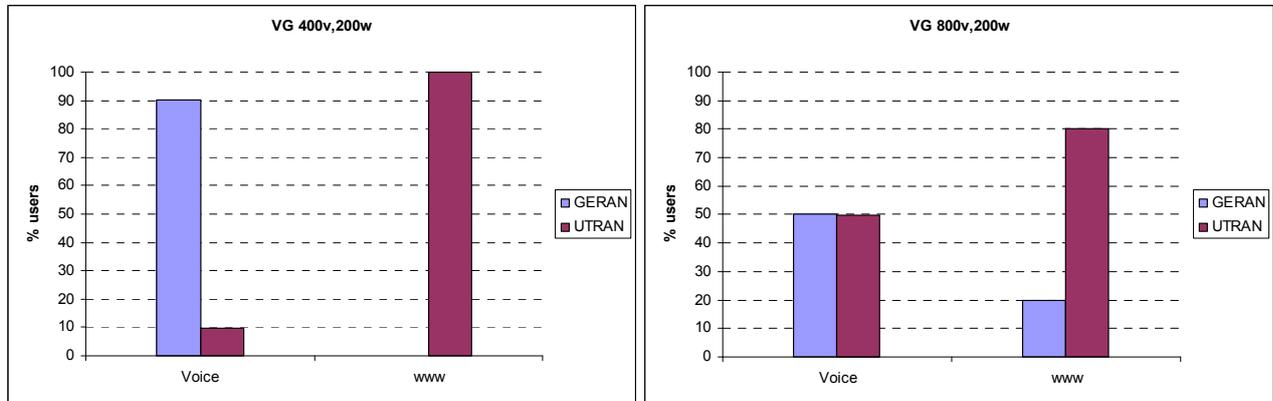


Figure 281 Traffic distribution between UTRAN and GERAN for two different traffic loads with the VG*VU strategy

4) VG_NCCB strategy: This case corresponds to the combination between NCCB and VG*VU, so that for low path losses the VG policy is applied while for high path losses users are allocated to GERAN. The corresponding flow diagrams for the initial RAT selection and the VHO conditions are shown in Figure 282 and Figure 283, respectively. The T-VHO vertical handover strategy is also considered. Figure 282 illustrates the traffic distribution in this case. It shows that, for low loads, voice traffic is mainly served through GERAN while www traffic is shared between UTRAN and GERAN (depending on path loss). Notice that the distribution of www users in GERAN/UTRAN is not 40/60% like in NCCB strategy. The reason is that in GERAN it takes longer to send the packets because of the high load, so that the duration of the sessions are longer in GERAN than in UTRAN, which affects the traffic distribution among the RATs. Notice that this effect is more noticeable when increasing the load, because in this case it takes even longer to complete the sessions for users in GERAN.

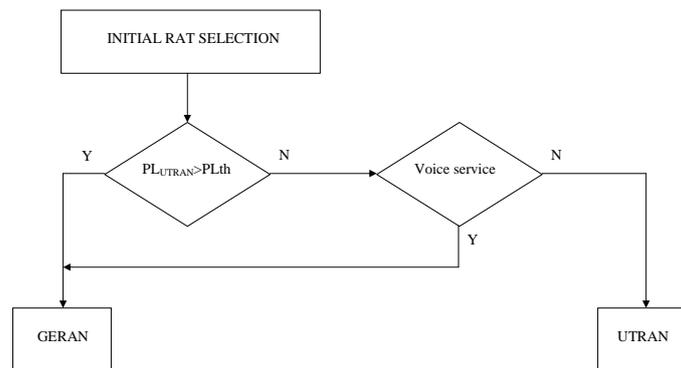


Figure 282 Initial RAT selection with the VG_NCCB strategy

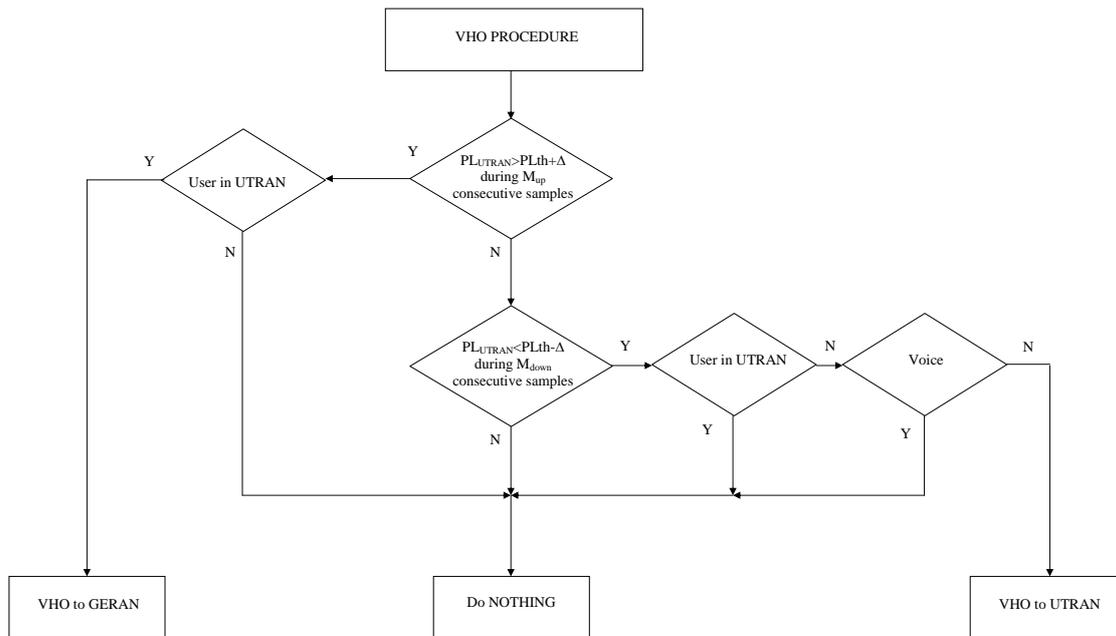


Figure 283 VHO with the VG_NCCB strategy

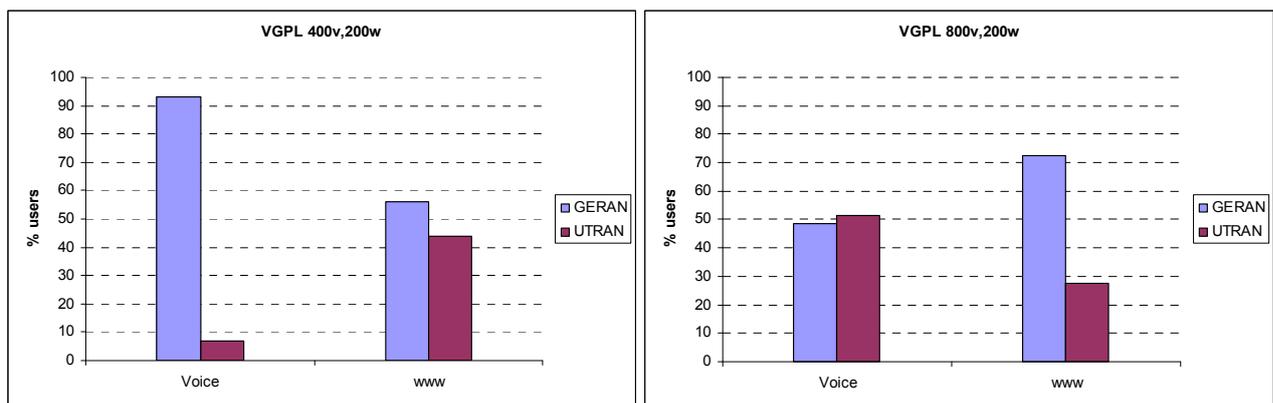


Figure 284 Traffic distribution between UTRAN and GERAN for two different traffic loads with the VG_NCCB strategy

Figure 285 and Figure 286 plot the corresponding load factors in UTRAN and GERAN as a function of the number of www users for two different voice traffic loads, namely 400 and 800 users, respectively. For the 400 voice case, Figure 285 shows that NCCB and NCCB_voice achieve a better load balancing among the two RATs, so that there is more room in GERAN to accommodate www users and therefore their delay will not increase so much (see the packet delay in Figure 287). In turn, with the service-based strategies VG*VU and VG_NCCB, GERAN is overloaded. For the 800 voice users case shown in Figure 286 the load increases significantly in both RATs for all the cases. Notice also that in all the cases, the highest load factor in UTRAN occurs with NCCB_voice because it is the strategy with higher number of voice and www users in UTRAN (i.e. all voice users with path loss below PLth and all the www users will mainly be allocated in UTRAN, as shown in Figure 280).

The results in terms of www packet delay are shown in Figure 287 and Figure 288 for the cases with 400, 600 and 800 voice users in the scenario. The results of the VG_NCCB strategy are not shown in the graphs because the average delay reaches very high values (around 10s) in this case. The reason is that, with this strategy, GERAN is completely overloaded because it contains the voice users (according to VG policy) as well as the www users with path loss above PLth. As a result, these www users experience very few

opportunities to transmit their packets, turning into a very significant delay degradation. With respect to the rest of policies, clearly the pure service-based policy VG*VU achieves the lowest delay because in this case www users are mainly served through UTRAN. In turn, the performance with NCCB is worst because of the www users that are served through GERAN, which experience a higher delay, mainly for the 600 and 800 voice users cases (due to the large load existing in GERAN). Finally, NCCB_voice achieves a delay performance closer to VG*VU for the cases of 400 and 600 voice users because www users are mainly served through UTRAN. For the 800 voice users case, and for high www levels, the delay of NCCB_voice reaches closer values to NCCB, because of the larger load factor existing in UTRAN (see Figure 286).

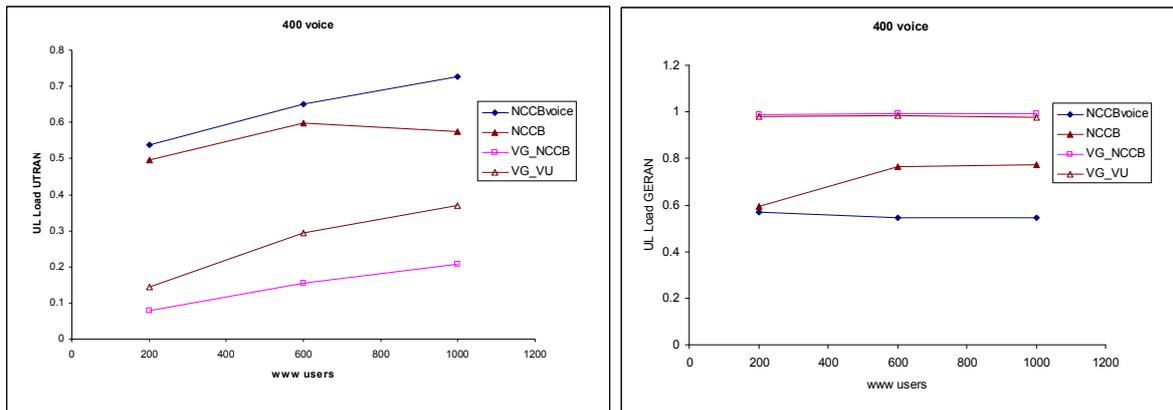


Figure 285 Load in UTRAN and GERAN for the 400 voice users case with the different strategies

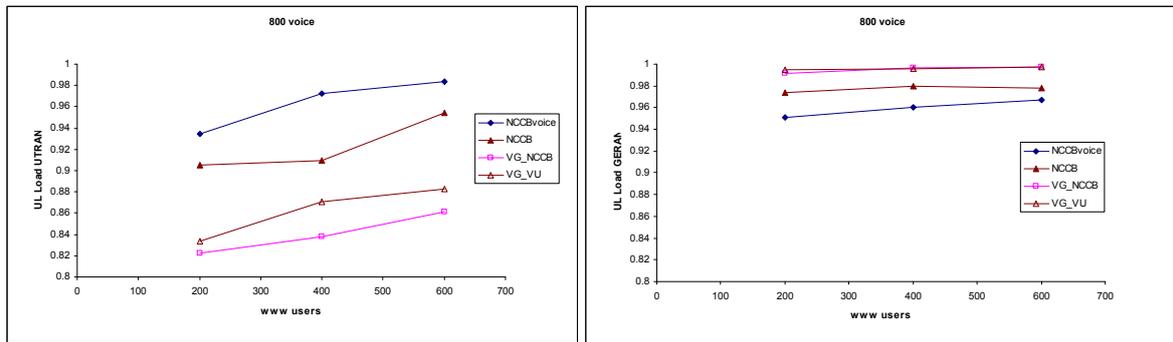


Figure 286 Load in UTRAN and GERAN for the 800 voice users case with the different strategies

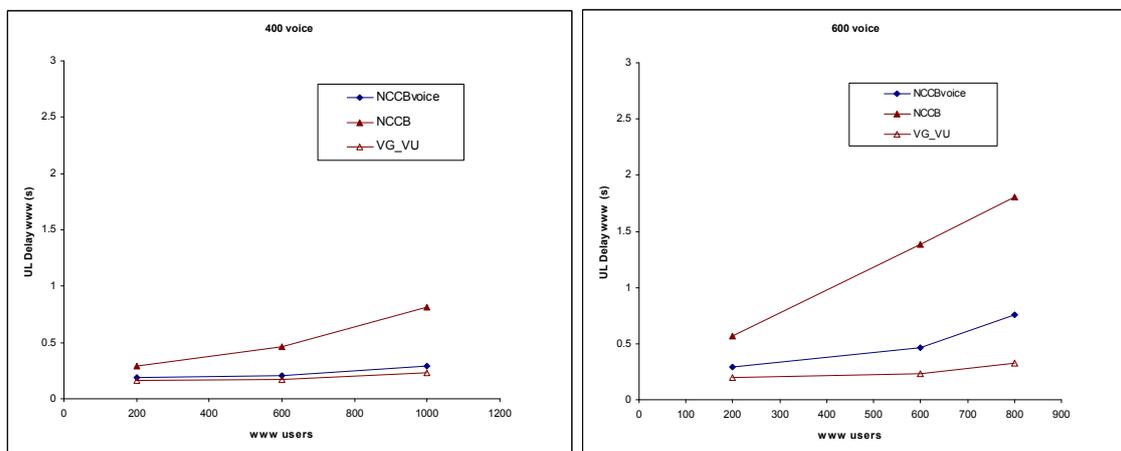


Figure 287 UL average packet delay for www users for the cases with 400 and 600 voice users

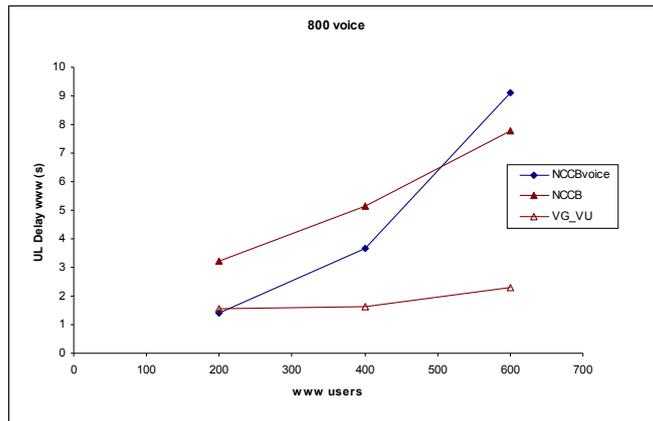


Figure 288 UL average packet delay for www users for the cases with 800 voice users

Figure 289, Figure 290 and Figure 291 plot the BLER for voice users in UTRAN and GERAN as a function of the number of www users for the cases with 400, 600 and 800 voice users, respectively. While the differences in GERAN are small, the worst performance in UTRAN is achieved with the VG*VU policy, which suffers a high BLER degradation for large path losses. Some BLER reduction is achieved with the VG_NCCB strategy, which reduces the load factor in UTRAN (see Figure 285 and Figure 286) by allocating high path loss www users to GERAN. In turn, the NCCB and NCCB_voice strategies achieve the best performance thanks to allocating the high path loss users in GERAN. In this case a slightly worst behaviour is observed with NCCB_voice because of the larger load factor in UTRAN. For the 800 voice users case, the large load factor existing in UTRAN with NCCB_voice (see Figure 286) causes a degradation in terms of BLER for high www loads. However, as it will be shown in the next sub-section, this effect can be smoothed by a proper setting of the PLth threshold.

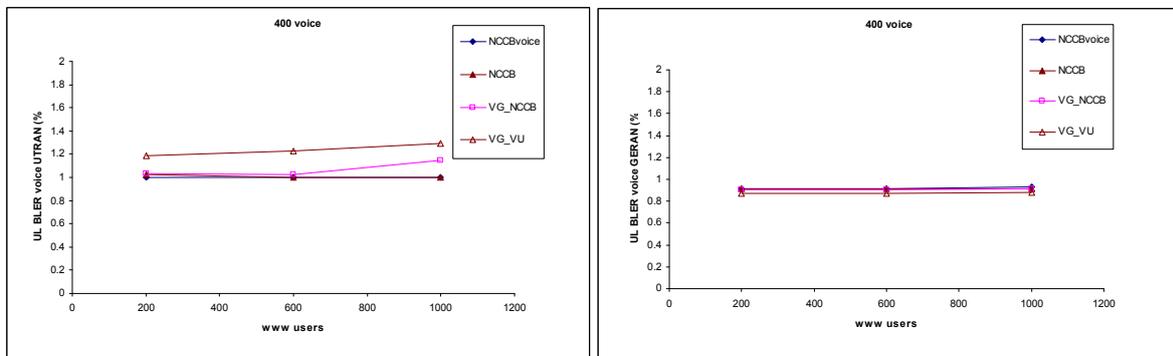


Figure 289 UL BLER in UTRAN and in GERAN for the 400 voice users case

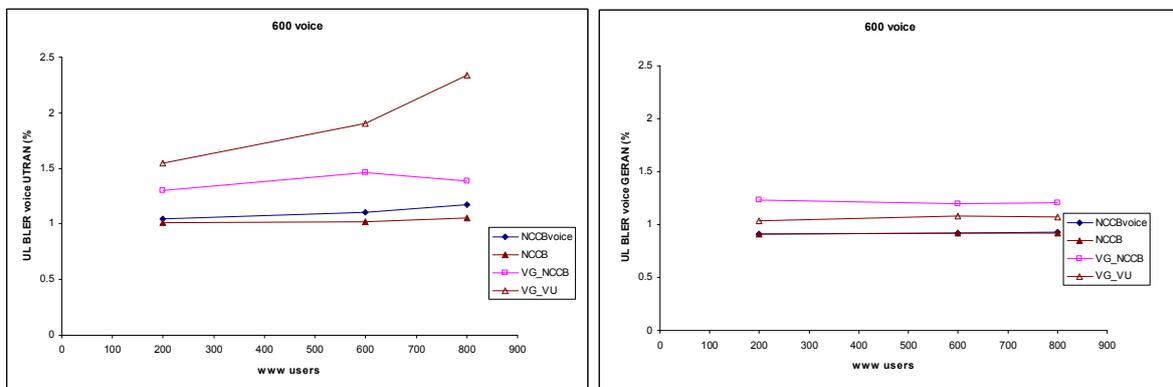


Figure 290 UL BLER in UTRAN and in GERAN for the 600 voice users case

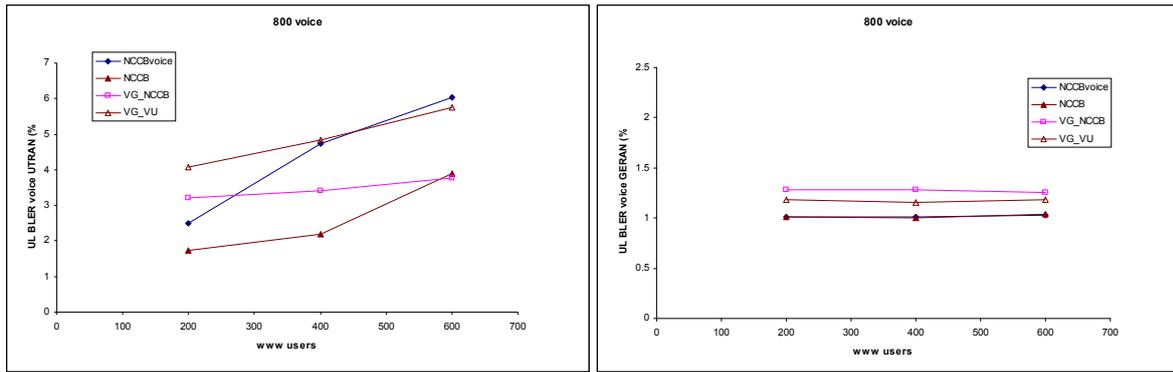


Figure 291 UL BLER in UTRAN and in GERAN for the 800 voice users case

Figure 292 plots the aggregated UL throughput for the different strategies in the 400 and 800 voice users case. Clearly, the worst throughput is achieved with the VG_NCCB strategy, mainly due to the large delay of www users with high path losses allocated in GERAN, which at the same time is overloaded due to the VG policy. With respect to VG*VU it achieves a good performance for the 400 voice users case. However, in the 800 voice users case, there is some throughput reduction because of the high BLER of the voice users in UTRAN, together with the larger load existing in GERAN (see Figure 286), which gives less room to make handovers and therefore it causes some droppings and blockings, as shown in Figure 293. Finally, the best performance is achieved by NCCB and NCCB_voice strategies, thanks to the better user distribution according to the path loss and the better load balancing existing between UTRAN and GERAN. The differences between NCCB and NCCB_voice are very small, and arise mainly for large numbers of voice and www users, due to the larger load factor existing in UTRAN.

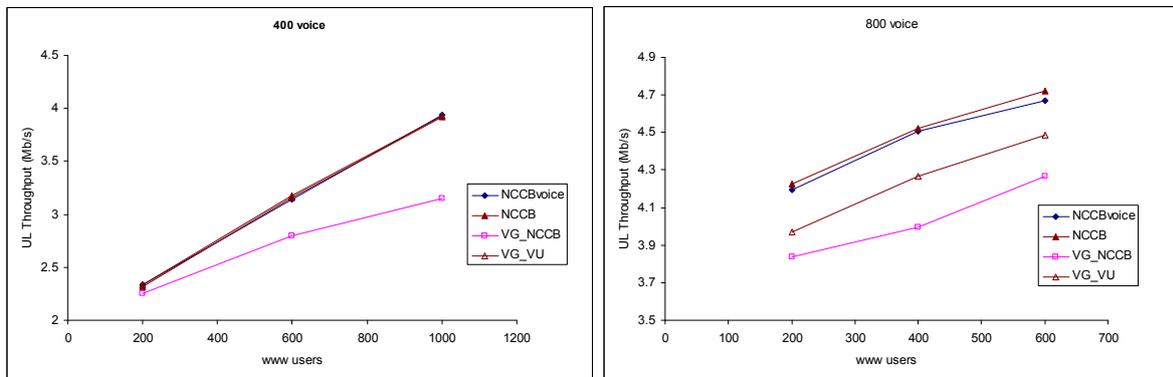


Figure 292 UL Aggregated throughput for the different policies

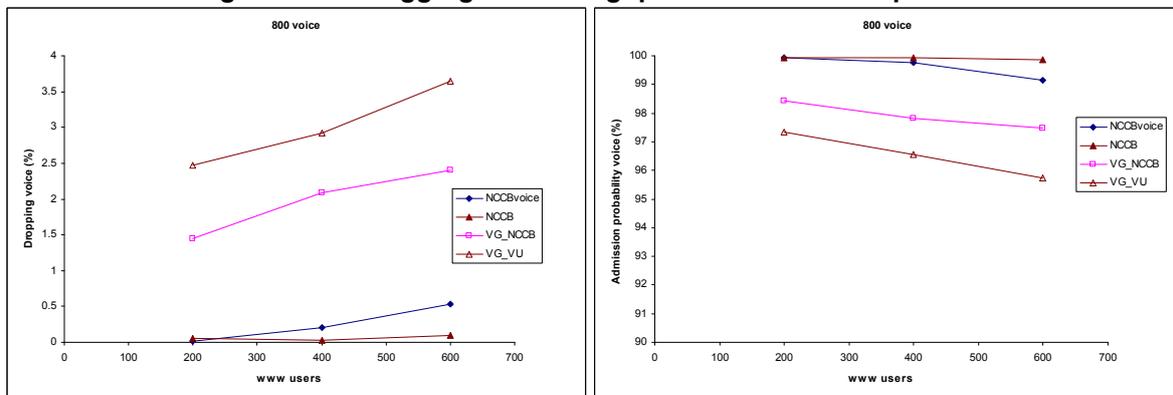


Figure 293 Dropping and admission probability for voice users

Figure 294 and Figure 295 plot the vertical handover ratio for voice and www users, respectively. For voice users, as shown in Figure 294, in all the cases VG*VU provides the lowest rate because there are no VHO due to path loss reasons. In turn, the highest rate is obtained with NCCB and NCCB_voice, which originate VHO when users cross the path loss regions. In turn, VG_NCCB has a lower increase with respect to VG*VU it only triggers a VHO to GERAN for voice users when going from a region with low path loss to another one with high path loss (see Figure 283). As a result of that, the increase is more noticeable in the VHO to GERAN case than in the VHO to UTRAN case.

With respect to www users, shown in Figure 295, the lowest VHO rate is again obtained with VG*VU. Then, the NCCB_voice strategy, which tends to affect only voice users, produces also a very low handover rate. The largest increase is obtained with VG_NCCB and NCCB, because they are the policies that force VHO for www users when crossing the path loss limits. Between these two, the differences come from the fact that with VG_NCCB there is more saturation in GERAN, so that a horizontal handover failure occurs more likely for users with an established TBF that cannot keep the TBF in the new cell.

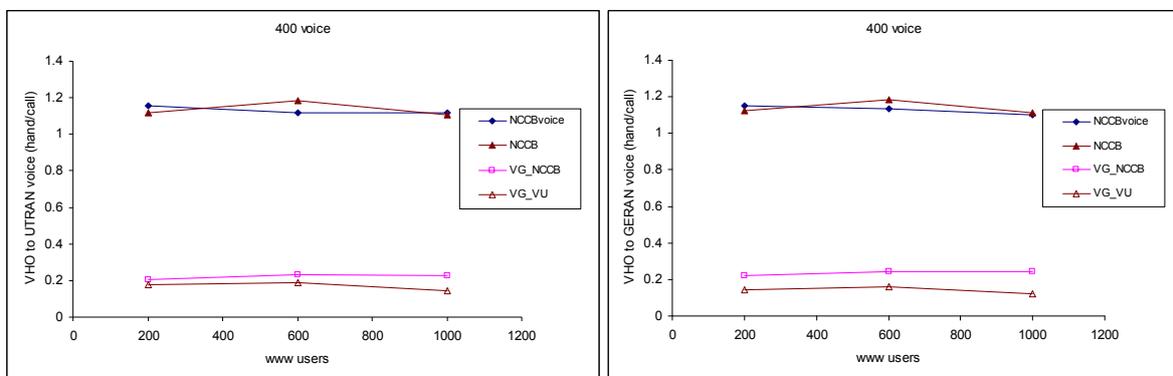


Figure 294 Vertical Handover Rate for voice users

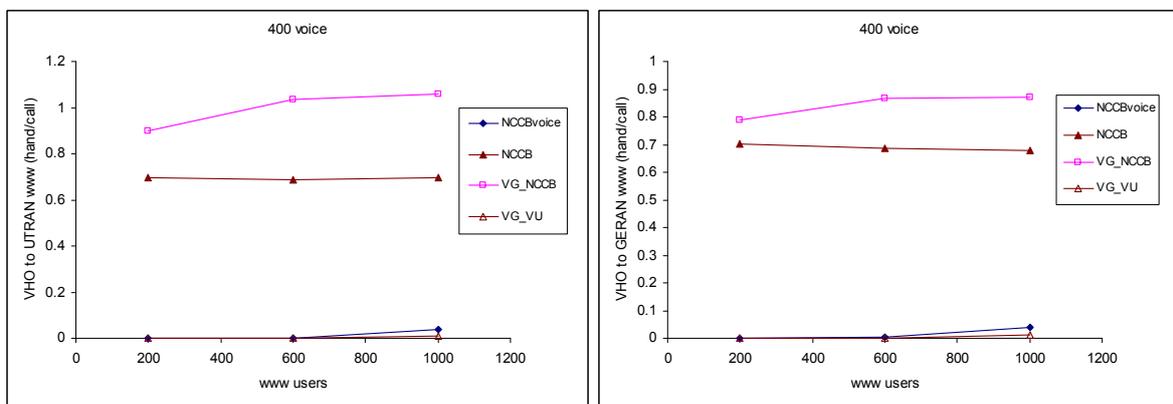


Figure 295 Vertical Handover Rate for www users

As a general conclusion of this section, it has been observed that from the delay point of view the best performance is achieved by the pure service policy VG*VU, while from the point of view of BLER and throughput, the best performance is achieved by NCCB and NCCB_voice, and the latter performs better in terms of delay and requires a lower number of VHOs. Consequently, NCCB_voice appears to have the best behaviour among the considered strategies.

5.5.3.3.2 Setting of PLth

After having concluded in the last sub-section that it seems appropriate to execute NCCB_voice strategy, this section is devoted to establish the appropriate setting of the path loss threshold PLth to be used in the algorithm. To this end, some simulations have been performed by considering different values of PLth according to the path loss distribution. They are PLth=110 dB (20-th percentile), PLth=115 dB (40-th percentile), PLth=120 dB (60-th percentile) and PLth=125 dB (80-th percentile).

Figure 296 plots the BLER in UTRAN of voice users for the considered situations and for the cases of 400 and 800 voice users in the scenario. It can be observed that large values of PLth (e.g. 125 dB) increase a lot the BLER, due to the large number of voice users allocated in UTRAN. Notice that in this case, the load factor is higher and also users in UTRAN will experience higher values of the path loss. In turn, when reducing PLth, the number of voice users in UTRAN reduces, and also their path loss, while at the same time the load in GERAN increases, thus unbalancing the two RATs. If the number of users in the system is moderate (e.g. 400 voice users), reducing PLth can be beneficial. However, if PLth is very low (110 dB) and GERAN becomes very saturated (e.g. in the 800 voice users case) the NCCB policy cannot be properly followed which turns into some degradation in terms of BLER. Notice that the high load level in GERAN will originate that some high path loss users cannot be allocated in GERAN and they should be moved to UTRAN. Notice that reducing a lot PLth would lead to some type of service-based policy since all the voice users would tend to be allocated in GERAN.

Figure 297 plots the average packet delay of www users. According to the NCCB_voice strategy, www users are allocated mainly in UTRAN because the NCCB policy applies only to voice users. As a result, increasing the number of voice users in UTRAN (i.e. by increasing PLth) turns into a higher load factor that degrades the performance of www delay. So from the point of view of www users, it is better to have reduced values of PLth.

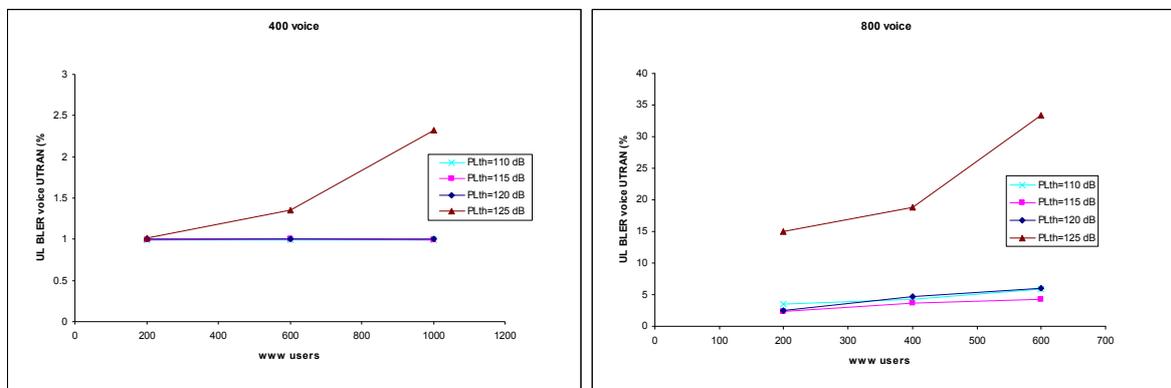


Figure 296 UL BLER in UTRAN for different values of PLth

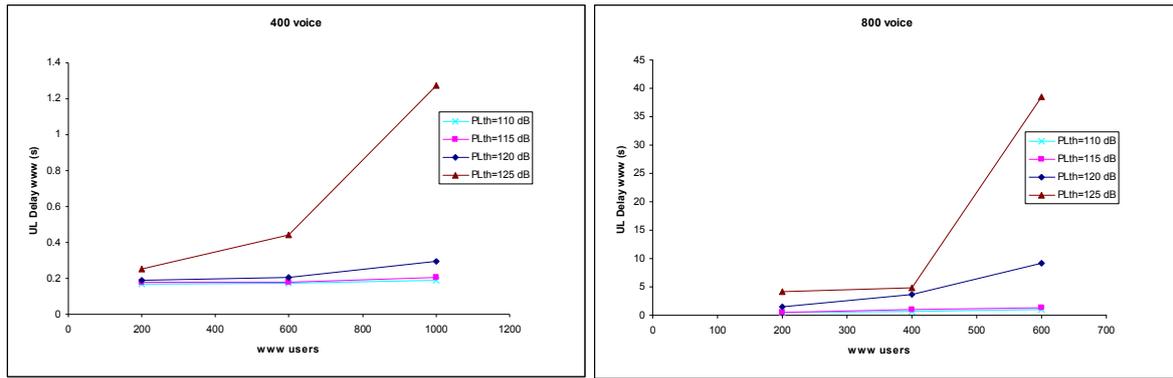


Figure 297 UL www packet delay for different values of PLth

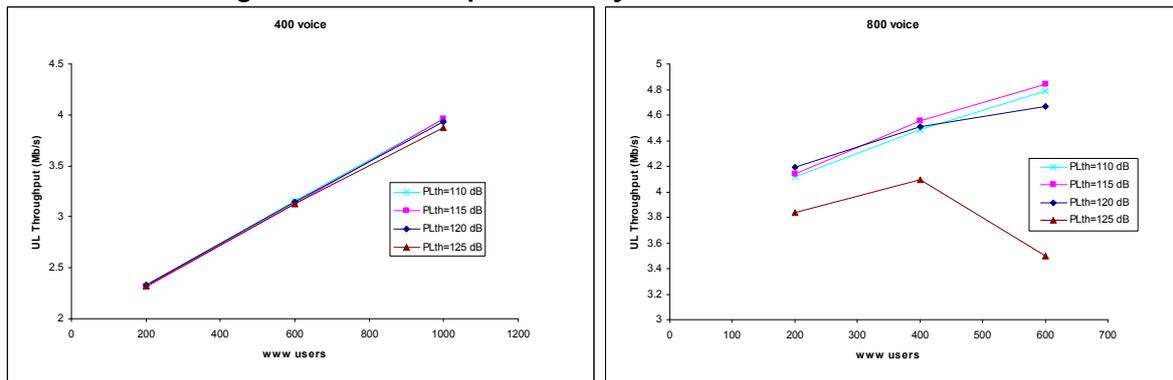


Figure 298 UL aggregated throughput for different values of PLth

Figure 298 plots the total aggregated throughput for different values of PLth and for two load levels, corresponding to 400 and 800 voice users. From the observation of this figure and the BLER and delay performances shown in Figure 296 and Figure 297, it can be concluded that the setting of the PLth threshold depends on the existing traffic mix in the scenario. Particular, the following cases can be distinguished:

- For moderate loads of voice and www users (e.g. 400 voice and 200 www) there are not significant differences among the values of PLth.
- For moderate voice load and high www load (e.g. 400 voice and 1000 www) users the best setting would be a value around PLth=110 dB, mainly because of the lowest delay achieved by www users, although the differences against PLth=115 dB are small.
- For high voice load and moderate www load (e.g. 800 voice and 200 www) the best performance is achieved with PLth=120 dB, like in the case with only voice users analysed in sub-section 5.5.3.2. This value achieves a good trade-off between balancing the load among the two RATs and at the same time distributing users according to the path loss.
- For high loads of voice and www users (e.g. 800 voice and 600 www) the best performance is achieved by PLth=115 dB, which provides a good trade-off between balancing the load among the RATs while at the same time keeping a low delay for www users.

5.5.4 Conclusions

This section has presented the network-controlled cell-breathing (NCCB) strategy for CRRM in heterogeneous TDMA/CDMA scenarios that achieves a reduction in the interference level of the CDMA system by controlling the its effective cell radius through initial RAT selection and vertical handover policies. The strategy has been evaluated by means of system level simulations in a scenario with UTRAN and GERAN as two examples of access networks using the CDMA and FDMA/TDMA technologies.

When considering a single voice service, the NCCB strategy has been compared against a classical load balancing strategy that tries to keep the same load level in both RATs. Results reveal that a significant improvement in terms of capacity for both uplink and downlink is achieved with the proposed strategy. It has been also shown that, by a proper setting of the maximum path loss PL_{th} , load balancing principles can also be achieved in NCCB, thus obtaining the benefits in terms of flexibility of load balancing while at the same time exhibiting a higher capacity than a pure load balancing.

When considering a mix between voice and www users, different combinations of the NCCB strategy with service-based policies have been tested. It has been observed that the best performance is achieved with the so-called NCCB_voice, corresponding to applying the NCCB strategy only to voice users and allocating www users in UTRAN according to the service-based policies. It has been also observed that the adequate setting of the threshold PL_{th} depends on the existing traffic mix and the trade-off among www delay and voice BLER.

5.6 RAT PRIORITY LIST-BASED RAT SELECTION

5.6.1 Introduction

This section provides some results in a scenario with a multiplicity of services, as corresponds to the traffic mix in the identified EVEREST scenarios in [24]. The considered algorithm intends to distribute the traffic between the different RATs following several targets and makes use of service-based policies in the form of a RAT priority list for each specific service, while at the same time it introduces load-balancing concepts. The objective of this analysis is the study of the algorithm performance in terms of capacity (number of attended users).

In the first subsection, the algorithm proposed initially in [1] and the subsequent modifications are described. Then, the different simulated strategies and the results obtained are presented. Finally, the conclusions are shown. On a second step, different strategies of load sharing are considered. The scenario described in [24], 2b, has been used, with business kind of users and the service mix defined for this kind of users.

5.6.2 Initial RAT Selection Algorithm

The algorithm that is being studied looks for an agreement between two main aspects:

- RATs preferences for the requested service, set by the operator according to the QoS offered by each RAT.
- Balance among the radio resource occupation of the different RATs, which is achieved by a suitable traffic load balance among them.

It is a particularization of the defined in section 5.2.5 of [1], just for UTRAN and GERAN.

Several parameters need to be defined before the explanation of the algorithm:

RAT Occupation Target

Network operator can set a target occupation for each RAT's radio resources (U_{GERAN} , U_{UMTS}).

Initial RAT selection algorithm shall distribute users between the different RATs trying to reach the target occupation established for each RAT. Target occupation is defined globally,

since it is related with all the bearers of each RAT. However, it is possible that the local resource usage doesn't meet the objective in some parts of the network.

In UMTS system, radio resources occupation in each cell can be measured with the uplink load factor, which is defined as:

$$FC_{UMTS} = (I_{inter} + I_{intra}) / (I_{inter} + I_{intra} + NoisePower) \quad (80)$$

where I_{inter} : intercell interference; I_{intra} : intracell interference.

UMTS network radio resource occupation will be calculated as the average of the uplink load factors of all the bearers in the network.

In GERAN system, radio resources usage can be measured as the percentage of timeslots used (*TSLutilization*) in the network. Timeslots may be used for voice and/or data.

GERAN network radio resource occupation will be calculated as the average of the timeslots utilization factors of all the bearers in the network.

RAT priorities list per service

A list composed by all the RATs capable of providing each service is defined. The network will try initially to provide the service through the first technology in the list. If it is not possible, the network will successively try with the rest of RATs in the list. In case of the list consists of just one technology, the service can only be provided by that technology.

Decision thresholds

Network operator defines decision thresholds (D_{GERAN} , D_{UMTS}) in order to decide RATs occupation intervals where the priority list is taken into account. When a RAT is much less loaded than the others, the service will be provided by that technology, regardless of the RAT priority list of the service.

Algorithm

The steps followed by the proposed Initial RAT Selection algorithm are listed below:

Step 1

The first step of the algorithm is to make a list of candidate serving cells for each technology involved in the heterogeneous network. These lists contain all the candidate cells capable of providing the requested service and they are ordered in such a way that the first cell is the one that better fulfils the requirements in each technology.

GERAN candidate cells are those that provide radio coverage at the point where the user is located in. In 2G the radio coverage is reached when the BCCH Rxlev value of the cell is greater than a threshold fixed by the network operator attending to its global quality objectives.

UMTS candidate cells have to fulfil two requirements:

- In the point where the user is located in, the E_c/N_t of the cell is greater than a specific value set by the operator attending to its global quality objectives.
- The transmission power requested to the user by the power control loop of the cell has to be lower than the maximum allowed value for that user.

Step 2

If there are only GERAN candidate cells and the service can be provided by this technology, GERAN will be the selected RAT. If GERAN can't provide the service then it will be rejected.

Step 3

If there are only UMTS candidate cells and the service can be provided by this technology, UMTS will be the selected RAT. If UMTS can't provide the service then it will be rejected.

Step 4

If there are candidate cells of different technologies:

- If service's RAT priority list only contains GERAN and the list of GERAN candidate cells is not empty, GERAN will be the selected RAT. If GERAN candidate list is empty, then the service will be rejected.
- If service's RAT priority list only contains UMTS and the list of UMTS candidate cells is not empty, UMTS will be the selected RAT. If UMTS candidate list is empty, then the service will be rejected.
- If service's RAT priority list contains several technologies:
 - If there are only candidate cells of just one of the technologies included in service's RAT priority list, that technology will be selected.
 - Otherwise, there are candidate cells belonging to different technologies that are included in service's RAT priority list. Then go to step 5.

Step 5

Several steps are followed:

1. The instantaneous radio resource occupation is calculated for all the technologies that have candidate cells and are included in service's RAT priority list:

$$TSL_utilization_GLOBAL = \frac{\sum_{i=1}^N Num_bearers_BSi \times TSL_utilization_BSi}{\sum_{i=1}^N Num_bearers_BSi}$$

where N= number of GERAN cells.

$$Fc_GLOBAL_UMTS = \frac{\sum_{i=1}^M Num_bearers_BSi \times Fc_BSi}{\sum_{i=1}^M Num_bearers_BSi}$$

where M= number of UMTS cells.

2. The Remaining Capacity (RC) which measures the difference between the target value set by the operator (U_{GERAN} , U_{UMTS}) and the current value of radio resource occupation is calculated.

$$RC_{GERAN} = U_{GERAN} - TSL_utilization_GLOBAL$$

$$RC_{UMTS} = U_{UMTS} - Fc_GLOBAL_UMTS$$

3. In the case that just one technology fulfils that its RC value is greater than its decision threshold, then the user will be assigned to this technology. This situation happens when the occupation of one of the technologies is farer away than the rest from its target value, and therefore the priority list is not taken into account. The user will be directly assigned to that technology.
4. If two or more technologies fulfil that their RC value is greater than their decision threshold, then the algorithm will try to assign the user to the first technology in the RAT priority list of the requested service. In case that none of the cell of this first RAT can process the service, the algorithm will try to assign the user to the second RAT, and so on.

The algorithm was modified in the sense that even if the target load occupation factor was reached in each RAT users were attended. Users were distributed then attending to the difference between RATs target load occupation factors. As an example, if the target occupation factor for GERAN is 70% and for UTRAN is 50%, once one or both technology reach to this target, the algorithm will distribute the users between the RATs trying to maintain a 20% (70% - 50%) difference between their load factors.

Then the following conditions were added to the algorithm:

- If both Fc_{GERAN} **and** Fc_{UTRAN} are lower than the objective:
 - If RC_{GERAN} is greater or equal to $GERAN_decission_threshold$ and RC_{UMTS} is lower than $UTRAN_decission_threshold$ then the user is assigned to GERAN.
 - If RC_{UTRAN} is greater or equal to $UTRAN_decission_threshold$ and RC_{GERAN} is lower than $GERAN_decission_threshold$ then the user is assigned to UTRAN.
 - If RC_{UTRAN} is lower than $UTRAN_decission_threshold$ and RC_{GERAN} is lower than $GERAN_decission_threshold$ then the user is assigned following the service priority list per RAT.
 - If the user cannot be served by the RAT decided in the algorithm, it will try to be served by the other RAT even though the load factor doesn't meet the objective.
- If Fc_{GERAN} **or** Fc_{UTRAN} have reached to their target value:
 - Relation between the target load occupation factors: $|U_{GERAN} - U_{UMTS}|$
 - Users will be distributed attending to his relation.
 - If the user cannot be served by the RAT decided in the algorithm, it will try to be served by the other RAT even though the load factor doesn't meet the objective.

5.6.2.1 Considered initial RAT Selection Strategies

Two different strategies for distributing the load between the RATs in the heterogeneous network (GERAN/UTRAN) have been tried and the capacity provided for each one has been analysed. Network capacity has been defined as the maximum number of users connected at the same instant of time.

The network has been simulated in each strategy for different network load conditions, from an initial situation of very low load to saturation.

The following list for service priority per RAT has been used:

- Voice: GERAN
- Video Streaming: UTRAN
- Video Telephony: UTRAN
- Web Browsing: UTRAN, GERAN
- E-mail: GERAN, UTRAN

One main conclusion inferred from the initial simulations done for the study is that ‘GERAN_threshold’ and ‘UTRAN_threshold’ are the most important factors to study. So, several values have been also tried for these parameters.

The following strategies have been simulated:

- No application of the proposed initial RAT selection algorithm. Users are served by the RAT of their best server at their point where they are located in.
- Application of the algorithm with the following objective values:
 - $U_{GERAN} = 90\%$
 - $U_{UTRAN} = 70\%$

Several values for the RAT thresholds have been also tried:

- Case 1: GERAN_threshold = 50% and UTRAN_threshold = 30%
- Case 2: GERAN_threshold = 40% and UTRAN_threshold = 20%
- Case 3: GERAN_threshold = 85% and UTRAN_threshold = 70%
- Case 4: GERAN_threshold = 1% and UTRAN_threshold = 1%
- Case 5: GERAN_threshold = 85% and UTRAN_threshold = 1%
- Case 6: GERAN_threshold = 1% and UTRAN_threshold = 65%
- Case 7: GERAN_threshold = 80% and UTRAN_threshold = 60%

The graphics below show the results obtained for each service:

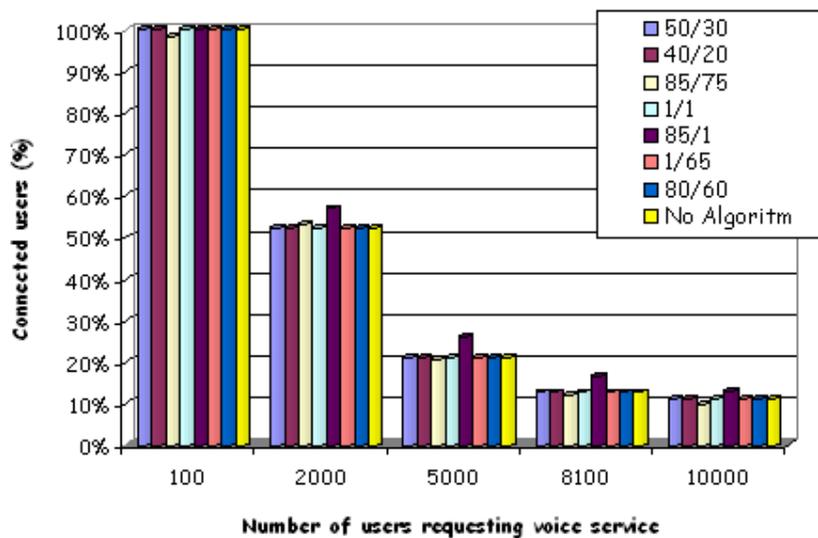


Figure 299 – Percentage of connected users for voice service.

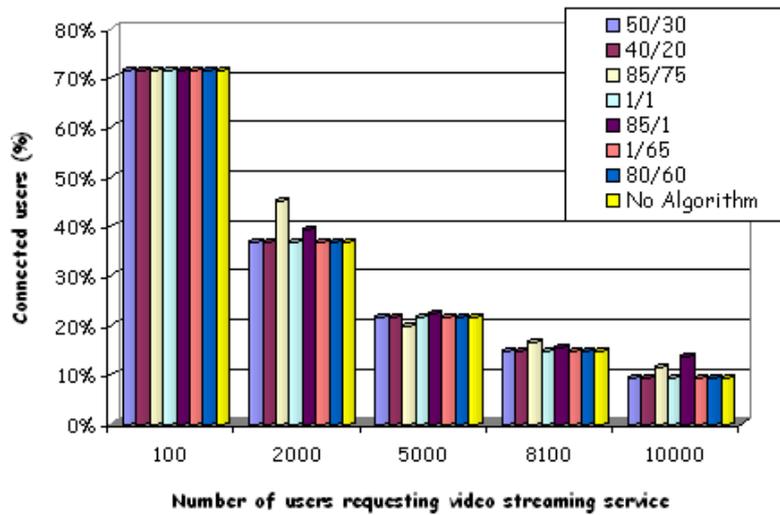


Figure 300 – Percentage of connected users for video streaming service.

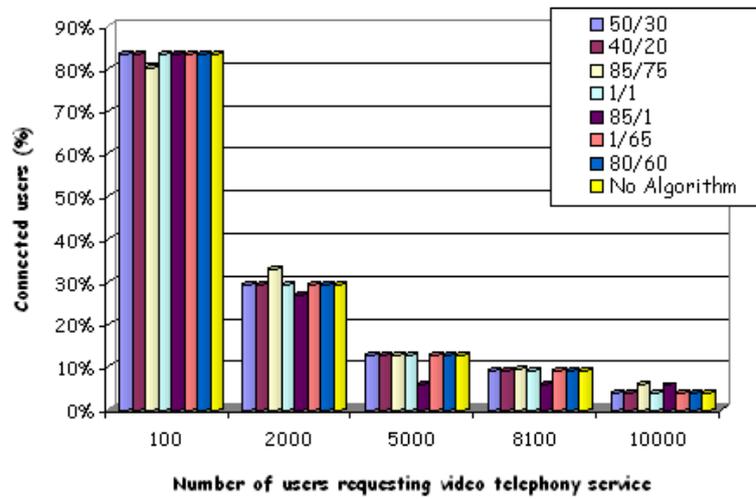


Figure 301 – Percentage of connected users for video telephony service.

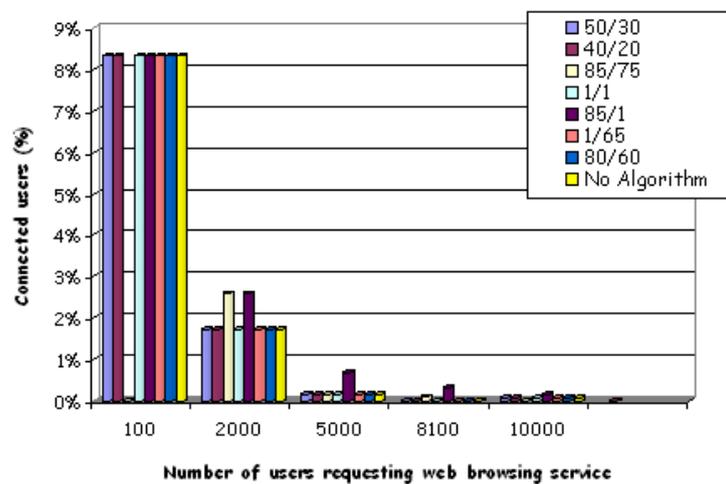


Figure 302 – Percentage of connected users for web browsing service.

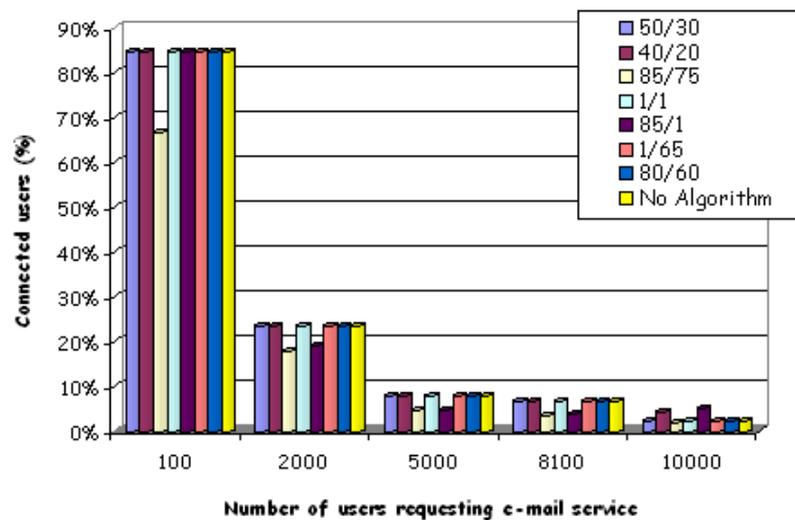


Figure 303 – Percentage of connected users for e-mail service.

5.6.2.2 Conclusions

The main conclusion obtained attending to the graphics is that results are not very different between both strategies, when taking into account the best server and when the algorithm is used. The parameters, GERAN_threshold and UTRAN_threshold, are the most influential parameters. They are defined and used to compare with the remaining capacity.

Even though the performance in both strategies is similar, in saturation situation (where the application of the algorithm has more sense) the strategy 2 (application of the algorithm) Case 5 (85% / 1%) is the one which shows better results for voice service and for some data services. This option models the situation in which UTRAN is loaded firstly. The worst results are shown in web browsing and e-mail services because of the priority list (firstly UTRAN and if not GERAN).

5.6.3 Study on sharing load and service prioritization

This section summarizes the most relevant results obtained during the analysis of how the load sharing between RATs impacts on the GoS of the heterogeneous (UTRAN and GERAN) network and how a joint admission control based on service prioritization affects to the final number of attended users per service.

5.6.3.1 Load sharing strategies: GERAN only can provide voice.

Regarding load sharing, a number of different strategies were tested in order to obtain as high capacity as possible (understanding capacity here as the number of connected users in the same instant of time); the following ones have been tried:

- Strategy 1: No control over the load sharing. When users may be served by both technologies, UTRAN and GERAN, the decision between both of them is taken in a random way, so no previous load sharing is decided by the operator.
- Strategy 2: Voice service can only be carried out by GSM and data services only by 3G, so that, 3G is able to spend all of its available resources among data users.
- Strategy 3: Data services can only be carried out by 3G and voice service may be carried out by both, GERAN and UTRAN but in a different percentage set a priori by the operator. The CRRM admission control shares voice load in a way defined by the operator, for example, giving priority GSM over UMTS

The following conclusions were obtained from the simulation results:

- When NodeBs were not saturated, the second strategy performed well. The numbers of total attended users for voice and for the rest of data services were maximized following the strategy of reserving voice service for GSM and data service for UMTS. This is due to the fact that NodeBs have still enough available resources to provide service to some data users.
- When in the scenario, the mean available power in NodeBs was reaching its maximum (41 dBm) and the NodeBs were reaching to their maximum allowed load factor, the strategy 1 performed better. The reason is that NodeBs had still available resources but not enough of them to attend any data service. In the second strategy these resources would be wasted but in strategy 1 are employed for voice service, which demands fewer resources.
- The third conclusion is that an operator strategy based on fixing an objective of load sharing between GSM and UMTS for voice service is not recommended because it doesn't perform as desired, as may be observed in results of strategy 3.

The detailed study may be found in section 4.4.2 in [17].

5.6.3.2 Admission Control based on Service Prioritization: GERAN only can provide voice.

In this section the global effect of prioritizing voice service over the rest of services in the case of joint admission control for GERAN and UTRAN was studied.

Regarding RAT selection, the results presented in this section were obtained in the same framework as in strategy 1 of previous section; when a user may be attended by GSM or UMTS, the serving RAT is selected in random way.

Two strategies were tried in order to analyse the results of this prioritization:

- Strategy 1. All services have the same priority when connecting.
- Strategy 2. Voice service is prioritized over the rest when connecting.

Several conclusions were inferred from the previous results.

CRRM Admission control strategies based on services prioritization have very little impact when offered traffic load is low. Therefore, all the other relevant conclusions are related to congestion situation.

A CRRM admission control based on case 2 strongly improved the GoS for voice users, while worsening the GoS for the rest of data services requested by the users. As system load grows, admission control keeps a good GoS for prioritized categories in a certain interval of traffic load, but finally GoS decreases as capacity runs out.

5.6.3.3 Admission Control Based on Service Prioritization: GERAN can provide both, voice and data services

5.6.3.3.1 Introduction

This section collects the most relevant results obtained from the study of the GERAN/UTRAN admission control taking into account the service prioritization between both RATs.

A previous analysis was done in [17] in which GERAN could only provide voice services. This study has been completed for this last deliverable so GERAN can provide both, voice and data services, same as UTRAN.

The scenario used is the one described in [24], 2b. The kind of users 'business' has been employed, with the same service mix defined also in [24]. The same simulation tool, URANO, has been used. It has been described in previous deliverables [24].

5.6.3.3.2 Admission Control Strategies

Three strategies for admission control in the heterogeneous network (UTRAN/GERAN) have been simulated for this analysis and the capacity provided by each one has been studied. For obtaining the results the system capacity has been defined as the maximum number of users connected at the same instant of time.

The objective of the study is the analysis of the possible impact in network capacity of a joint admission control 2G/3G which takes into account the kind of service. The network has been simulated in each strategy for different load network conditions, from an initial situation with very low load to saturation.

The strategies simulated are listed below:

- Strategy A: No control over admission control is applied. When users may be served by both technologies, GERAN and UTRAN, the decision is taken in a random way, without no predefined decision from the the operator. That is, both RATs have the same priority for all the services.
- Strategy B: Voice service users are only served by GERAN. Data services users are served by UTRAN if is possible; if not, they are served by GERAN.
- Strategy C: Voice service users are served by GERAN if is possible; if not, they are served by UTRAN. Data services users are served by UTRAN if is possible; if not, they are served by GERAN.

In the graphics below the results for each service are shown:

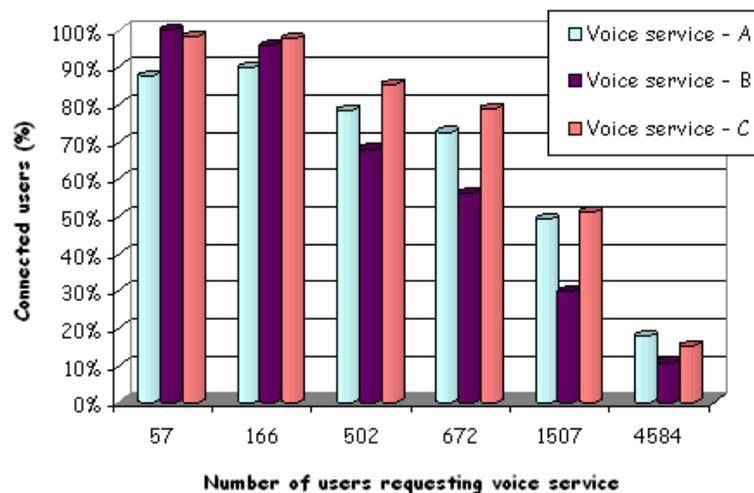


Figure 304 Percentage of attended users for voice service

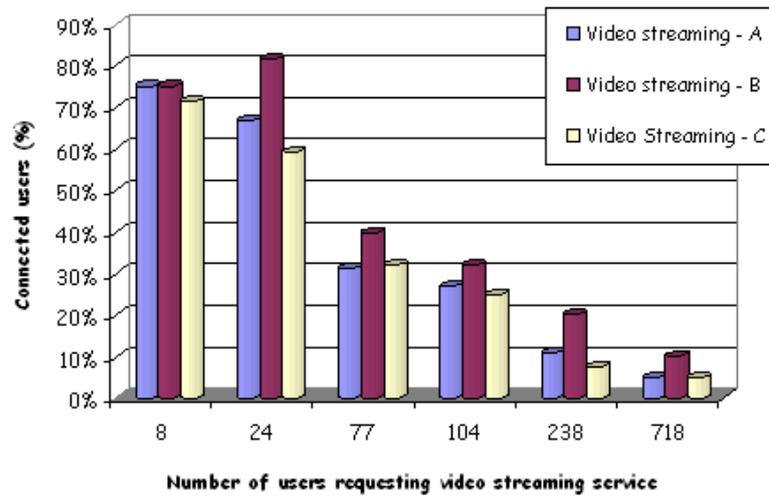


Figure 305 Percentage of attended users for video streaming service

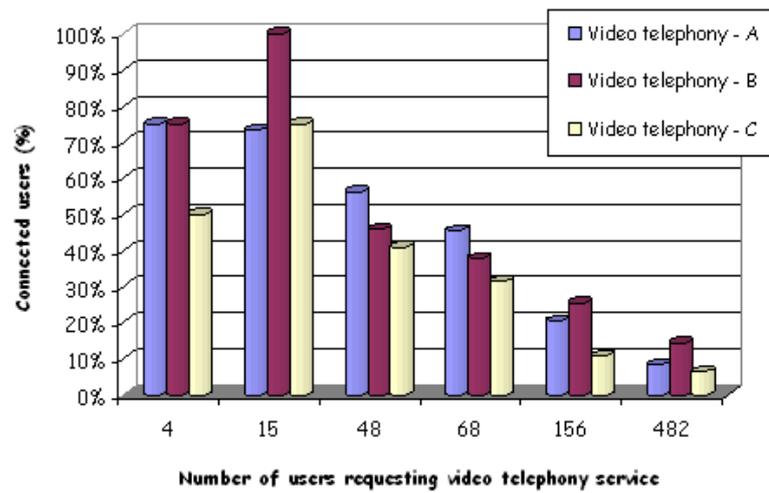


Figure 306 Percentage of attended users for video telephony service

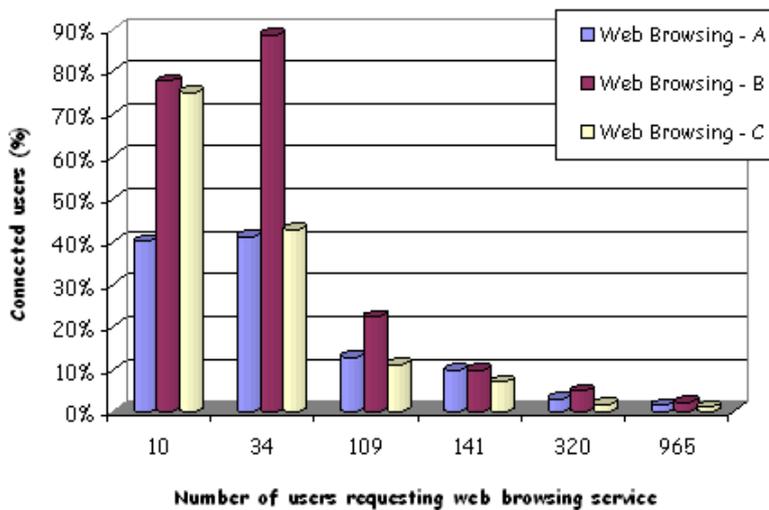


Figure 307 Percentage of attended users for web browsing service

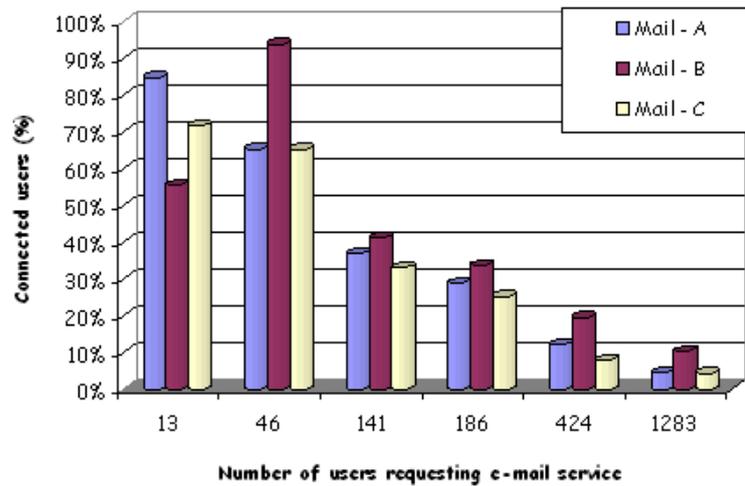


Figure 308 Percentage of attended users for e-mail service

In the graphics below, the radio resource usage for each technology is also shown. The RR usage for GERAN has been measured by means of “Load Factor”, i.e, the percentage of busy PDCCHs. The RR usage for UTRAN has been measured by means of the transmitted traffic transmission power (without taking into account the pilot power and the control channels power):

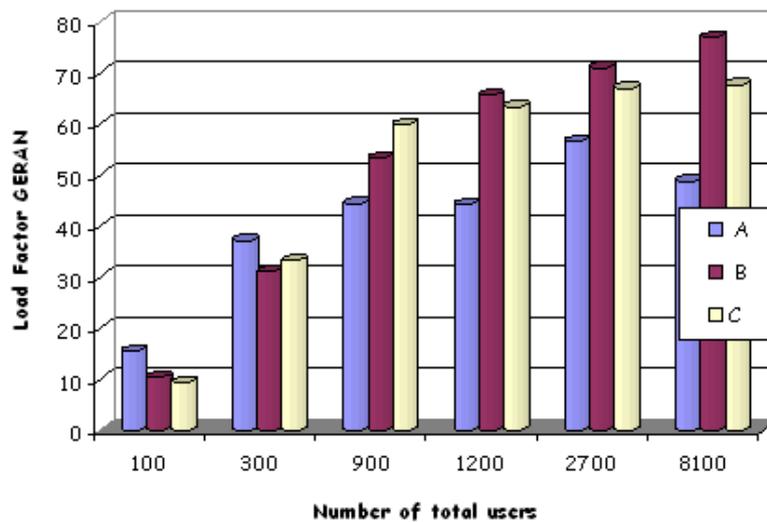


Figure 309 Radio resource usage for GERAN

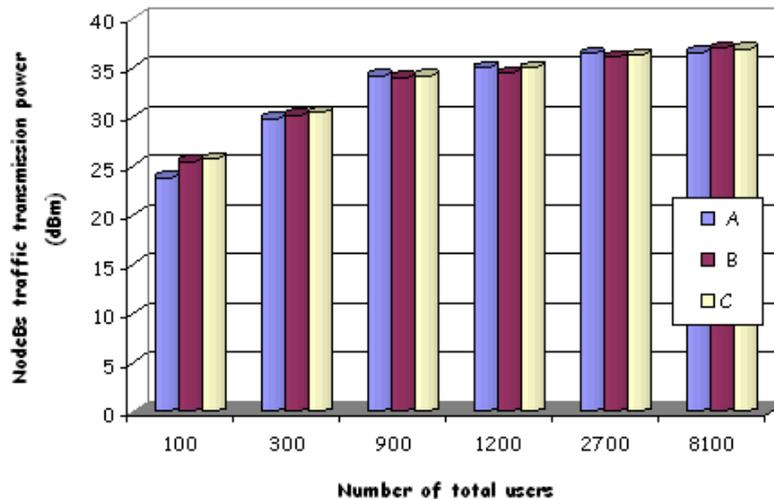


Figure 310 Radio resource usage for UTRAN

5.6.3.3.3 Conclusions

The following conclusions may be inferred from the results of the simulations:

- For low load situations, the strategy B (voice service only provided by GERAN and data services for UTRAN if is possible and, if not, for GERAN) shows the better performance for all the services.
- For high load situations, the strategy C (voice service provided by GERAN if is possible and if not for UTRAN, and data service provided by UTRAN and if not for GERAN) shows the better behaviour in general because the network capacity is maximized in the work conditions (strong presence of voice service in the service mix). In this situation, the strategy B is which shows the worst performance with a high number of voice users not served. However, it is possible that strategy B would perform better in a scenario with a strong presence for data services.

In conclusion, it may be inferred that the level of network load where services are going to be demanded should be taken into account when assigning services priorities to technologies.

5.7 PERCEIVED TCP THROUGHPUT IN CRRM FRAMEWORK

Some algorithms for balancing the load between RATs deployed in a heterogeneous network are investigated in this section. The algorithms must both take into account the commercial strategies and policies given by the operator and the user preferences as well as the instantaneous network performance.

5.7.1 Simulation assumptions

We have set the RAT preferences based on the parameters described in Section 5.2. Specifically the preferences are according to the service mix, where data services are preferably allocated to WLAN while voice is split between 2G and 3G until we reach the maximum load factor per cell of each RAT. The algorithms take into account link data rates and priorities according to QoS and SLA. The service mix and bit rates per service and service usage is according to [24], which have been used as input in the two theoretical scenarios hotspot in urban and hotspot along main road in suburban environment.

We used the perceived TCP throughput on top of the radio bearers for comparisons between RATs. We base the user TCP throughput on the time needed for a certain TCP connection including channel set-up and other network delays.

In [1] we showed an example of the perceived user throughput as a function of the physical layer bit rate for a data session size of 1 MB and a severe link where the BLER is 2% for all RATs. We showed that the system round trip time (RTT) is the dominating limiting factor when the physical layer bit rate is increased. Thus, 2G and 3G networks would limit the perceived user throughput at some 10 to 100 kb/s even if the physical layer bit rate would be increased. See [1] for more details on the simulation setup.

In [1] we also used this simulation tool to analyse the performance of a CRRM strategy compared to a manual RAT selection strategy where the service consuming the largest data load are scheduled on WLAN and smaller service loads are put on the 3G and 2G networks respectively. E.g. a large web download is always manually scheduled on WLAN, a smaller streaming session is manually scheduled on 3G and Mail or MMS to the 2G network. We showed that when all radio resources are utilized as a common pool by the CRRM the optimum load balancing gives a higher throughput. For these specific CRRM simulations we found that CRRM improved the throughput with 0 to 460% in different network load conditions.

A simulator structure as shown in Figure 311 is used for basic common radio resource management algorithm evaluations. The simulator handles incoming information from the Everest scenario documentation [16] and combines this with specific inputs regarding CRRM, such as meeting operator specific quality targets (e.g. service level agreements to special customers etc). Example of possible operator specific QoS strategies could be:

- Worst case allowed data rates to maintain a minimum QoS level
- Preferred RAT for users
- Preferred RAT for certain services
- Prioritised services

The scenario used from [16] is the hotspot in urban area. We have used the worst case data rate limitation as well as the prioritisation between services in order to meet specific QoS performance results.

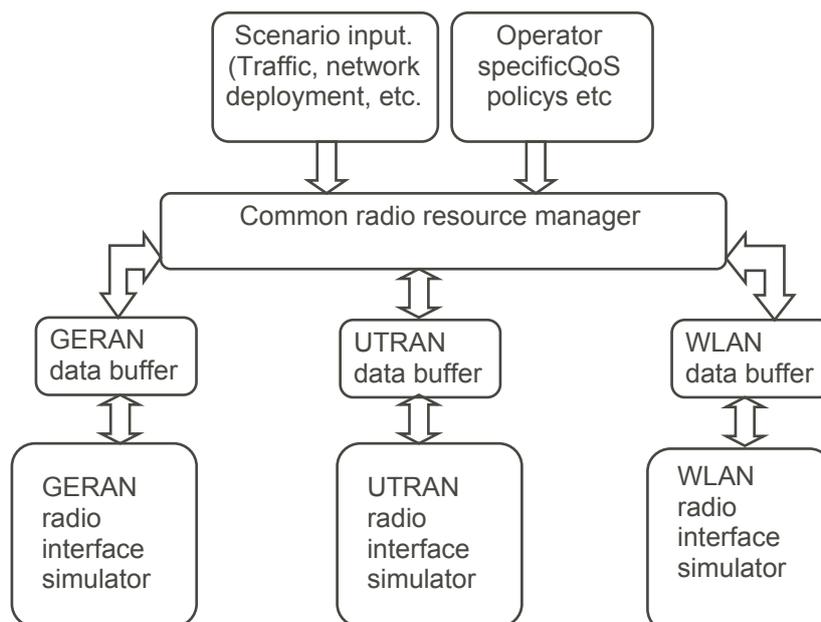


Figure 311 Simulator structure for basic CRRM evaluations with UTRAN (incl HSDPA), GERAN and WLAN radio access technologies.

Each radio access technologies (RAT) is individually simulated, assuming current network load, target BLER level and internal signalling delays etc. The RAT performance is then simulated based on the payload on each data transmission to be made of the specific RAT. A data transmission is generated by the following three steps:

1. A connection is established between the terminal and the corresponding network node
2. The transport channel setup procedure is conducted
3. The data transmission is performed assuming a TCP connection

Once the TCP connection is established and transmission is performed, the CRRM algorithm can at any point of time make an inter-RAT handover (CRRM re-allocation). The transmission procedure is then performed from point number one again with the remaining payload. The transmission process including possible CRRM re-allocations continues until the entire payload has been transmitted. This procedure is shown in Figure 312. This process will then result in a perceived TCP throughput value used for the CRRM evaluations.

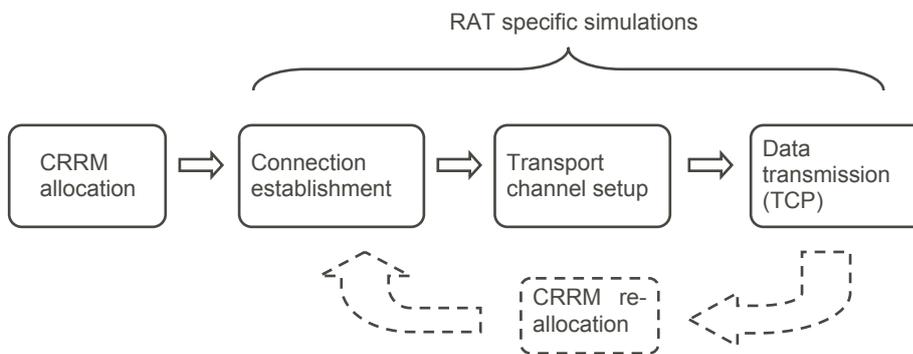


Figure 312 Model for the perceived TCP throughput performance simulations.

The connection establishment and transport channel setup procedures are RAT specific and includes GERAN aspects such as the GPRS attach procedures as well as the UTRAN connection request and connection setup delays. PDP context activation and radio bearer setup are also modelled.

The TCP modelling is based on a slowstart followed by average TCP throughput modelling based on the current status of specific parameters for each RAT. The RAT specific parameters are e.g. current cell load, average block error rates, system round trip times, core network signalling delays and general radio bearer information such as the physical layer maximum bit rate.

Common radio resource manager

The CRRM entity in Figure 312 handles all incoming traffic and by utilising different CRRM algorithms combined with possible additional inputs from e.g. an operator specific policy the incoming payloads are distributed to the RAT buffers. Admission control is applied on a RAT level in order not to overload the RAT transmission. The CRRM algorithm is event triggered where any new decision for re-allocation or new incoming traffic is managed based on one of the events used for a new CRRM evaluation. The possible events implemented are a finalised transmission of one payload (one data session) or one RAT buffer empty.

The different CRRM algorithms implemented are described in Table 66 below.

Table 66 Four different CRRM algorithms used for CRRM performance evaluations.

	Type of algorithm	Algorithm description
Algorithm 1	Long term CRRM optimisation criteria	At each evaluation the CRRM decisions are made trying to minimise the time to empty all buffers assuming the current load.
Algorithm 2	Short term CRRM optimisation criteria	At each evaluation the CRRM decisions are made trying to minimise the time until the first data session is finalised.
Algorithm 3	RAT prioritisation for CRRM	According to service prioritisation due to SLA or other operator specific QoS strategies, the service with highest priority is allocated to the RAT with highest system capacity, i.e. WLAN in this scenario. If the high priority data would overload the high capacity RAT, the next RAT (e.g. UTRAN) is also filled with high priority data etc. If the system can handle more traffic, the payload with the second highest priority is allocated in the CRRM using the same principle, i.e. with WLAN usage as best choice.
Algorithm 4	Manual RAT selection	The manual RAT selection used is that all payload with highest priority is allocated to WLAN, while the lowest priority payload is allocated to GERAN. The payload from other "middle priority services" is then allocated to UTRAN. This CRRM algorithm is used as the benchmark for the other algorithms.

5.7.2 Performance results

In [1] we showed an example of the perceived user throughput as a function of the physical layer bit rate for a data session size of 1 MB and a severe link where the BLER is 2% for all RATs. We showed that the system round trip time (RTT) is the dominating limiting factor when the physical layer bit rate is increased. Thus, 2G and 3G networks would limit the perceived user throughput at some 10 to 100 kb/s even if the physical layer bit rate would be increased. See [1] for more details on the specific simulation setup.

In [1] we also used this simulation tool to analyse the performance of a CRRM strategy compared to a manual RAT selection strategy where the service consuming the largest data load are scheduled on WLAN and smaller service loads are put on the 3G and 2G networks respectively. E.g. a large web download is always manually scheduled on WLAN, a smaller streaming session is manually scheduled on 3G and Mail or MMS to the 2G network. We showed that when all radio resources are utilized as a common pool by the CRRM the optimum load balancing gives a higher throughput. For these specific CRRM simulations we found that CRRM improved the throughput with 0 to 460% in different network load conditions.

In order to further evaluate the CRRM throughput gains for different settings we have made simulations with four different setups as shown in Table 67. The parameters we have changed are

- HSDPA capability in the 3G network. The WCDMA network in the hotspot in urban scenario can be equipped with HSDPA capability.
- Service prioritisation. We have set the service prioritisation to either prioritise small payloads or to prioritise large payloads.
- Operator specific minimum QoS level setting. The common radio resource manager input consists among others of a minimum QoS level, to be set by the operator. We have blocked users that would have been unsatisfied, only getting a perceived throughput of 1% (low) or 5% (high) of the maximum equivalent radio bearer available.

Table 67 Four different simulation setups based on the hotspot in urban scenario.

	HSDPA capability	Service prioritisation	QoS level
Case a	No	Large packets	Low
Case b	Yes	Large packets	Low
Case c	No	Small packets	Low
Case d	No	Large	High

As a reference case we consider Case a. The performance utilising the four different CRRM strategies are shown in Figure 313 below. The performance for each algorithm is presented as the relative total CRRM throughput compared to the “manual RAT selection algorithm”, algorithm 4. It can be seen that the relative CRRM throughput gain using the algorithms one to three varies between 35 to 40%. As will be seen in all figures for this section the first three algorithms will always result in a significant higher total throughput than the manual RAT selection, since the RAT utilisation is more effective. However, the most straightforward algorithm filling in each RAT to it’s maximum according to service prioritisation seems to be just slightly better than the other two more advanced algorithms, even through they all perform fairly similar. Most likely is the reason for this that the future payload is not predicted. Hence, the results indicate that it could be more important to fill all RAT buffers as much as possible at all times than to predict that it would be more efficient to have a longer perspective in time.

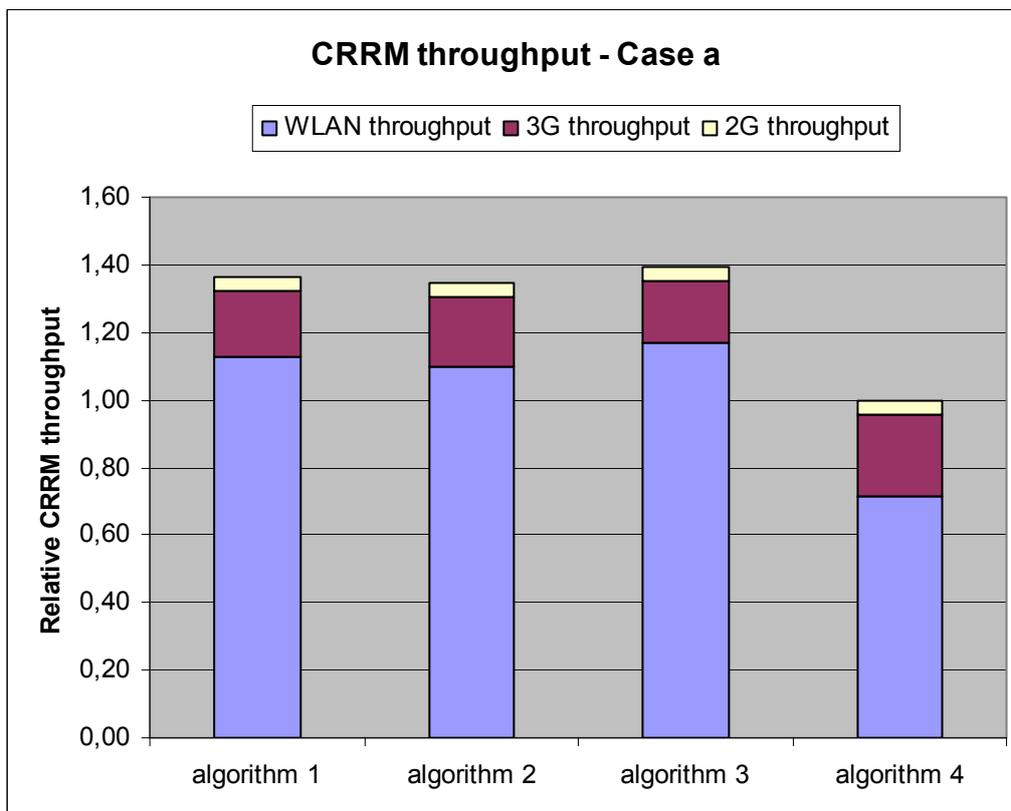


Figure 313 Relative CRRM throughput for Case “a” with four different CRRM algorithms.

Next we consider the performance impact when introducing HSDPA in the network. Figure 314 shows the relative CRRM throughput using the same setup with or without HSDPA in the

WCDMA network, i.e. the performance of the reference case (case “a”) vs the HSDPA setup (case “b”).

It can be seen that the throughput in the 3G network is increased significantly, which is a result of the reduced network delays in HSDPA compared to the ordinal release 99 WCDMA system. A reduced round trip time improves both the initial setup delays, but also the TCP performance significantly. Hence, the 3G throughput is in general improved significantly when utilising the inter-RAT handovers in a CRRM scenario.

Summarising the performance with HSDPA in the 3G network we can see that the total CRRM throughput gain compared to the manual RAT selection procedure based on service prioritisation is over 60% for all of the three other CRRM algorithms

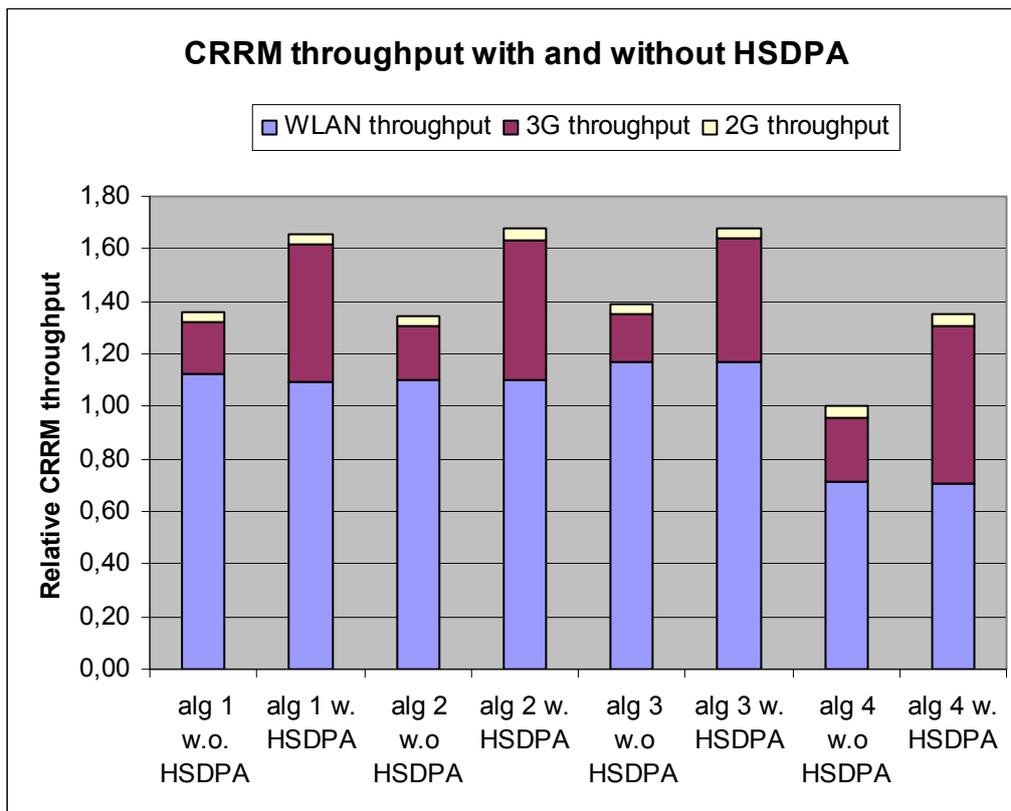


Figure 314 Relative CRRM throughput for Case “a” (without HSDPA) and “Case “b” (with HSDPA) using the four different CRRM algorithms.

The packet sizes in each service payload will impact the CRRM performance since a large packet will utilise more of the network radio resources. Hence the time spent transmitting data will be longer and the time spent for system setups will be shorter. However the impact is somewhat reduced since the setup time when making an inter-RAT handover is not negligible. The total CRRM throughput is therefore improved because of the larger ratio of payload compared to service initiations, but reduced because of more inter-RAT handovers. In order to quantify the CRRM throughput in the hotspot within urban area we consider a scenario where the operator specific service prioritisation is based on payload sizes. Hence, a service generating large packets can e.g be prioritised higher or alternatively lower compared to small packet payloads.

Figure 315 shows the impact of this type of service prioritisation. We compare case “a” where the services generating large packets are prioritised with case “c” where instead small

packet sizes are prioritised. We can see that the performance when prioritising small packets is slightly lower than the large packet prioritisation. In general the impact of setup delays on the 2G network performance is larger than the other networks due to the large system delays. This is seen especially for algorithm four where the payload allocated to the 2G network is approximately the same for the two cases but the performance for small packets is much lower than the large packet performance.

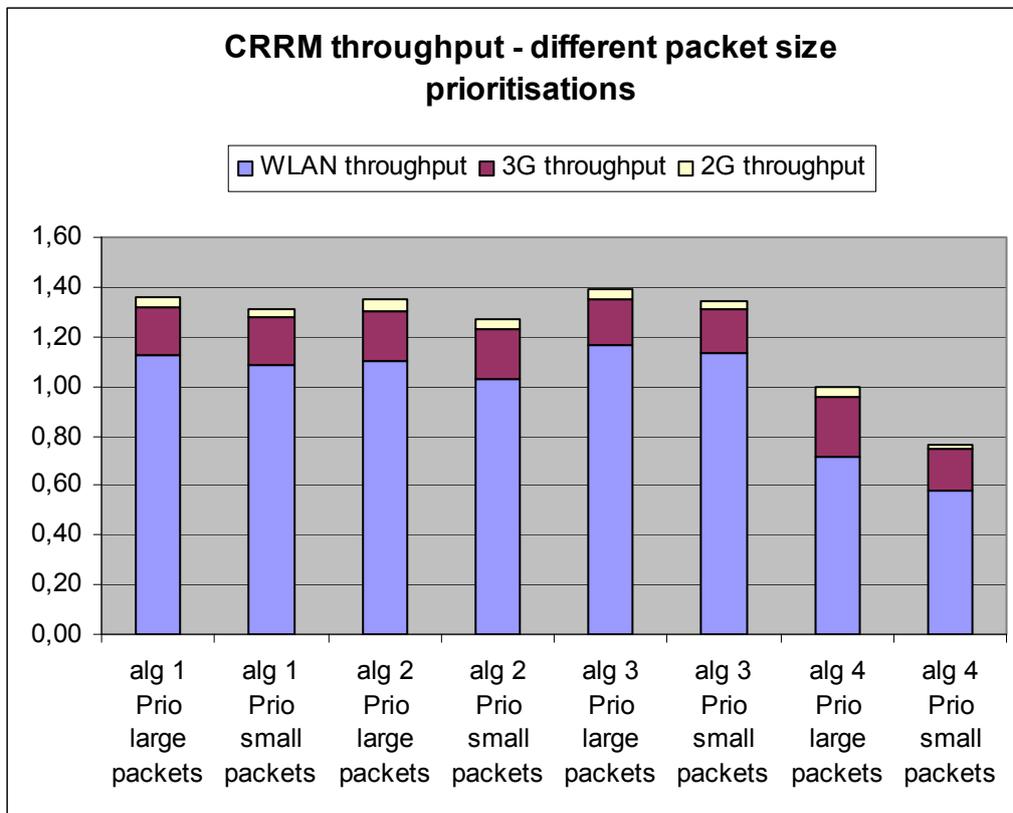


Figure 315 Relative CRRM throughput for Case “a” (prioritisation of large packets) and “Case “c” (prioritisation of small packets) using the four different CRRM algorithms.

As a last comparison we also consider the impact of the minimum QoS level operator specific parameter implemented as possible input for this common radio resource manager. The minimum QoS level defines a ratio between the maximum user bit rate available for each RAT and the expected perceived user throughput for a user awaiting a data transmission. If the expected perceived user throughput is below the threshold the user would not be satisfied with the achieved CRRM QoS, and the user will instead get access to another RAT or wait a few slots until more resources are available.

The two levels of minimum QoS used for these simulations are 1% and 5% of the maximum user bitrate (equivalent radio access bearer) for each RAT. The 1% target is denoted as the low QoS level, and the 5% target as the high QoS level.

Figure 316 shows the performance comparing case “a”, which has the low QoS level with case “d” utilising the high QoS level for the common radio resource manager input. It can be seen that when increasing the QoS level in the networks the total CRRM throughput is slightly reduced.

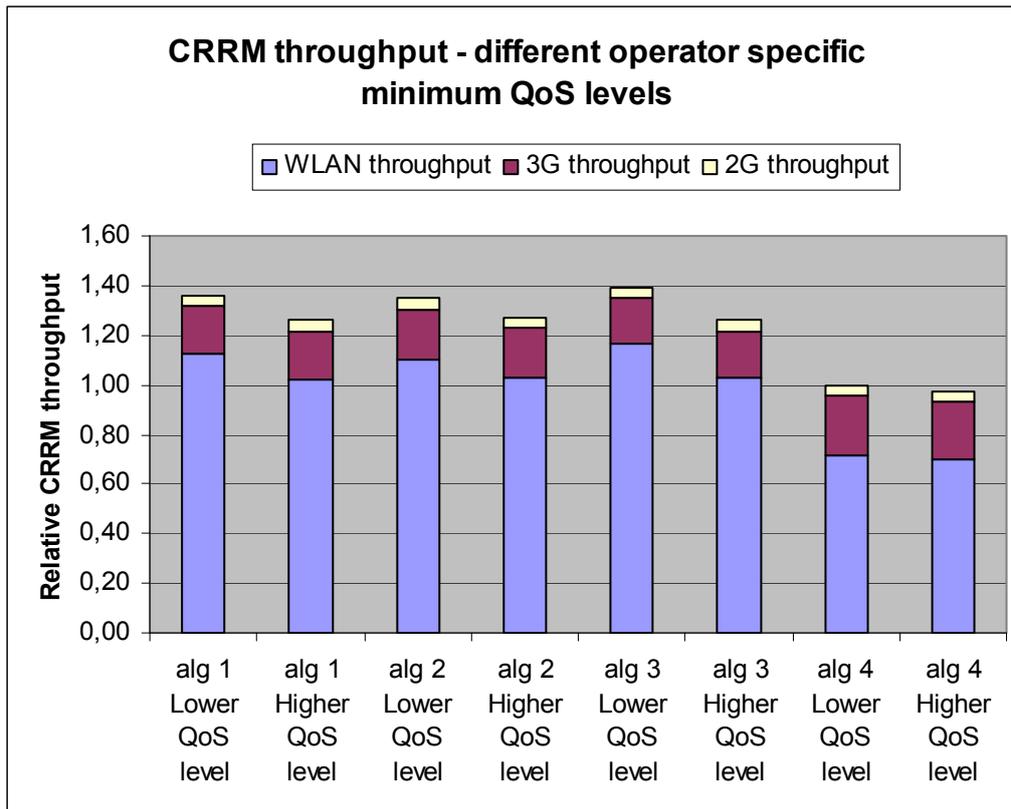


Figure 316 Relative CRRM throughput for Case “a” (low QoS level) and “Case “d” (high QoS level) using the four different CRRM algorithms.

5.8 IMPACT OF MULTI-MODE TERMINALS ON CRRM PERFORMANCE

5.8.1 Introduction

Today’s wireless communications comprise a broad variety of Radio Access Technology (RAT) standards. In order to provide the end-user with the requested service and corresponding QoS (Quality of Service) requirements in an Always Best Connected framework [150], beyond 3G (B3G) networks encompass the notion of integration and heterogeneity among different networks. In this way, heterogeneous networks may provide a larger set of available resources allowing users to seamlessly connect, at any time and any place, to the access technology that is most suitable according to some user/operator specified criteria.

Under the previous statements, and in order to take full advantage of B3G networks, existing mobile terminal capabilities need to be extended. Particularly, to provide connectivity to a variety of underlying access technologies is a must. In this sense, multi-mode terminals, which are able to operate via different RATs, are devised [151]. The assumption that 2G/2.5G/3G multi-mode terminals are available for most users in 2009-2010 with a penetration reaching 90% is still valid [152]. Furthermore, it is expected that the penetration of multi-mode 2G/2.5G/3G/WLAN terminals in the same timeframe will reach 50% of the population [152]. Therefore, in the short term, both single-mode (2G only) and multi-mode terminals will co-exist. On the other hand, the increasing complexity of multi-mode terminal devices may in turn result in a price increase. Consequently, some users may prefer simpler, smaller and cheaper devices for their basic needs such as, e.g., voice and short messaging service (SMS). Then, single-mode 2G or 2.5G terminals may not become extinct. All these

facts, together with many other factors will cause differentiation in terminals and will cause segmentation in the terminal market to grow even further.

The Third Generation Partnership Project (3GPP) has identified several issues concerning multi-mode terminals. Specifically, [153] identifies multi-mode User Equipments (UE) categories as well as describes the general principles and procedures for the multi-mode operation. In [154] the parameters of the UE radio access capabilities are addressed and some reference configurations are provided for utilization in test specifications. As for considering multi-mode terminals in Radio Resource Management (RRM) procedures, load sharing among different RATs was already devised between 1G and 2G systems, like e.g., AMPS and CDMA-based IS-95, as a form of improving flexibility and lowering infrastructure costs. In [155] the expected gain obtained through statistical multiplexing effect (trunking gain) in multi-mode multicarrier CDMA/AMPS deployment is investigated by allocating multi-mode terminals to CDMA and single-mode terminals to AMPS. This study did not consider service differentiation when allocating users to different RATs, only voice calls and no vertical handovers (i.e. seamless roaming between RATs during call/session lifetime) either. More recently, Lincke et al. proposes in several papers, e.g. [156] and references therein, that capacity in a cellular network can be expanded by rearranging traffic (both voice and data) between different RATs, where only multi-mode terminals are capable of doing so. By means of reallocating multi-mode terminals (through vertical handover) new incoming users with single-mode terminal capabilities may experience lower blocking probabilities. Reference [156] compares several substitution policies and evaluates them by means of simulations.

This section aims to analyse the impact of multi-mode terminal mixing in an EDGE/UMTS heterogeneous network with a policy-based initial RAT selection algorithm. This access selection is based on the demanding service-class and simulations will be performed considering different service-class mixings as well as multi-mode terminal mixings.

According to [153], a multi-mode UE is considered to be a terminal with at least one UMTS Radio Access Mode (UTRA FDD and/or TDD). In addition, the multi-mode UE supports one or more other RATs, e.g., GSM, (E)GPRS, WLAN, etc. In particular, in our study we will consider multi-mode terminals to be those with connectivity to GERAN and UTRAN radio interfaces. On the contrary, single-mode terminals are those that support GERAN RAT only. Moreover, [153] defines several types of UEs, namely Type 1 through Type 4. The assumed multi-mode terminal type in this section is a Type 2 terminal, which can, when utilizing one RAT, perform monitoring of another RAT and report it using the current RAT.

In the following, the service-based policy VG*VU explained in sub-section 5.3 will be considered in a scenario with voice and interactive www users. Specifically, this service policy first attempts to assign voice users to GERAN and interactive users to UTRAN. If no capacity is available in GERAN, voice users try admission to UTRAN. Similarly, rejected interactive users in UTRAN will attempt admission in GERAN. If no capacity is available in any of the RATs, the user gets blocked. Note that all this will apply provided the terminal has the required capabilities to operate with the suitable RAT, otherwise GERAN is selected as the default RAT. Figure 317 illustrates how the VG*VU policy is modified to consider multi-mode terminal capabilities.

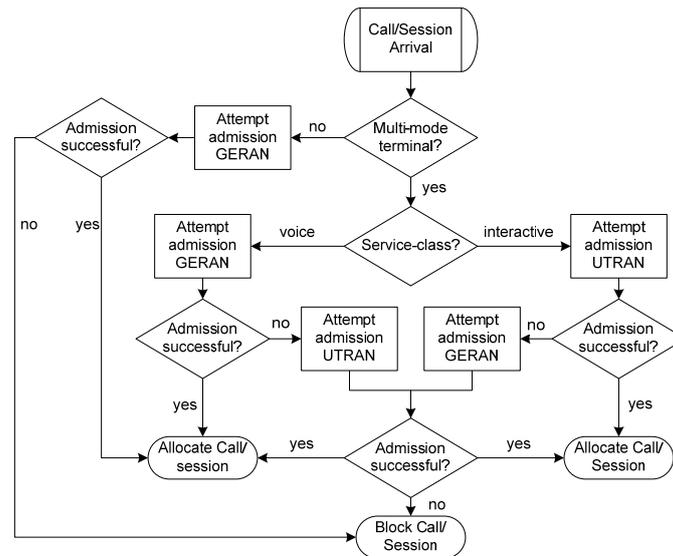


Figure 317 Initial RAT selection flow chart considering multi-mode terminals.

5.8.2 Simulation results

In the following, the performance of the previously defined initial RAT selection policy considering different multi-mode terminal availabilities and service mixing is evaluated by means of simulation. The simulation conditions are the same as defined in sub-section 5.3.2.1, with cell radius 500m. Representative values of multi-mode terminal availabilities consider, 25%, 50%, 75% and 100%, which indicate the percentage of terminals that support both RATs, i.e. GERAN and UTRAN. As for the service class mixing, several sets consisting of voice and interactive users are considered. Let $\bar{u}_i = (VU, WU)$ represent a service-class mixing set i consisting of a number of voice users (VU) and interactive users (WU). It is worth noticing that strain is placed on GERAN, not only having to cope with voice users being assigned by the aforementioned policy, but also with interactive users having single-mode (only GERAN) terminals.

5.8.2.1 Throughput Performance.

Table 68 shows the total aggregated throughput for different values of multi-mode terminal availability and different sets of service class mixings. A first expected result is that maximum throughput is achieved, for each service mix set, when all terminals are multi-mode (see rightmost column in Table 68). Results show that, as long as GERAN can handle its share of users, i.e. voice and single-mode terminal interactive users, no throughput degradation is noted when decreasing the number of multi-mode terminals. This is the case, e.g. for $VU = 200$.

Table 68 Total UL aggregated throughput (Mb/s)

$\bar{u}_i = (VU_j, WU_j)$		Multi-mode Terminal Availability (%)			
VU_j	WU_j	25	50	75	100
200	200	1,35	1,37	1,39	1,39
	400	1,76	1,79	1,77	1,81
	600	2,18	2,18	2,19	2,20
400	200	2,06	2,10	2,14	2,17
	400	2,29	2,40	2,50	2,55
	600	2,41	2,71	2,87	2,97
600	200	2,17	2,39	2,55	2,70
	400	2,27	2,57	2,87	3,07
	600	2,38	2,78	3,19	3,46

Since the maximum aggregated throughput is achieved in each service class mixing for 100% of multi-mode terminal availability, it can be useful to measure the degradation introduced by single-mode terminals. Accordingly, we define the throughput degradation $D_{i,j}$ for a given service class mixing i and a multi-mode terminal availability j as:

$$D_{i,j}(\%) = \frac{C_{100}^i - C_j^i}{C_{100}^i} \cdot 100 \quad (81)$$

where C_j^i is the total aggregated throughput for service class mixing i and multi-mode terminal availability j .

Figure 318 shows the throughput degradation as defined previously. Notice that, for different multi-mode terminal availabilities and different number of users requesting service, throughput degradation exhibits different trends. In particular, for $\bar{u} = (200, 200)$, no big differences are noticed, meaning that, in this case, GERAN is able to manage voice and single-mode terminals with ease. Certainly, the average timeslot utilisation factor in GERAN (properly defined in [133]) reveals an occupation of resources below 70%. The increase of users requesting to be served is translated into a bigger degradation in terms of throughput as multi-mode terminal availability decreases. While for $\bar{u} = (400, 400)$ degradation starts to get noticeable for multi-mode availabilities of 25% and 50%, for $\bar{u} = (600, 600)$ this degradation is already perceptible at 75% of multi-mode availability.

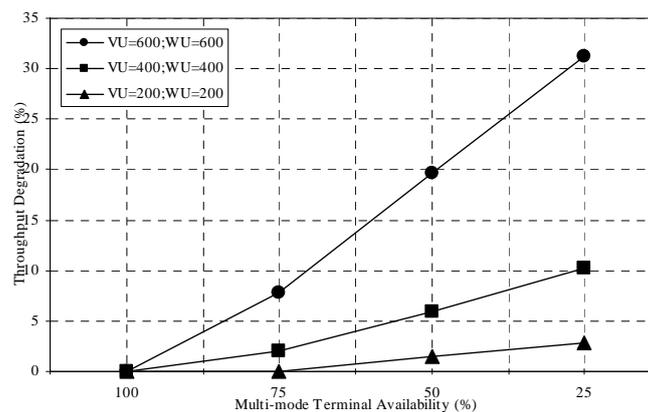


Figure 318 Uplink throughput degradation due to multi-mode terminals.

5.8.2.2 Delay Performance

Packet delay statistics for interactive users also reveal degradation when considering a scenario with mixed multi-mode and single-mode terminals. This degradation impacts directly in the perceived QoS by the data user and it is therefore important to keep its value as low as possible.

Figure 319 shows the uplink average packet delay for interactive users being served through GERAN for different mixings of multi-mode terminals and service-classes. Notice the increasing packet delay when multi-mode terminal availability decreases for $\bar{u} = (400, 400)$. As for $\bar{u} = (200, 200)$, the average packet delay remains almost constant with multi-mode terminal availability and also at an acceptable level, therefore exhibiting no degradation in this sense. Recall that the same behavior was also observed when analyzing throughput performance earlier on. A look at the average timeslot utilization factor reveals that, for

$\bar{u} = (400, 400)$, this value is over 90% while for $\bar{u} = (200, 200)$ this value is kept below 70%. This explains the big difference between the average packet delays of both service mixings. Observe that a multi-mode terminal availability of 100% is not taken into account because the considered service policy does not allocate any, or hardly any, interactive user in GERAN, so no statistics are available in this case.

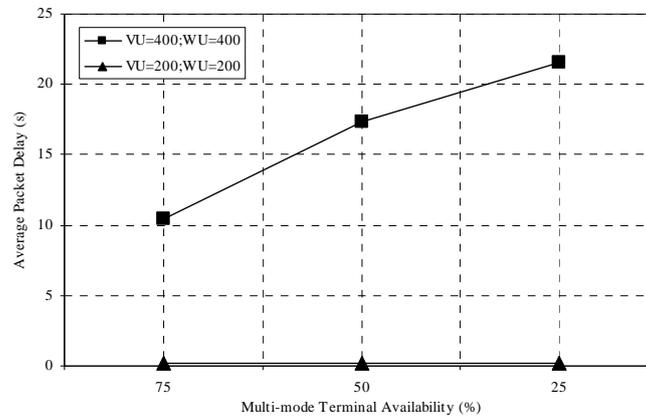


Figure 319 Uplink average packet delay for interactive users.

5.8.2.3 Approach using EGPRS dedicated slots

The suggested approach to overcome the delay increase problem, shown in Figure 319, resides in the introduction of dedicated EGPRS timeslots for interactive users. In this way, some resources may be reserved for interactive users and therefore packet delay performance can be improved with respect to not having any reservation scheme. As for throughput performance, it can be foreseen that contribution of voice users to the total aggregate throughput might diminish. This reduction of voice throughput may be compensated, given certain conditions, by interactive users allocated to the reserved slots. This entails a trade-off between the number and the applicability of reserved resources (slots) for interactive users in GERAN and other parameters, such as offered load and multi-mode terminal availability, as we will see in the following.

Without loss of generality we assume a resource reservation scheme by dedicating 3 EGPRS slots per cell for interactive users. Recall that, three carriers per cell are available and that slot 0 of first carrier is devoted to control and signaling. Thus, 23 slots are left of which 3 are EGPRS-only slots (13% of total resources) and the rest are reversible slots for both voice and interactive services (87% of total resources).

In order to evaluate the suitability of using dedicated slots, we can define a gain parameter, G , as:

$$G_{i,j}(\%) = \frac{C_0^{i,j} - C_3^{i,j}}{C_3^{i,j}} \cdot 100 \quad (82)$$

with $C_k^{i,j}$ being the total aggregated throughput for service mixing set i , multi-mode availability j and k dedicated slots for EGPRS services. This gain indicates if aggregate throughput performance is improved ($G > 0$), degraded ($G < 0$) or unconcerned ($G = 0$), when using 3 dedicated EGPRS slots for data traffic as opposed to not using any.

Table 69 shows the gain introduced by dedicated resources as defined in (82). It can be seen that for 200 voice users and the considered set of interactive users no especial improvement or degradation is noted. This is because GERAN can manage with 200 voice users even with the shortage of resources introduced by dedicated data slots and still can

also handle single-mode terminal users requesting interactive service either by allocating them to dedicated slots or to reversible slots. On the other hand, UTRAN is able to serve the set of multi-mode terminal interactive users and therefore does not make use of resources in GERAN. Thus, no degradation or improvement is observed in terms of throughput.

As for 400 voice users, the improvement in throughput due to the reservation scheme depends on both the service-class mixing and the multi-mode terminal availability. In general, the overall trend when the number of multi-mode terminals increases is reflected in the gain reduction of throughput due to the reservation scheme. In particular, for $\bar{u} = (400, 200)$, degradation in throughput is observed, which gets more severe as the number of multi-mode terminals increases. On the contrary, increasing the number of interactive users results in a throughput gain, particularly for low multi-mode terminal availability. This is explained due to the fact that, for voice loads above 87% (i.e. the percentage of reversible resources), gain is only achieved if the interactive users are sufficiently high in order to contribute with throughput in the remaining 13% of dedicated resources and compensate for the lack of voice throughput contributions in those reserved resources.

For 600 voice users, the behaviour is similar to the case of 400. However, resource reservation gain is achieved even for higher multi-mode terminal availabilities.

Table 69 EGPRS Slot Reservation Gain (%).

$\bar{u}_i = (VU_j, WU_j)$		Multi-mode Terminal Availability (%)			
VU_j	WU_j	25	50	75	100
200	200	0.00	0.00	0.72	0.00
	400	1.14	0.56	1.13	0.55
	600	0.00	0.46	-0.46	-0.91
400	200	-3.88	-4.76	-6.07	-7.83
	400	3.49	-0.42	-4.40	-5.10
	600	8.71	2.21	-1.39	-4.71
600	200	1.38	-2.10	-2.75	-1.48
	400	13.22	7.00	-0.35	-4.23
	600	13.87	12.23	3.45	-1.16

Finally, Figure 320 depicts the uplink average packet delay when considering 3 dedicated EGPRS slots for interactive users. Clearly, these users benefit from the dedicated slots exhibiting lower packet delays than in the case of not having any reservation scheme (see Figure 319).

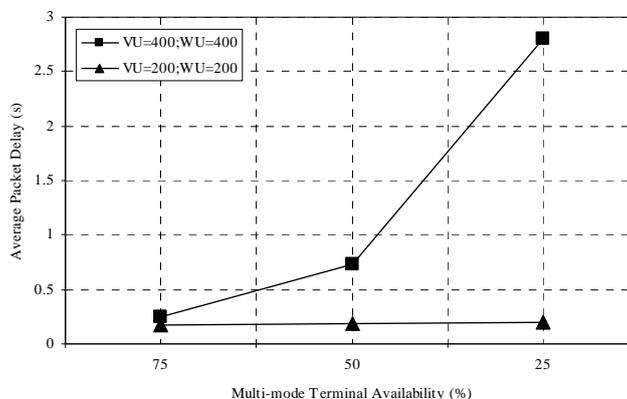


Figure 320 Uplink average packet delay for interactive users for different multi-mode terminal availability and 3 dedicated EGPRS slots.

5.8.3 Conclusions

This sub-section has covered several aspects on the impact of multi-mode terminals in the framework of heterogeneous networks with an initial RAT selection policy. Results indicate degradation in terms of throughput introduced by the limited operation of single-mode terminals. By considering a reservation scheme in GERAN for interactive users, we can improve the average packet delay for such users. While for a high multi-mode terminal availability results indicated that the reservation scheme was not necessary, for lower multi-mode terminal availabilities this scheme improved both aggregated throughput and packet delay figures, particularly for high number of interactive users.

6 CONCLUSIONS

This deliverable has presented a wide range of studies in the field of RRM/CRRM strategies that have been developed in the framework of the IST-project EVEREST. The proposed and analysed solutions envisage an optimised utilisation of the scarcely available radio resources for the support of mixed services within heterogeneous networks. The management of radio resources can be seen as a problem with multiple dimensions. Every RAT is based on specific multiple access mechanisms exploiting in turn different orthogonal dimensions, such as frequency, time and code. Then, RRM mechanisms are needed for every considered RAT, namely GERAN, UTRAN and WLAN in the case of EVEREST. CRRM is based on the picture of a pool of radio resources, belonging to different RATs but commonly managed. Then, the additional dimensions introduced by the multiplicity of RATs available provide further flexibility in the way radio resources can be managed and, consequently, overall improvements may follow.

As a final report of EVEREST's activities on RRM and CRRM for the whole duration of WP3, this document has captured the different developments achieved from March 2004 to October 2005. For those aspects already included in previous deliverables (D11 "First report on the evaluation of RRM/CRRM algorithms", issued October 2004, and D15 "Report on the evaluation of RRM/CRRM algorithms", issued March 2005) only some relevant concepts and a few representative results have been included here, so as to maintain the self-contained nature in the present final report. Strategies and algorithms developed since March 2005 have been described with a higher level of detail, so that these activities are suitably reported in an official deliverable. Thus, the extension devoted to every single topic covered in the above sections is not necessarily indicative of its relevance for the whole EVEREST project.

In the following, the different covered aspects and the main extracted conclusions are summarised:

UMTS

- An innovative mathematical framework capturing the air interface coupling among the different cells in the scenario has been developed. This framework has been presented in a compact formulation for both uplink and downlink, which is claimed to be novel for both directions. The multiple issues impacting the radio interface behavior and the much higher degree of coupling among them deriving from the WCDMA nature has been shown on a more visible form than more classical formulations. It is based on computing the gradient of the uplink load factor and the downlink transmitted power among cells. Several applicability examples of the proposed gradient characterisation have been developed and assessed.
- An analytical framework has been developed for an integrated voice/data CDMA cellular system. To model the close interaction between inter-cell and intra-cell

interference, an iterative process has been applied. Moreover, since congestion control function is applied to differentiate the voice and data traffic, i.e. in congestion state, data traffic is delayed to guarantee the real-time delivery for voice traffic, the iterative process includes the congestion function that furthers the correlation between inter-cell and intra-cell interference. With this framework, the system performance has been evaluated in terms of packet delay and voice outage probability with and without power control errors. The results on effects of control threshold in congestion control function on the system performance suggest that a tighter threshold is necessary to guarantee voice quality as the power control errors get larger. And further, the data delay performance degrades quite fast as power control errors increase even with congestion control, so an accurate power control for data transmission is necessary to sustain data throughput.

- Indoor traffic is very important in 2G networks, as remarkable traffic load is originated and terminated inside buildings. The implications of indoor traffic in 3G WCDMA based systems have been demonstrated to be significantly different from 2G TDMA-based solutions because transmitted power levels are the key radio resources in WCDMA. In particular, the impact on system capacity deriving from different percentages of indoor traffic in the scenario has been covered, firstly, by means of a simple analytical model and, secondly, by analyzing different radio bearers with different asymmetry levels by means of system level simulations. The complete set of obtained results may help to provide the indications on how and when new infrastructure and/or new cell hierarchies need to be deployed in a given scenario as user density and demanded services evolve.
- A formula comparing the power increase in uplink and downlink due to indoor traffic was derived. From this, it was concluded that the power increase will be higher in the uplink direction and the higher the load in the system the higher the difference with respect to the downlink will be. Consequently, a lower degradation caused by indoor traffic is expected in the downlink when compared to the uplink. Particularly, when half of the users are indoor, the reduction in the uplink is 88% while in the downlink it is only 15%.
- In order to assure the user QoS (Quality of Service) requirements in a hotspot, it is necessary a proper radio network planning. However, hotspots characteristics (such as geographical location, etc.) are not always known a priori, so an unexpected increase in the demand of resources by a hotspot can have a relevant impact on network performance. Therefore, it is prime important a proper evaluation of the effect of these hotspot peculiarities on the network behaviour. Moreover, the existence of dynamic hotspots (i.e. a group of mobile users that move following a certain mobility pattern, such as the way out of a railway station, the exit of a football match, etc) and its impact on system performance is as well an important issue, especially on the forward link, where the user location distribution affects directly on the base station power allocation. These aspects have been evaluated and some proposals on suitable solutions have been formulated and supported with simulation results. In particular, a pilot adjustment algorithm reducing the differences in the load of the different cells by shedding traffic from overloaded cells to low loaded cells has been presented. The load balancing technique reduces the power limitation probability of the different base stations, and this, in turn, reduces the dropping probability.
- The exploitation of the usual static nature of data traffic in WCDMA for admission control purposes has been addressed. In this case, the adoption of an advanced admission control policy, denoted as PLEBAC, which takes path loss reports into account results in a significant improvement of the system performance. The throughput gain of the PLEBAC admission control with respect to PABAC for different

cell radii can be as high as 35%. As a result, it can be concluded that the PLEBAC strategy is better adapted to user's distribution in the network.

- The impact of using repeaters within UTRAN has been analysed, taking into account an outdoor environment. In order to estimate the capacity gain, a hot-spot scenario has been considered. It has been verified that the capacity gain is rather low, unless a high ratio between the mean number of users within the hot-spot and the users within the cell is considered. The achieved simulation results suggest using repeaters only in UTRAN macro cells used to offer the service in area with high density of users.
- The possible usage situations for repeaters in a WCDMA system are: coverage extension, in order to cover the so called "dead spot" (areas not covered during the first deployment of the network); capacity extension, in order to increase the capacity of a base station with an increased traffic load; or soft-handover region reduction: thanks to the repeaters it is possible to reduce the soft-handover areas for already-connected users. When a repeater is introduced in a WCDMA system, the first effect that it is possible to observe is the increase of the noise figure of the base station. Considering the coverage area, with the increase of the noise figure of the base station we have the receiver sensitivity making worse, with the consequence that the base station coverage decreases, although the total cell radius is increased due to the introduction of the repeater.
- We have simulated an UMTS operator using two frequencies in an indoor traffic hot spot within an urban area. Different strategies to distribute the load of the different services between the two carriers have been analysed. The different methods affect on blocking ratio, dropping ratio and bit rate.
 - The results indicate that PS blocking is sensitive to the load allocation on the frequencies, whereas speech blocking and dropping, and PS bit rate, is less sensitive to different load distributions. HTTP blocking is improved and the speech QoS is worsening when all speech is allocated to frequency f1. Speech QoS is improved when all speech is allocated to frequency f2 (assuming that inter-frequency handover does not cause dropped calls). If soft handover is tuned off on frequency f2, the PS bit rate is increased, but, unfortunately, the PS blocking is also increased.
 - On the other hand, both speech and HTTP gain when speech services are redirected. The speech blocking is however reduced marginally. When redirecting PS services, the PS blocking is reduced significantly but the PS bit rate is not improved. The speech service does not improve its QoS when HTTP requests are redirected. We note that redirecting service requests is as pooling together the radio resources of the frequencies. A service that requires a small amount of radio resources will have a smaller trunking gain of redirection than a service which requires a large amount of radio resources. When pooling more and more resources together the trunking gain will be less and less.
- The project has also studied the setting of different parameters and strategies in Hierarchical Cell Structures. In that respect, the main conclusions are:
 - In the study devoted to assess and to compare the capacity offered by a multi-layer deployment versus a single-layer layout, two situations have been considered: macro- and micro- cellular layers with the same radio frequency (layer discrimination is only *power-based*) and macro- and micro- cellular layers with different radio frequencies (layer discrimination by means of radio

frequency). For the analysed scenario, in the first case an increasing of capacity of about 86% has been estimated, comparing the values of capacity of both the situations (without and with the microcellular layer). Using two layers with different frequencies (second case), the improvement is about 133%. According to the achieved results, HCS can be considered as a valid solution for network operators for trying to solve some congestion problems due to areas where there is a great amount of traffic (hot spots).

- Using intelligent link adaptation to obtain extra capacity by adding embedded hot spot in a macro-cell has also been studied. In this approach, the same frequency is reused in both macro-cell and hot spot. By taking advantage of the interference variation in time, packet switched data service is supported in the hot spot with link adaptation. The link adaptation is design to schedule data transmission and data rate for users in hotspot based on the interference conditions especially the interference from macro-cell in hotspot area, i.e. increase the transmissions when interference low and decrease the transmissions when interference is high. And further a power constraint is updated from time to time for hot spot base station, which ensures that the radiation from hot spots does not affect macro-cell performance at all, and thus make sure that any throughput gained in the hot spot is the extra gain harvested from the spectrum. Our simulation results show that it is possible to obtain some extra capacity gain by this implementation.
- The complete frequency reuse in WCDMA simplifies in the frequency domain the planning exercise. Nevertheless, in the usual case that more than one WCDMA carrier is available for a given operator, frequency assignment to cells also plays a key role. Using the derivative framework developed in the theoretical studies has revealed to be a good approach to cope with this issue. In particular, it has been shown that the derivative-based algorithm performs better than other reference algorithms.
- With respect the topic on mobility issues within HCS, the activity's goals have been to identify how to set the thresholds that control the algorithm applied by the terminal for the identification of the high/low mobility state by means of the number of cell reselections performed in a fixed period of time. The results of the analysis highlights how a network operator could manage the settings of a plethora of parameters in order to segregate the traffic according to its necessities. In the situation described in this document, the segregation has been based on the mobility class of the users. In addition, it has been also showed that an operator may set the trade-off between the negative effects of the segregation (e.g. impacts on the interference level due to the increase of the "*near-far*" effect) and the benefits of it, according to its objectives. The decision of which solution has to be used is left to be assessed case by case.
- The RRM algorithms controlling the transport channel type switching procedure can have a very important role when users request VBR (Variable Bit Rate) services. This is the case, for example, of the World Wide Web browsing: this service implies a very discontinuous data transfer due to the reading time spent by each user to analyze the received WEB page. In cases like this, it is very important to be able to optimize the usage of dedicated transport channel (DCH), preventing channelization code shortage in the downlink without degrading the end-to-end quality of service experienced by the user in an appreciable manner. As a general rule, even if the common channels are used only when a very small amount of data (or nothing at all) have to be transferred, the QoS is better for DCH-only than for TCTS because of the switching time. DCH offers higher transfer speeds (throughput) but requires a

significant setup time, whereas shared channels have a low throughput but also a low setup time. Due to these considerations the usage of common channels is more efficient only when the traffic is sporadic or for short time, otherwise for long and frequent data sessions, the usage of a dedicated channel is highly recommended. Another very important issue consists in the proper setting of the parameters controlling the reporting of the traffic measurements performed by the users, in order to prevent an overhead of useless signalling traffic.

- User's admission in a realistic indoor/outdoor scenario has also been studied. The simulations have allowed corroborating what was being observed by the mobile operator by means of measurements: the percentage of users located in a certain floor attended by stations situated in a different floor is very low. Furthermore, the study has concluded that, for low load situations, the effect of prioritizing services is faintly noticed, whereas as the network load grows up, the percentage of connected users for data services considerably increases. The percentage of attended voice service users decreases, but not dramatically. On the contrary, the loss is almost 10% with respect to no-prioritization strategy. Then, it may be inferred from the graphics that this planning scheme, with outdoor macro stations, would have a good performance because, when the indoor microcells tend to saturation (because of the number of attended users and also of the interference level), the macro stations start absorbing this traffic. This would be a proper solution when a hotspot in the building happens and the situation outside remains constant. Then, the solution would be locating macrostations outside the building and prioritizing data services.
- The DiffServ aware link adaptation has been studied to improve the efficiency to deliver DiffServ traffic over CDMA-based air-interface under the EVEREST end-to-end QoS architecture. In this study, to cope with both the multi-class traffic and multi-dropping levels in each AF class, the resource differentiation is carried out at both inter-class level and intra-class level for AF traffic. Inter-class level is designed to allocate the resource to each class. Through this inter-class differentiation, a better QoS balance is achieved between classes and thus a larger admission region is obtained for AF classes. And the intra-class differentiation is made by color aware joint the buffer management and rate adaptation for each AF classes in order to improve the QoS performance from the DiffServ point view i.e. guarantee the throughput for in-profile packets by penalizing the throughput for out-profile packets in congestion state. Finally, through the delta modulation feedback scheme, the proposed intra-class and inter-class schemes are integrated to form a close interaction loop. This guarantees a rather strong self-adaptation capability of this proposed link adaptation. We also examined the important designed parameter $\Delta\phi$ for the delta modulation adaptation. It suggests that a small value of $\Delta\phi$ favors the system performance especially when the EF class traffic load is high. The admission region provided by this study can be useful for further admission control study to support IP service over CDMA-based air-interface.
- The project has analysed also the following aspects in relation with the high-speed downlink packet access (HSDPA) evolution of the WCDMA:
 - The main benefit with HSDPA is that it reduces the user packet call delay and it increases the system capacity in downlink. This increased capacity can be used to either increase the number of users in each cell, and/or to provide the existing users with higher average data rates. Typically we can increase the link throughput by some 30-50% at a given cellular deployment geometry G-factor. The G-factor describes the ratio between average received power from wanted base station to the average inter cell interference power.

- In what schedulers concerns, it has been shown that the Max (C/I) scheduler obtained best throughput results. With the available CQIs the system achieves the service capacity of near 4 Mbps. Max (C/I over average C/I) and Round robin achieve 3Mbps and 2.5Mbps of service throughput respectively. CDF of user average bit rate shown that Max (C/I over average(C/I)) have most equilibrated values, with bit rates varying from about 70 kbps to 215 kbps. Max (C/I) obtained average bit rate varying from about 5kbps to 430Kbps. Concerning BLER, Max C/I scheduler obtained average value below $3E-1$. Worst average BLER in cell was obtained by the Round Robin, with $4E-1$.
- Comparing the two link adaptation schemes, one that will aim to provide best system throughput the other that selects higher CQI that provide probability of correct transmission of 0.9. Results show that, in fact, best throughput result is obtained with the 'Maximum throughput' link adaptation scheme. However this scheme has associated higher BLER. With probability correct transmission of 0.9, less block error probability and less transmission attempt per block is obtained. Less transmission attempts will lead to less block delivery delay.
- Code multiplexing that enables multiple users transmit on the same frame to facilitate HSDPA was studied from various perspectives including throughput, fairness, and packet delay. A proportionally fair code multiplexing scheduler that greedily exploits available power and code resources have been exploited. The effects of code multiplexing on the features of HSDPA, including AMCS and HARQ are also taken into consideration with channel and SIR estimation errors. Further, the use of interference cancellers, which potentially improves the overall performance in favour of the single server scenario due to a larger amount of interference that can be cancelled, was also studied. Our simulation results on a multi-cell scenario elucidated that the code multiplexing is able to improve the system performance regardless of the interference cancellation. All important performance measures, including the user and system throughputs, fairness, average packet delay, as well as delay jitter, were improved significantly.
- Observing that the channel estimate is less accurate when the channel SIR is low, we further introduced a biased AMC scheme. A larger transmission power is imposed on each MCS to lower error rates, such that the number of HARQ retransmissions and the consequent delay are reduced. A power offset of 3 dB increased the system capacity by nearly 30 percent, consuming about 1 dB of extra power on aggregate. To derive an appropriate power offset, we analysed the statistics of the channel estimation error including the CQI feedback delay. Our simulation results showed that adapting the offset by the average SIR (representing location) produces a marginal significance, and a fixed bias is sufficient to improve QoS. This implies that the CQI feedback delay is the main evil that must be mitigated. Our results showed that the biased AMC increases the system capacity by over a three-fold at 80 km/h.
- Key issues in supporting heterogeneous traffic in HSDPA were discussed, and were elaborated through system simulations. With the prospect for DiffServ support in UMTS, supporting heterogeneous traffic segregates into two key issues, i.e., to map the DiffServ traffic classes onto the priority queues in HSDPA, and to schedule the different priority queues. A simulation study was performed to develop a viable scheduler, in which three services were considered, i.e., voice over IP (VoIP), variable bit rate video streaming (MPEG), and web-browsing (HTTP). The proposed scheduler was shown to support the heterogeneous traffic efficiently, without causing any particular

service being a bottleneck in providing capacity. The simulation results further showed that the proposed scheduler is robust to changes in the traffic constitution, and supports call arrival rates of up to 3.5 calls/s in guaranteeing 5% outage to all the services in the default traffic scenario. This is substantially larger than some conventional schedulers, as strict priority queuing schemes support only up to 2.8 calls/s, whereas single queue schemes fail to secure 1 call/s, according to the simulation results.

- Sharing spectrum can be very attractive. For example, in rural areas UMTS coverage can be offered with much lower investment costs, but also in urban areas and hot spot areas capacity gains can be achieved, due to the increased trunking efficiency as channels are pooled together between the operators. Three admission control methods, which allow some operator resource usage control, were tested. A new method proposed, which uses the bit rate elasticity of TCP flows in an attempt to achieve a fair QoS between the operators, revealed to achieve the best fairness, but not any impressively higher fairness than the reference method. It also gives the best capacity, and the highest bit rate for PS services. Furthermore, considering RAN sharing on a *mult-cell* basis revealed to be a promising solution. It provides higher gain when the percentages of users per operator in each hotspot are different.
- A resource reservation algorithm that makes use of location-aware techniques in order to assure service to high priority users has been proposed. The proposed algorithm optimizes the existing trade-off between the dropping probability and the blocking probability in a WCDMA system. The proposed algorithm takes advantage of the predictability in the movement of users along a main road in order to determine the most adequate instant of time when the resource reservation for handover users should be made. A too large reservation region may cause that a handover user ends its connection before starting the handover procedure, resulting in a high false reservation ratio. Moreover, the blocking probability of new connection requests would increase because a large number of resources would be devoted to reservations for users inside the reservation region. On the other hand, a too small reservation region increases the number of handover failures (i.e. the dropping ratio) because there is not time enough to obtain the available resources. Then, an optimization of the reservation distance has been made by minimizing the system GoS. Moreover, it has been shown that, for shorter call durations, the reservation distance must be lower in order to reduce the false reservation probability. Similarly, higher service bit rate require higher reservation distance because more time is needed in order to obtain the required resources.

GERAN

- The most important radio resource management mechanisms offered by GPRS/EGPRS to support video streaming services have been identified and investigated, analyzing the main procedures related to QoS handling and channel administration for the services belonging to the streaming class, as well as all the standard features and parameters of the system able to affect the quality of these types of services. The aim of the carried out simulation analysis was to set the value of all the most important parameters which are not imposed by the GPRS specifications, in order to optimize the performances both of the network and the streaming clients. The simulation results show that low-loaded GPRS cells are able to support video streaming with an offered throughput from 24 kbps to 30 kbps, even in presence of GSM users (if a sufficient amount of radio resource are available in the cell and these ones are managed in the most appropriate way). On the contrary, when the total amount of radio resource offered by the cell is not enough to support

all the active users (GSM and GPRS), the level of decrease in the QoS of the streaming sessions has been measured. Other simulation results stress the importance to consider some specific mechanisms able to maximize the quality perceived by the video streaming end-user, like client-side pre-bufferization and header compression techniques.

- In particular we found very useful to introduce a new parameter that synthetically defines the satisfaction degree of a user on a scale from 0 to 1; this parameter is linked to the percentage of lost bytes and to the number of blocks. The general idea is that interruptions during a session are perceived as a major clue of a poor quality system, whereas the percentage of lost bytes, even if important, has a lower weight because it does not prevent from enjoying the video. We also define a “fully satisfied” user of a single session the user whose session has no interruptions at all and with a lost bytes probability $\leq 10^{-5}$ (user satisfaction = 1.0).
- The situation of blocking for higher bit rates shows that the blockage is nearly irrelevant for codec bit rate less or equal to 28 kbps whereas for greater bit rates it grows with a linear law. As a consequence, most of subsequent simulation experiments are done by using codec bit rate not greater than 32 kbps, which is assumed to be the limit value for bit rates of the application codec.
- Another important degree of freedom, which has been tuned in our simulation experiments, is the RTP packet size, which is the amount of segmentation introduced by RTP protocol. Experiments have shown that by configuring a dimension of RTP packet size equal to 500 bytes will provide high offered throughput while maintaining low jitter level.
- Another important parameter to be considered for the support of video streaming services is the limitation of the receiver initial buffering time. In order to determine the minimum initial buffering time, we must assume an acceptable level for user satisfaction. From the results we appreciate that, taking into account the desired levels of user satisfaction (about 90%), a safe value for the initial buffering time must be 5 seconds.

WLAN

- A novel Admission Control policy for IEEE 802.11 family of standards has been introduced. The policy can be applied in order to make the wireless LANs able to support also real-time data services (i.e. conversational and streaming classes). For these services, a minimum amount of offered throughput must be guaranteed as well as a specified maximum level of mean delay for packet transmission. For these reasons, a specific amount of radio resources should be reserved for real-time services and, consequently, only a maximum number of users should be supported by the wireless LAN system. The analytical model for WLANs IEEE 802.11b/a/g used in the AC policy has been validated by means of simulations. In particular the behavior of the Distributed Coordination Function (DCF) was investigated with Basic Access and with the RTS/CTS handshaking and the effect of the different factors not considered in the analytical model, such as the AckTimeout, the EIFS interval, etc., was highlighted. In the simulations, additional performance statistics as respect to the throughput offered by the hot-spot have been collected; due to the rapid growth of the average delay as respect to the number of users, the admission control policies applied by the network, have to take into account this parameter especially when applied to users that require time-bounded services.

- On the other hand, the proposed Enhanced Distributed Admission Control algorithm for IEEE 802.11e is able to protect already active flows of continuous nature (conversational and streaming) and allows controlling low priority bursty traffic by means of minimum average guaranteed load. It provides a dynamic control of time spend on transmissions from each access category managed by Transmission Time Threshold parameter and controlled by AP. Moreover, the proposed algorithm only introduces slight changes in the current draft of IEEE 802.11e standard. In that sense its applicability is guaranteed.
- The benefits of using the Transmission Opportunity and fragmentation mechanisms have been evaluated and demonstrated through simulations. By means of this option destructive influence of downward transmission rate shift can be control at some extend, giving higher priority, through higher TXOP limit, to stations with high QoS requirements. Moreover, simulation results show that there exist optimum TXOP limit values in terms of system performance. However, as these values change with service mix distribution a novel dynamic TXOP configuration algorithm based on the number of packets in AP queues has been proposed. The developed algorithm provides TXOP limits that are within minimum most advantageous TXOP limits and hence quasi-optimal system performance is reached independently of system service mix. Finally, by using the TXOP limit to perform fragmentation of low priority packets, further enhancement of prioritization mechanism is obtain.

CRRM

- With respect to Common Radio Resource Management, a first contribution of the project has been the outline of a functional model for having a common management of the pool of radio resources in heterogeneous scenarios. In the presented model, the Common Radio Resource Management (CRRM) refers to the set of functions that are devoted to ensure a proper coordination between the different radio access networks to achieve the most efficient use of the available radio resources. Different approaches to the CRRM and RRM interaction have been presented, outlining the potential levels of coordination in the radio resource management decisions in the identified functionalities. Finally, the requirements in terms of interworking capabilities and considerations about the physical CRRM implementation have been detailed. According to this framework, and taking into account the scope and time-frame of the EVEREST project, the CRRM studies reported here assume the functionality split in which the CRRM takes charge of the RAT selection procedures, including both initial RAT selection and vertical handover, while the RAT-specific RRM algorithms, like admission control, congestion control or packet scheduling are executed locally at the RRM entities.
- A general policy-based framework for the specification of CRRM algorithms has been defined and different policies considering the service type as well as the fact that the users may be indoor or outdoor have been evaluated through simulations. It has been obtained that, in outdoor scenarios, VG basic policy turns into a higher throughput than VU, ensuring lower interactive packet delay. This is because of the higher efficiency for non-real time traffic transmission in UTRAN achieved in the VG case, since web browsing traffic is supported by means of dedicated channels whereas in VU a packet scheduling algorithm must be implemented in GERAN. In turn, in scenarios with a mix of indoor and outdoor users with different services, the performance of IN*VG policy improves when the voice load increases, the www load decreases and there is a high fraction of indoor users. On the contrary, for low voice loads and high www loads VG*IN achieves a better throughput. This suggests that

the suitable configuration of the RRM and CRRM entities according to specific policies depends on the existing traffic conditions and therefore it may be modified at e.g. different periods of the day.

- With respect to vertical handover, the interworking between horizontal and vertical handovers has been studied, with two considered approaches, namely the tight approach T-VHO, in which the vertical handover algorithm is executed at every time that a horizontal handover algorithm should be carried out, so that both possibilities are considered prior to taking a decision, and the loose approach L-VHO, in which the vertical handover algorithm is executed only when a horizontal handover fails or when a call dropping is about to occur due to bad propagation conditions. It has been shown that the traffic distribution among the considered RATs can be quite different with the two approaches. In that sense, the tight approach allows a better fulfilment of the initial RAT selection policy, due to having more chances to execute a vertical handover than in the loose case. Consequently, and for a service-based RAT selection policy, it has been observed that the tight approach offers a better performance in terms of lower delay for interactive www users than the loose approach because most of the www traffic is served through UTRAN. On the other hand, a higher number of vertical handover procedures are also required with the tight approach, which increases the signalling overhead.
- With respect to CRRM implementation, if the CRRM entity is implemented in every existing RNC/BSC, then HHO and VHO are tightly coupled in a natural way and the interaction is simply an internal matter of the RNC/BSC. On the contrary, if the CRRM entity is implemented only in some RNC/BSC, then delay in taking decisions plays a role and tends to impact more on T-VHO because of the signalling exchange required between the nodes where the CRRM and RRM entities reside. In turn, if the CRRM entity is implemented in a separate node of the network, then delay in taking decisions poses further constraints impacting even more on T-VHO, because in this case a signalling exchange with the CRRM entity is always required before a HHO.
- Load balancing (LB) is a possible guiding principle for resource allocation in which the RAT selection policy will distribute the load among all resources as evenly as possible. Taking this into account, the performance of load balancing principles in the RAT selection procedure has been covered and compared against a service-based policy.
 - With respect to the initial RAT selection based on load balancing without including VHO, results revealed a tight dependency between the suitability of load balancing RAT selection and service-class mixing. It has been shown that even though the overall throughput may increase with load balancing policies, this at the expense of interactive traffic performance degradation. Nevertheless, other service type mixings showed no type of throughput improvement at all.
 - In turn, the introduction of VHO capabilities allows higher flexibility in the allocation of multi-service users in a multi-access scenario. We have compared two initial RAT selection policies along with a tight approach for VHO procedures. Results indicate that no remarkable improvement is noted on the total aggregate throughput when using the LB policy as opposed to the VGVU policy. Moreover, with LB, interactive users undergo higher average packet delays which impact the user's perceived QoS. However, we have seen that load balancing procedures may improve the call dropping probability due to a more flexible allocation of users onto both RATs, which is also a key performance indicator to consider.

- The concept of network-controlled cell-breathing (NCCB) strategy for CRRM in heterogeneous TDMA/CDMA has been proposed. The envisaged algorithm achieves a reduction in the interference level of the CDMA system by controlling the effective CDMA cell radius through initial RAT selection and vertical handover policies. The strategy has been evaluated by means of system level simulations in a scenario with UTRAN and GERAN as two examples of access networks using the CDMA and FDMA/TDMA technologies.
 - When considering a single voice service, the NCCB strategy has been compared against a classical load balancing strategy that tries to keep the same load level in both RATs. Results reveal that a significant improvement in terms of capacity for both uplink and downlink is achieved with the proposed strategy. It has been also shown that, by a proper setting of the maximum path loss PLth, load balancing principles can also be achieved in NCCB, thus obtaining the benefits in terms of flexibility of load balancing while at the same time exhibiting a higher capacity than a pure load balancing.
 - When considering a mix between voice and www users, different combinations of the NCCB strategy with service-based policies have been tested. It has been observed that the best performance is achieved with the so-called NCCB_voice, corresponding to applying the NCCB strategy only to voice users and allocating www users in UTRAN according to the service-based policies. It has been also observed that the adequate setting of the threshold PLth depends on the existing traffic mix and the trade-off among www delay and voice BLER.
- The impact of multi-mode terminals in the framework of heterogeneous networks with an initial RAT selection policy has also been assessed. Results indicate degradation in terms of throughput introduced by the limited operation of single-mode terminals. By considering a reservation scheme in GERAN for interactive users, we can improve the average packet delay for such users. While for a high multi-mode terminal availability results indicated that the reservation scheme was not necessary, for lower multi-mode terminal availabilities this scheme improved both aggregated throughput and packet delay figures, particularly for high number of interactive users.
- The impact of CRRM strategies on TCP throughput has also been studied in a scenario including 3G with HSDPA, 2G and WLAN systems. Four different CRRM algorithms have been evaluated and compared against the manual RAT selection algorithm. Without HSDPA, throughput increases up to 40% have been observed thanks to CRRM. When introducing HSDPA capabilities, the improvement in 3G throughput leads to increases of up to 60%.

APPENDIX A

A.1 SIMULATION SETUP

In the dynamic system level simulator, Poisson arrivals of calls are assumed, with exponentially distributed call lengths having a 90 s mean. Voice activity was simulated with exponentially distributed spurt and gap durations having a mean of 1 s and 1.5 s, respectively. We assumed a constant delay from the source to NodeB for simplicity, thereby allowing the RTP packets of a single spurt to arrive at fixed intervals given by the AMR/RTP block size. We omitted any control packets other than the actual voice packets. The main system parameters are summarized in Table 70

Table 70 Simulation Parameters

Cell layout	7 cells, 3 sectors, wrap around
Site separation	1 km
Chip rate	3.84 Mc/s
Path loss exponent	3.44
Shadowing	spacially correlated lognormal (8 dB)
Path model	Vehicular-A
Mobile terminal velocity	3 km/h (unless otherwise noted)
Handover hysteresis	3 dB
Receiver noise figure	9 dB
Total Tx power	20 W
CPICH Tx power	2 W
HS-PDSCH spreading factor	16
Number of HS-PDSCH codes	10
HS-SCCH spreading factor	128
Number of HS-SCCH codes	4
AMC	25 MCS levels (68.5 kb/s to 7.2 Mb/s)
HARQ	6-channel SAW, Chase combining
CQI feedback delay	2 HS-PDSCH frames
Scheduling discipline	PF-LDF, $f = 1$ (unless otherwise noted)
AMR/RTP block size	80 ms (unless otherwise noted)
RTP packet due time	40 ms (unless otherwise noted)

A regular hexagonal cell layout was assumed with the so-called wrap around technique applied to avoid the boundary effect. The radio propagation was simulated as a concatenation of the Hata loss [20], lognormal shadowing (std. deviation= 8 dB, inter-site correlation = 0.5) [1], and multipath Rayleigh fading [22]. The AMC was performed using 25 MCS levels from 68.5 kb/s to 7.2 Mb/s (Category 8 in the 3GPP specification [18]), with a CQI feedback delay of 2 HSPDSCH frames. Moreover, the Chase combining HARQ [7] was modeled with 6-channel stop-and-wait (SAW) process per user.

The packet delay was measured per RTP packet, after reordering to fix the shuffle caused by the 6-SAW HARQ in the receiver. An RTP packet that took longer than its due time to transfer from NodeB to UE was considered as a packet loss. Moreover, any RTP packet that has not been fully or partly transmitted (and therefore not stored in one of the HARQ processes) was dropped at the NodeB. If a VoIP call experiences a packet loss rate (including dropping) of over 0.01, the call was considered as an outage. The call was dropped if any of the following occurred:

² The HARQ is unable to recover data after 20 retransmissions.

² After 10 s of initial monitoring period, the call experiences a packet loss rate over 0.1.

² The CPICH SIR falls below -10 dB for a consecutive 100 ms.

A dropped call was also counted as an outage.

A.2 SIR CALCULATIONS

The SIR was calculated per HS-PDSCH frame and was used to look up a frame error rate (FER) table, which was prepared for each MCS level by link level simulations, to generate random errors.

To derive an SIR calculation formula considering channel estimation error, we define variables:

c_k channel vector of k-th path ($\sum^k |c_k|^2 = 1$)

A_d HS-PDSCH received signal amplitude

A_c CPICH received signal amplitude

s_d HS-PDSCH symbol ($|s_d|^2 \equiv 1$)

s_c CPICH symbol ($|s_c|^2 \equiv 1$)

$n_{d,k}$ noise/interference vector of k-th path after despreading HS-PDSCH

$n_{c,k}$ noise/interference vector of k-th path after despreading CPICH

sf_d spreading factor of HS-PDSCH

sf_c spreading factor of CPICH

Note that all variables are functions of time, although we have omitted the time representation for simplicity. Using above notations the received HS-PDSCH of the k-th path after despreading is given by

$$r_{d,k} = A_d \cdot c_k \cdot s_d + n_{d,k} \quad (83)$$

and similarly for CPICH by

$$r_{c,k} = A_c \cdot c_k \cdot s_c + n_{c,k} \quad (84)$$

The channel estimate for the k-th path to demodulate $r_{d,k}$ is derived by averaging $r_{c,k}$ in the time vicinity. If we assume the channel is constant over this averaging interval, the estimated

channel vector \hat{c}_k is given by

$$\hat{c}_k = E[r_{c,k} \cdot s_c^*] = A_c \cdot c_k + \bar{n}_{c,k} \quad (85)$$

where $E[\cdot]$ indicates the ensemble average. The residual noise/interference component $\bar{n}_{c,k}$ is the channel estimation error. If we apply standard Gaussian approximations on the additive noise and interference, the variances of the vectors $n_{d,k}$, $n_{c,k}$ and $\bar{n}_{c,k}$ are given by

$$\sigma_{d,k}^2 \triangleq \text{Var}[n_{d,k}] = \frac{I_k}{sf_d} \quad \sigma_{c,k}^2 \triangleq \text{Var}[n_{c,k}] = \frac{I_k}{sf_c} \quad (86)$$

$$\sigma_{c,k}^2 \triangleq \text{Var}[\bar{n}_{c,k}] = \frac{I_k}{sf_c \cdot m} \quad (87)$$

$$\sigma_{c,k}^2 \triangleq \text{Var}[\bar{n}_{c,k}] = \frac{I_k}{sf_c \cdot m} \quad (88)$$

respectively, where m is the number of P-CPICH symbols in the ensemble of (85), and I_k is the wideband noise/interference power on the k-th path.

The rake output is thus given by

$$\begin{aligned}
z &= \sum_k r_{d,k} \cdot \hat{c}_k^* \\
&= A_d \cdot A_c \sum_k |c_k|^2 s_d + A_c \sum_k c_k^* \cdot n_{d,k} \\
&\quad + A_d \sum_k c_k \cdot \bar{n}_{c,k}^* + \sum_k n_{d,k} \cdot \bar{n}_{c,k}^*.
\end{aligned} \tag{89}$$

The first term indicates the desired signal and the second term represents the noise and interference. The third and fourth terms represent the additional noise caused by the channel estimation error. Hence, we can calculate the received instantaneous SIR, namely γ_d , by

$$\gamma_d = \frac{A_d^2 \cdot A_c^2 (\sum_k |c_k|^2)^2}{A_c^2 \sum_k |c_k|^2 \sigma_{d,k}^2 + A_d^2 \sum_k |c_k|^2 \bar{\sigma}_{c,k}^2 + \sum_k \sigma_{d,k}^2 \bar{\sigma}_{c,k}^2}. \tag{90}$$

The Chase combining HARQ combines the repetitive frames coherently with the original frame symbol-by-symbol at the rake output. The resulting signal is given by

$$\begin{aligned}
z &= \sum_q \sum_k r_{d,q,k} \cdot \hat{c}_{q,k}^* \\
&= \sum_q A_{d,q} \cdot A_{c,q} \sum_k |c_{q,k}|^2 s_d + \sum_q A_{c,q} \sum_k c_{q,k}^* \cdot n_{d,q,k} \\
&\quad + \sum_q A_{d,q} \sum_k c_{q,k} \cdot \bar{n}_{c,q,k}^* + \sum_q \sum_k n_{d,q,k} \cdot \bar{n}_{c,q,k}^*
\end{aligned} \tag{91}$$

where the suffix q indicates the repetitive frames. Consequently, γ_d after HARQ combining is given by

$$\gamma_d = \frac{\left\{ \sum_q \left(A_{d,q}^2 \cdot A_{c,q}^2 \sum_k |c_k|^2 \right) \right\}^2}{\sum_q \left\{ \begin{aligned} &A_{c,q}^2 \sum_k |c_k|^2 \sigma_{d,q,k}^2 \\ &+ A_{d,q}^2 \sum_k |c_k|^2 \bar{\sigma}_{c,q,k}^2 \\ &+ \sum_k \sigma_{d,q,k}^2 \bar{\sigma}_{c,q,k}^2 \end{aligned} \right\}}. \tag{92}$$

A.3 CQI ESTIMATION

The CQI is derived by estimating the received SIR of the CPICH. From this SIR estimate, the mobile terminal selects the MCS that satisfies an FER of 0.1 on the HS-PDSCH, and reports the corresponding CQI to the serving base station [18]. In a practical mobile terminal, the signal power and the noise/interference power are estimated individually, and their ratio is calculated to obtain the SIR.

The signal power estimate \hat{S} is given by

$$\hat{S} = \left(\sum_k |\hat{c}_k|^2 \right). \tag{93}$$

We can generate $|c_k|^2$ in a Monte-Carlo simulation by

$$|\hat{c}_k|^2 = A_c^2 |c_k|^2 + 2A_c |c_k| |\bar{n}_{c,k}| \cos \theta_k + |\bar{n}_{c,k}|^2 \tag{94}$$

where θ_k is the angle between the vectors c_k and $n_{c,k}$, which is uniformly distributed in the range $(0, 2\pi]$. The noise/interference power $|n_{c,k}|^2$ follows an exponential distribution having a mean $I_k / (sf_c \cdot m)$.

The noise/interference power estimate \hat{I} is given by

$$\hat{I} = \sum_k |\hat{c}_k|^2 \cdot \hat{I}_k \tag{95}$$

where

$$\begin{aligned} \hat{I}_k &= E [|r_{c,k} - \hat{c}_k|^2] \\ &= E [|n_{c,k} - \bar{n}_{c,k}|^2]. \end{aligned} \tag{96}$$

Taking into account that $n_{c,k}$ itself is included in the ensemble average $\bar{n}_{c,k}$, after some mathematics, we obtain

$$\hat{I}_k = \frac{I_k}{sf_c} \cdot \left(1 - \frac{1}{m} \right) \cdot \frac{u_\Gamma(w)}{w} \tag{97}$$

where w is the ensemble size in (96), and $u_i(w)$ is a gamma distributed random variable having w -degree of freedom.

Finally, the P-CPICH SIR estimate, namely $\hat{\gamma}_c$, is calculated by

$$\hat{\gamma}_c = \hat{S} / \hat{I}. \tag{98}$$

The SIR estimate of the HS-PDSCH, namely $\hat{\gamma}_d$, is given by scaling $\hat{\gamma}_c$ by the relative transmission powers and spreading factors.

APPENDIX B

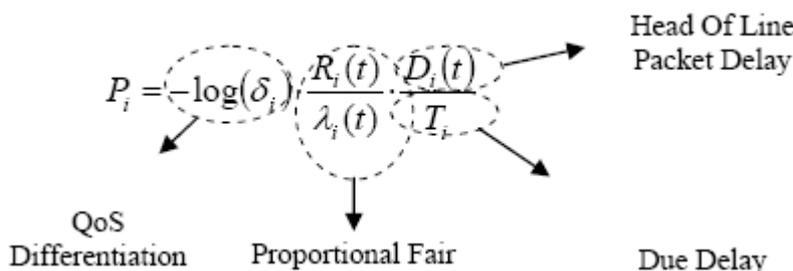
Proportional Fair (PF): Proportional Fair scheduler serves the user with largest relative channel quality, in which the priority is given as:

$$P_i = \frac{R_i(t)}{\lambda_i(t)} \quad i = 1, \dots, N$$

where $P_i(t)$ denotes the i th user priority at time t , $R_i(t)$ is the instantaneous data rate experienced by the user i , and λ_i is the user's throughput.

This algorithm intends to serve users under very favourable instantaneous radio channel conditions relative to their average ones, thus taking advantage of the temporal variations of the fast fading channel. In theory presented in [70], the instantaneous rate $R_i(t)$ is linear with the instantaneous Eb/No. In HSDPA, due to the discrete property of AMC, it does not hold. However, this relationship still can be reflected by the relationship between the CQI and MCS levels.

M-LWDF: The so-called Modified Largest Weighted Delay First (M-LWDF) is well studied in [71], which is developed based on proportional fairness, but also takes each flow delay property and QoS such as dropping rate requirements into account, as shown in the following picture,



Note that the term $D_i(t)/T_i$ ranges from zero to one, becoming closer to the latter when the head of line packet delay approaches the due delay. This algorithm does not only take advantage of the multiuser diversity available in the shared channel through the Proportional Fair algorithm, but also increases the priority of flows with head of line packets close to their deadline violation. In this way, the algorithm is able to control the flow delays to provide QoS for real time traffic or streaming type of services which has certain delay requirements. The algorithm also adds a QoS differentiation term (δ_i in our study, outage probability requirements) to imbalance the priority between users with different application demands on error rate.

7 REFERENCES

- [1] O. Sallent (editor) et al. "First report on the evaluation of RRM/CRRM algorithms", Deliverable D11 of the EVEREST IST-2002-001858 project, November, 2004. Available at <http://www.everest-ist.upc.es/>.
- [2] H. Holma, A. Toskala (editors), W-CDMA for UMTS, John Wiley and Sons, 2000.
- [3] K.S. Gilhousen, I.M. Jacobs, R. Padovani, A.J. Viterbi, L.A. Weaver, "On the Capacity of a Cellular CDMA System," IEEE Trans. Veh. Technol., vol 40. pp.303-312, May 1991.
- [4] M.G. Jansen and R. Prasad, "Capacity, throughput, and delay analysis of a cellular DS-CDMA system with imperfect power control and imperfect sectorization," IEEE Trans. Veh. Technol., vol.44, pp. 67-75, Feb 1995.
- [5] F.D. Priscoli and F. Sestini, "Effects of imperfect power control and user mobility on a CDMA cellular network," IEEE J-SAC, vol. 14, no. 9, pp1809-1817, Dec 1996.
- [6] E. Kudoh, "On the capacity of DS/CDMA cellular mobile radios under imperfect transmitter power control," IEICE Trans. Commun., vol.E76-B, no.8, pp886-892, Aug 1993.
- [7] B. Hashem, and E.S. Sousa, "Reverse-link capacity and interference statistics of a fixed-step power-controlled DS/CDMA system under slow multi-path fading," IEEE Trans. Commun, 1999, 47, (12), pp. 1905–1912.
- [8] D.K. Kim, and D.K. Sung, "Capacity estimation for an SIR-based power-controlled CDMA system supporting ON-OFF traffic," IEEE Trans. Veh Technol., 2000, 49, (7), pp 1094-1101
- [9] S. Manji and W. Zhuang, "Power control and capacity analysis for a packetized indoor multimedia DS-CDMA network," IEEE Trans. Veh. Technol., vol. 49, no.3, pp. 67-74, May 2000.
- [10] D.K. Kim, and F. Adachi, "Theoretical Analysis of Reverse Link Capacity for an SIR-based power-controlled cellular CDMA system in a multipath fading environment," IEEE Trans. Veh. Technol., vol. 50, no.2, pp452-464, March 2001.
- [11] 3GPP TS 25.331 v5.0.0 "RRC Protocol Specification", March 2002.
- [12] J. Pérez-Romero, O. Sallent, R. Agustí, "Impact of User Location in W-CDMA Downlink Resource Allocation", IEEE 7th International Symposium on Spread Spectrum Technologies and Applications, Prague, Czech Republic, Sept 2-5, 2002
- [13] Kim, D., et al, "Pilot Power Control and Service Coverage Support in CDMA Mobile Systems", Proceedings of the 49th IEEE Vehicular Technology Conference, pp 1464-1468, 1999.
- [14] H. Zhu, T. Buot, R. Nagaïke, S. Harmen, "Load Balancing in WCDMA Systems by adjusting Pilot Power", International Symposium on Wireless Personal Mobile Communications, Vol.3, pp: 936-940, October 2002
- [15] Valkealahti K., Höglund A., Parkkinen J., Flanagan A., "WCDMA common pilot power control with cost function minimization", Proc. IEEE Vehicular Technology Conference, Vancouver, Canada, October 2002.
- [16] EVEREST D13 – "Target Scenarios specification: vision at project stage 2". (already listed in ref section)
- [17] EVEREST IST-2002-001858, D15, "Report on the evaluation of RRM/CRRM algorithms"

- [18] 3GPP TS 25.304, UE Procedures in Idle Mode and Procedures for Cell Reselection in Connected Mode
- [19] Performance Study for a Microcell Hot Spot Embedded in Macrocell CDMA Systems”, IEEE Transactions on Vehicular Technology, January 1999.
- [20] J. Laiho, A. Wacker, T. Novosad, “Radio Network Planning and Optimisation for UMTS“, John Wiley & Sons LTD
- [21] T. Rautiainen, “Breaking the Hierarchical Cell Structure in WCDMA Networks”, IEEE 55th VTC Spring conference, Birmingham, USA, 2002.
- [22] I. Jami, H. Tao, “Micro-cell planning within macro-cells in UMTS: downlink analysis”, 3rd International conference on 3G communication technologies (3G 2002), London, 2002.
- [23] EVEREST IST-2002-001858, D12,
- [24] P. Karlsson et al., D05 *Target Scenarios specification: vision at project stage 1*, Deliverable of the EVEREST project IST-2002-001858 (available at <http://www.everest-ist.upc.es/>)
- [25] An Approach for Inter-Operator Spectrum Sharing for 3G Systems and beyond, M. K. Pereirasamy, J. Luo, M. Dillinger, C. Hartmann. In Proceedings of the 15th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC'2004, Barcelona, Spain, Sept 2004.
- [26] M. K. Pereirasamy, J. Luo, M. Dillinger, C. Hartmann, “An Approach for Inter-Operator Spectrum Sharing for 3G Systems and Beyond”, 15th IEEE International Symposium on Personal Indoor and Mobile Radio Communications 2004 (PIMRC'04), september 2004.
- [27] M. K. Pereirasamy, J. Luo, M. Dillinger, C. Hartmann, “Dynamic Inter-Operator Spectrum Sharing for UMTS FDD with Displaced Cellular Networks”, IEEE Wireless Communications and Networking Conference 2005 (WCNC'05), march 2005.
- [28] J. Pérez-Romero, O. Sallent, R. Agustí, M.A. Díaz-Guerra, “Radio Resource Management Strategies in UMTS”, Ed. Wiley and Sons, 2005.
- [29] K. Johansson, M. Kristensson, U. Schwarz, “Radio Resource Management in Roaming Based Multi-Operator WCDMA Networks”, 59th IEEE Vehicular Technology Conference 2004 (VTC-spring '04), may 2004.
- [30] TR 101 112 V3.2.0 (1998-2004) Selection procedures for the choice of radio transmission technologies of the UMTS (UMTS 30.03 version 3.2.0).
- [31] 3GPP TR 34.108 v4.2.1 “Common Test Environments for User Equipment Conformance Testing”, March 2002.
- [32] J.J. Olmos, S. Ruiz, “Transport Block Error Rates for UTRA-FDD Downlink with Transmission Diversity and Turbo Coding”, PIMRC-2002, Vol 1, 2002, pp. 31-35.
- [33] Lee W.C.Y, Lee D.J.Y., “The Impact of Repeaters on CDMA System Performance”, in Proceedings of VTC 2000 Tokio, Spring 2000.
- [34] S. Blake, et. al., “An architecture for differentiated services,” *RFC 2475, Internet Request for Comments*, Dec. 1998
- [35] Markku J. Heikkilä, Kari Majonen, “INCREASING HSDPA THROUGHPUT BY EMPLOYING SPACE-TIME EQUALIZATION”, PIMRC 2004, Barcelona, Spain.
- [36] Johan Gystavsson and Johan Lewin, “Impact of MAC and RLC Protocols on UMTS release 5”, M.Sc thesis, Lund Institute of Technology, November 2003.
- [37] IST-2002-001858: EVEREST, D07, “Simulation tools: inherited features and newly implemented capabilities”, [http:// www.everest-ist.upc.es/](http://www.everest-ist.upc.es/).

- [38] Xiaoxin Wang, "3G HSDPA Performance in Mobile Internet Connections", M.Sc thesis, KTH, March 2004.
- [39] 3GPP, "TR25.214: Physical Layer Procedures", Version 6.3.2, 2004-09.
- [40] EVEREST IST-2002-001858, WP3-PTIN-I00-Int-1.doc, "HSDPA Link Level Simulator Description", 10/2004
- [41] 3GPP, "TR25.848: Physical layer aspects of UTRA High Speed Downlink Packet Access", Version 4.0.0, March/2001.
- [42] 3GPP, "Feasibility Study for Orthogonal Frequency Division Multiplexing (OFDM) for UTRAN enhancement", TR25.892, v6.0.0
- [43] D. Chase, "Code combining - a maximum-likelihood decoding approach for combining an arbitrary number of noisy packets," IEEE Trans. Commun., vol. 33, no. 5, pp. 385-393, May 1985.
- [44] M. Gudmundson, "Correlation model for shadow fading in mobile radio systems," IEE Electronics Letters, vol. 27, no. 23, pp. 2145-2146, Nov. 1991.
- [45] W. C. Jakes, Microwave Mobile Communications, John Wiley & Sons, New York, 1974.
- [46] M. Iwamura, L. Wang, V. Friderikos, and A. H. Aghvami, "Channel estimation error compensated adaptive modulation/ coding to support VoIP over HSDPA," IEEE Trans. Veh. Technol., submitted.
- [47] ETSI, TR 101 112, "Selection procedures for the choice of radio transmission technologies of the UMTS," UMTS 30.03, April 1998.
- [48] 3GPP, TS 25.214, "Physical layer procedures (FDD)," v6.6.0, June 2005.
- [49] L. Kleinrock, "Queueing Systems-Volume I: Theory ,", 1976,Wiley.
- [50] T. Moulosley, "Performance of UMTS high speed downlink packet access for data streaming applications," IEE Proc.3G2002, pp. 302-307, London, May 2002.
- [51] P. Ameigeiras, Packet scheduling and quality of service in HSDPA, Ph.D. Thesis, Aalborg Univ., Oct. 2003.
- [52] 3GPP, TS 23.207, "End-to-end quality of service (QoS) concept and architecture," v6.5.0, June 2005.
- [53] M. Iwamura, L. Wang, V. Friderikos, and A. H. Aghvami, "Biased adaptive modulation/coding to provide VoIP QoS over HSDPA," Proc. IST Mobile &Wireless Communications Summit 2005, Dresden, June 2005.
- [54] 3GPP, TS 25.321, "Medium access control (MAC) protocol specification," v6.5.0, June 2005.
- [55] 3GPP, TS 25.322, "Radio link control (RLC) protocol specification," v6.4.0, June 2005.
- [56] 3GPP, TS 25.323, "Packet data convergence protocol (PDCP) specification," v6.2.0, June 2005
- [57] R. Love, A. Ghosh, R. Nikides, L. Jalloul, M. Cudak, and B. Classon, "High speed downlink packet access performance," IEEE Proc. VTC Spring 2001, vol. 3, pp. 2234-2238, Rhodes, May 2001.
- [58] Y. Ofuji, S. Abeta, and M. Sawahashi, "Comparison of packet scheduling algorithms focusing on user throughput in high speed downlink packet access," IEICE Trans. Commun., vol. E86-B, no. 1, pp. 132-141, Jan. 2003.
- [59] 3GPP, TR 25.933, "IP transport in UTRAN," v5.4.0, Dec. 2003.

- [60] 3GPP, TS 23.207, "End-to-end quality of service (QoS) concept and architecture," v6.5.0, June 2005.
- [61] S. Blake, D. Black, M. Carlton, E. Davies, Z. Wang, and W. Weiss, "An architecture for differentiated services," IETF RFC 2475, Dec. 1998.
- [62] R. Love, A. Ghosh, R. Nikides, L. Jalloul, M. Cudak, and B. Classon, "High speed downlink packet access performance," IEEE Proc. VTC Spring 2001, vol. 3, pp. 2234-2238, Rhodes, May 2001.
- [63] Y. Ofuji, S. Abeta, and M. Sawahashi, "Comparison of packet scheduling algorithms focusing on user throughput in high speed downlink packet access," IEICE Trans. Commun., vol. E86-B, no. 1, pp. 132-141, Jan. 2003.
- [64] M. Iwamura, L. Wang, V. Friderikos, and A. H. Aghvami, "Channel estimation error compensated adaptive modulation/ coding to support VoIP over HSDPA," IEEE Trans. Veh. Technol., submitted.
- [65] ETSI, TR 101 112, "Selection procedures for the choice of radio transmission technologies of the UMTS," UMTS 30.03, April 1998.
- [66] M. Hata, "Empirical formula for propagation loss in land mobile radio services," IEEE Trans. Veh. Technol., vol. 29, no. 3, pp. 317-325, Aug. 1980.
- [67] COST231, Digital mobile radio: towards future generation systems, Final Report, EUR18957, Ch. 4, 1999.
- [68] M. Gudmundson, "Correlation model for shadow fading in mobile radio systems," IEE Electronics Letters, vol. 27, no. 23, pp. 2145-2146, Nov. 1991.
- [69] 3GPP, TS 25.214, "Physical layer procedures (FDD)," v6.6.0, June 2005.
- [70] Holtzman J.M. CDMA Forward Link Waterfilling Power Control. Vehicular Technology Conference, 2000. VTC 2000 Spring. Volume 3. pp. 1663-1667.
- [71] Andrews M., et al. Providing Quality of Service over a Shared Wireless Link. IEEE Communications Magazine, Volume 39, Feb 2001. pp. 150-154.
- [72] Y. Ishikawa, T. Hayashi, and S. Onoe, "W-CDMA downlink transmit power and cell coverage planning," IEICE Trans. Commun., vol. E85-B, no. 11, Nov. 2002.
- [73] 3GPP, TS 25.213, "Spreading and modulation (FDD)," v6.0.0, Dec. 2003.
- [74] J. H. Horng, G. Vannucci, and Z. Jinyu, "Capacity enhancement for HSDPA in W-CDMA system," IEEE Proc. VTC, vol. 2, pp. 661-665, Vancouver, Sept. 2002.
- [75] Grosicki E., Abed-Meraim K., Loubaton P., Chaufray J-M., "Comparion of Downlink Mobile Positioning Methods for the UMTS FDD Mode without using IPDL periods", Proceedings of the 7th International Symposium on Signal Processing and its Applications, vol: 2, pp_347-350, july 2003
- [76] Spiegel S.J., Kovacs I.G., "An efficient integration of GPS and WCDMA radio front-ends", IEEE Transactions on Microwave Theory and Techniques, vol: 52, issue 4, april 2004.
- [77] Y. Ma, J.J. Han, K.S. Trivedi, "Call Admission Control for Reducing Dropped Calls in Code Division Multiple Access (CDMA) Cellular Systems", IEEE Infocom 2000.
- [78] J.W. Chang, D.K. Sung, "Adaptive Channel Reservation Scheme for Soft Handoff in DS-CDMA Cellular Systems", IEEE Transactions on Vehicular Technology, Vol 50., Issue 2, pp. 341-353, march 2001.
- [79] S. Choi, G. Shin, "Predictive and adaptive bandwidth reservation for handoffs in QoS-sensitive cellular networks", Proceedings of ACM SIGCOMM'98, pp: 155-166.

- [80] F.D. Priscoli, F. Sestini, "Fixed and adaptive blocking thresholds in CDMA cellular networks", IEEE Personal Communications, vol. 5, april 1998.
- [81] M.H. Chiu, M.A. Bassiouni, "Predictive Schemes for Handoff Prioritization in Cellular Networks Based on Mobile Positioning", IEEE Journal on Selected Areas in Communications, Vol 18, Issue 3, pp: 510-522, march 2000.
- [82] S. Naghian, "Hybrid Predictive Handover in Mobile Networks" IEEE Vehicular Technology Conference 2003 (VTC-03), Vol 3, pp: 1918-1922, october 2003 .
- [83] 3GPP TS 25.213 v5.0.0, "Spreading and Modulation (FDD)", march 2002.
- [84] 3GPP, General Packet Radio Service (GPRS); Service description; Stage 2, 3GPP TS 23.060 version 6.6.0 Release 6, September 2004.
- [85] H. G. Perros and K. M. Elsayed, Call admission control schemes: a review, IEEE Communications Magazine, Volume: 34, Issue: 11, Nov. 1996, Pages:82 – 91.
- [86] 3GPP, General Packet Radio Service (GPRS); Mobile Station (MS) - Base Station System (BSS) interface; Radio Link Control/Medium Access Control (RLC/MAC) protocol, 3GPP TS 44.060 version 6.9.0 Release 6, September 2004.
- [87] EVEREST IST-2002-001858 D14 – "Simulation tools: final version capabilities and features"
- [88] J.L. Sobrinho and A.S. Krishnakumar, "Real-time traffic over the IEEE 802.11 medium access control layer," Bell Labs Technical Journal, pages 172-187, Autumn 1996.
- [89] A. Banchs and X. Perez, "Distributed Weighted Fair Queuing in 802.11 Wireless LAN", IEEE ICC '02, vol. 5, Apr. 2002, pp. 3121-27
- [90] N. Vaidya, P. Bahl and S. Gupta "Distributed Fair Scheduling in a Wireless LAN", Mobicom 2000, Boston, USA, August 2000, 167-178
- [91] MARCHETTA, Stefano. "Standard Wireless LAN IEEE 802.11: valutazione analitica delle prestazioni del protocollo di accesso in condizioni di traffico asimmetrico". Tesi di Laurea Università degli Studi Roma Tre, Facoltà di Ingegneria, Corso di Studi in Ingegneria Elettronica, 2001.
- [92] K. Pahlavan, A. H. Levesque, "Wireless Data Communications", Proc. of the IEEE, Vol. 82, No. 9, September 1994, pp. 1398-1430.
- [93] A. De Simone, S. Nanda, "Wireless Data: Systems, Standards, Services", Journal of Wireless Networks, Vol. 1, No. 3, February 1996, pp. 241-254.
- [94] ANSI/IEEE, 802.11-1999, "Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications", IEEE.
- [95] L. Kleinrock, F. Tobagi, "Packet Switching in radio channels, part II - the Hidden Terminal Problem in Carrier Sense Multiple Access and the Busy Tone Solution", IEEE Trans. Comm., Vol. COM-23, No. 12, December 1975, pp. 1417-1433.
- [96] J. Weinmiller, M. Schlager, A. Festag and A. Wolisz, "Performance study of access control in wireless LANs IEEE 802.11 DFWMAC and ETSI RES 10 HIPERLAN", Mobile networks and applications, vol. 2, no. 1, pp. 55-67, 1997.
- [97] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function", IEEE Journal on Selected Areas in Communications, vol. 18, no. 3, pp. 535-547, 2000.
- [98] H. Wu, Y. Peng, K. Long, S. Cheng and J. Ma, "Performance of reliable transport protocol over IEEE 802.11 wireless LAN: analysis and enhancement", Proceedings of

IEEE INFOCOM '02, New York, USA, 2002.

- [99] R. Litjens, F. Roijers, J.L. van den Berg, R.J. Boucherie and M. Fleuren, "Performance analysis of wireless LANs: an integrated packet/flow level approach", Memorandum No. 1676, Department of Applied Mathematics, Faculty of EEMCS, University of Twente, The Netherlands, May, 2003.
- [100] Stefano Marchetta, " Standard Wireless LAN IEEE 802.11: valutazione analitica delle prestazioni del protocollo di accesso in condizioni di traffico asimmetrico", Tesi di laurea, Facoltà di Ingegneria, Università degli Studi Roma Tre, 2002.
- [101] IEEE 802.11e/D4.0 Draft Supplement to Part 11: Wireless Medium Access Control (MAC) and physical layer (PHY) specifications: Medium Access Control (MAC) Enhancements for Quality of Service (QoS), November 2002.
- [102] Y. Xiao and H.Li "Evaluation of Distributed Admission Control for the IEEE 802.11e EDCA", IEEE Radio Communications, September 2004
- [103] D. Gu and J.Zhang, "A New Measurement-based Admission Control Method for IEEE 802.11 Wireless Local Area Networks ", Mitsubishi Elec. Research Lab., Tech report TR-2003-122, Oct. 2003
- [104] M. Barry, A.T. Campbell, and A. Veres, "Distributed Control Algorithms for Service Differentiation in Wireless Packet Networks", Proc. IEEE INFOCOM '01, vol. 1, Anchorage, AK, 2001 pp.582-90
- [105] L. Zhang and S. Zeadally, "HARMONICA: Enhanced QoS Support with Admission Control for IEEE 802.11 Contention-based Access", Proc. IEEE RTAS '04, Toronto, Canada, May 2004, pp.64-71
- [106] D. Pong and T. Moors "Call Admission Control for IEEE 802.11 Contention Access Mechanism", Proc. IEEE GLOBECOM'03, vol. 1, San Francisco, CA, Dec.2003, pp.174-78
- [107] A. Banchs, X. Perez-Costa and D. Qiao, "Providing Throughput Guarantees in IEEE 802.11e Wireless LANs", Proc. 18th Int. Teletraffic Cong., Berlin, Germany, Sept.2003
- [108] J. Majkowski, F. Casadevall "Admission Control in IEEE 802.11e EDCA", IWS'05, Aalborg, Denmark Sept. 2005
- [109] A. Grilo, M. Macedo and M. Nunes "A scheduling Algorithm for QoS Support in IEEE802.11e Networks", IEEE wireless Communications, June 2003
- [110] Nitin H. Vaidya, Paramvir Bahl, Seema Gupta, "Distributed fair scheduling in a wireless LAN," Proceedings of the sixth annual international conference on Mobile computing and networking, p.167-178, August 06-11, 2000, Boston, Massachusetts, United States.
- [111] J.L. Sobrinho and A.S. Krishnakumar, "Real-time traffic over the IEEE 802.11 medium access control layer," Bell Labs Technical Journal, pages 172-187, Autumn 1996.
- [112] I. Aad, C. Castelluccia, "Differentiation mechanisms for IEEE 802.11," Proceedings. of IEEE INFOCOM 2001, vol. 1, pp. 209 -218, April 2001.
- [113] ANSI/IEEE, 802.11-1999, "Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications", IEEE.
- [114] Martin Devera, "Hierarchical Token Bucket", <http://luxik.cdi.cz/~devik/qos/htb/>
- [115] Bert Hubert, Gregory Maxwell, Remco van Mook, Martijn van Oosterhout, Paul B Schroeder, Jasper Spaans, Pedro Larroy, "Linux Advanced Routing & Traffic Control HOWTO", <http://lartc.org/howto/>

- [116] Martin Heusse, Franck Rousseau, Gilles Berger-Sabbatel, Andrzej Duda, "Performance anomaly of 802.11b", INFOCOM 2003.
- [117] W. Pattara-Atikom, S. Banerjee, and P. Krishnamurthy, "Starvation Prevention and Quality of Service in Wireless LANs", Proc. 5th Int'l. Symp. Wireless Pers. Multimedia Commun., Oct. 2002
- [118] ANSI/IEEE Std 802.1D, 1998 Edition IEEE Standard for Information technology. Telecommunications and information exchange between systems. Local and Metropolitan area networks. Common specifications Part 3: Media Access Control (MAC) Bridges
- [119] Kishor S. Trivedi "Probability and Statistics with Reliability, Queuing, and Computer Science Applications", Prentice-Hall, 1982.
- [120] 3GPP TR 43.901. "Feasibility study on Generic Access to A/Gb interface"
- [121] 3GPP TR 25.881 v5.0.0 "Improvement of RRM across RNS and RNS/BSS"
- [122] 3GPP TR 25.891 v0.3.0 "Improvement of RRM across RNS and RNS/BSS (Post Rel-5) (Release 6)"
- [123] J. Pérez-Romero, O. Sallent, R. Agustí, P. Karlsson, A. Barbaresi, L. Wang, F. Casadevall, M. Dohler, H. González, F. Cabral-Pinto "Common Radio Resource Management: Functional Models and Implementation Requirements", 16th Annual IEEE International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC), Berlin, 2005.
- [124] F. Cabral-Pinto (editor) "Testbed Specification" , Deliverable D06 of the EVEREST IST-2002-001858 project, April, 2004. Available at <http://www.everest-ist.upc.es/>
- [125] A. Tölli, P. Hakalin, H. Holma, "Performance Evaluation of Common Radio Resource Management (CRRM)", IEEE International Conference on Communications (ICC 2002), Vol. 5, April, 2002, pp. 3429-3433.
- [126] R. Agusti, O. Sallent, J. Pérez-Romero, L. Giupponi "A Fuzzy-Neural Based Approach for Joint Radio Resource Management in a Beyond 3G Framework", 1st International Conference on Quality of Service in Heterogeneous Wired/Wireless Networks, Qshine'04, Dallas, USA, October, 2004.
- [127] R. Mukerjee, J. Luo, M. Dillinger, E. Mohyeldin "Joint Scheduling Algorithms over Heterogeneous networks in a Reconfigurable Environment", 5th European Personal Mobile Communications Conference, April, 2003, pp. 591-596.
- [128] G. Fodor, A. Furuskar, and J. Lundsjo, "On access selection techniques in always best connected networks," in ITC Specialist Seminar on Performance Evaluation of Wireless and Mobile Systems, Aug. 2004.
- [129] J. Pérez-Romero, O. Sallent, R. Agustí, "Policy-based Initial RAT Selection algorithms in Heterogeneous Networks" accepted at MWCN '05. Marrakech-Morocco 19-21 Sept. 2005.
- [130] 3GPP TR 25.942 "Radio Frequency (RF) system scenarios"
- [131] UMTS 30.03 v3.2.0 TR 101 112 "Selection procedures for the choice of radio transmission technologies of the UMTS", ETSI, April, 1998.
- [132] J.J. Olmos, S. Ruiz, "Transport Block Error Rates for UTRA-FDD Downlink with Transmission Diversity and Turbo Coding", 13th IEEE International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC), Lisbon, 2002, pp. 31-35.
- [133] T. Halonen, J. Romero, J. Melero, GSM,GPRS and EDGE Performance, John Wiley & Sons, 2002.

- [134] 3GPP TS 34.108 "Common Test Environments for User Equipment (UE); conformance testing"
- [135] 3GPP TS 25.331 "Radio Resource Control (RRC); Protocol Specification"
- [136] 3GPP TS 04.60 General Packet Radio Service (GPRS); "Mobile Station (MS) - Base Station System (BSS) interface; Radio Link Control/ Medium Access Control (RLC/MAC) protocol"
- [137] 3GPP TS 05.08 Technical Specification Group GSM/EDGE Radio Access Network; "Radio subsystem link control"
- [138] J. Pérez-Romero, O. Sallent, R. Agustí, "On The Capacity Degradation in W-CDMA Uplink/Downlink Due to Indoor Traffic", VTC in Fall 04 conference, Los Angeles, USA, 2004.
- [139] S. Lincke-Salecker, "The Benefits of Load Sharing when Dimensioning Networks", Proceedings of the 37th Annual Simulation Symposium (ANSS'04), April, 2004.
- [140] S. Lincke-Salecker, "Performance and Service Issues in Selecting Adaptive Placement as a Load Distribution Technique", IEEE 59th Vehicular Technology Conference, VTC 2004-Spring, Milan, 2004.
- [141] M. Siebert, M. Schinnenburg, M. Lott, "Enhanced Measurement Procedure for Vertical Handover in Heterogeneous Wireless Systems", 14th IEEE International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC), Beijing, 2003.
- [142] J. Pérez-Romero, O. Sallent, R. Agustí, "On Evaluating Beyond 3G Radio Access Networks: Architectures, Approaches and Tools", IEEE 61st Semiannual Vehicular Technology Conference VTC Spring, Stockholm, 2005.
- [143] G. Cybenko, "Dynamic Load Balancing for Distributed Memory Multiprocessors", IEEE Journal on Parallel and Distributed Computing, Vol. 7, Oct. 1989, pp. 279-301.
- [144] T. Chu, S. Rappaport, "Overlapping Coverage with Reuse partitioning in Cellular Communication systems", IEEE Trans. on Vehicular Technology, vol. 46, no. 1, pp.41-54, Feb 1997.
- [145] A. Tölli, P. Hakalin, "Adaptive load balancing between multiple cell layers", VTC 2002-Fall. 2002 IEEE 56th, Vol. 3, 24-28 Sept. 2002 pp.1691 – 1695.
- [146] A. Pillekeit, F. Derakhshan, E. Jugl, A. Mitschele-Thiel, "Force-based load balancing in co-located UMTS/GSM networks", VTC 2004-Fall. 2004 IEEE 60th Vol. 6, 26-29 Sept. 2004 pp. 4402 – 4406.
- [147] H. Holma, A. Toskala (Eds.) WCDMA for UMTS: Radio Access for Third Generation Mobile Communications, John Wiley & Sons, 2002.
- [148] R.Jain, D-M. Chiu and W. Hawe. "A Quantitative Measure of Fairness and Discrimination For Resource Allocation in Shared Computer Systems," Technical Report TR-301, DEC Research Report, September, 1984.
- [149] J. Pérez-Romero, O. Sallent, R. Agustí, "Impact of Indoor Traffic over W-CDMA Capacity", IEEE Personal Indoor and Radio Communications Conference (PIMRC), Barcelona, Spain, September, 2004.
- [150] E. Gustafsson and A. Jonsson, "Always best connected," IEEE Wireless Communications, vol. 10, no. 1, pp. 49-55, Feb. 2003.
- [151] R.E. Schuh, P. Eneroth, and P. Karlsson, "Multi-Standard Mobile Terminals," Proceedings of the IST Mobile & Wireless Telecommunications Summit '02, pp. 174-178, June 2002.

- [152] P. Karlsson (editor) et al. "Target Scenarios specification: vision at project stage 2" Deliverable D13 of the EVEREST IST-2002-001858 project, February, 2005. Available: <http://www.everest-ist.upc.es>
- [153] 3GPP TR 21.910 3.0.0, "Multi-mode UE issues; categories, principles and procedures".
- [154] 3GPP TR 25.306 5.9.0, "UE Radio Access capabilities definition".
- [155] T.W. Wong and V.K. Prabhu, "Multi-Mode / Multi-Carrier Resource Management in CDMA / AMPS Deployment", IEEE 49th Vehicular Technology Conference, 1999, Vol. 3, pp. 1861-1865.
- [156] S. J. Lincke "Vertical handover policies for common radio resource management", International Journal of Communication Systems 2005; Published Online: 15 Mar 2005. DOI: 10.1002/dac.715.

8 ABBREVIATIONS

3GPP	3 rd Generation partnership project
AAC	Advanced Audio Coding
AC	Access Category
AC	Admission Control
ACK	Acknowledgement
AEDCF	Adaptive Enhanced Distributed Coordination Function
AF	Application Function
AIFS	Arbitration Inter-frame Space
AM	Acknowledge Mode
AMCS	Adaptive Modulation/Coding Scheme
AMR-WB	Adaptive Multi-Rate WideBand
AP	Access Point
ARQ	Automatic Repeat Request
AS	Active Set
ASF	Advanced Streaming Format
AVC	Advanced Video Coding
B3G	Beyond 3 rd Generation
BB	Bandwidth Broker
BCCH	Broadcast Control CHannel
BCH	Broadcast CHannel
BER	Bit Error Rate
BHCA	Busy Hour Call Arrival
BLER	BLock Error Rate
BS	Base Station
BSS	Basic Service Set
BSS	Base Station Subsystem
BSSGP	Base Station System GPRS Protocol
BTS	Base Transceiver Station
CAS	Code Allocation Scheme
CBR	Constant Bit Rate
CCPCH	Common Control Physical CHannel
CCCH	Common Control Channel
CDF	Cumulative Distribution Function
CF	Contention Free
CIF	Common Intermediate Format
CIR	Committed Information Rate
CM	Connection Management
CN	Core Network
CP	Contention Period
CPBAC	Combined PLEBAC/PABAC Based Admission Control
CPCH	Common Packet CHannel
CPICH	Common Pilot CHannel
CQI	Channel Quality Information
CRMS	Common Resource Management Server
CRRM	Common Radio Resource Management
CS	Circuit Switched
CSMA/CA	Carrier Sense Multiple Access/Collision Avoidance
CTCH	Common Traffic CHannel
CTS	Clear to Send
CW	Contention Window
DC	Deficit Counter
DCA	Dynamic Channel Allocation

DCF	Distributed Coordination Function
DCCH	Dedicated Control Channel
DCH	Dedicated Channel
DDRR	Distributed Deficit Round Robin
DFS	Distributed Fair Scheduling
DFT	Defer First Transmission
DIFS	DCF Inter-frame Space
DL	Downlink
DSCH	Downlink Shared Channel
DS-CDMA	Direct Spread Code Division Multiple Access
DSCP	Diffserv Code Point
DTCH	Dedicated Traffic Channel
DWFQ	Distributed Weighted Fair Queuing
EDCA	Enhanced Distributed Channel Access
ETSI	European Telecommunications Standards Institute
FACH	Forward Access Channel
FDD	Frequency Division Duplex
FEC	Forward Error Control
FER	Frame Error Rate
FIFO	First-in-First-out
GANC	Generic Access Network Controller
GERAN	GPRS-EDGE Radio Access Network
GGSN	Gateway GPRS Support Node
GPRS	General Packet Radio System
GSM	Global System for Mobile Communications
GTP	GPRS Tunnelling Protocol
H-ARQ	Hybrid Automatic Repeat reQuest.
HC	Hybrid Coordinator
HCCA	HCF Controlled Channel Access
HCF	Hybrid Coordination Function
HCS	Hierarchical Cell Structure
HHO	Hard Handover
HLR	Home Local Register
HO	Handover
HSDPA	High Speed Downlink Packet Access
HS-DPCCH	High Speed Dedicated Physical Control CHannel
HSPDSCH	High Speed Physical Downlink Shared CHannel
HSSCCH	High Speed Shared Control CHannel
HTB	Hierarchical Token Bucket
HTTP	HyperText Transfer Protocol
IEEE	Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
IF	Interference Factor
IMS	IP Multimedia Subsystem
IP	Internet Protocol
ITU	International Telecommunications Union
L1	Layer 1
L2	Layer 2
L3	Layer 3
LCD	Long Constrained Delay
LLC	Logical Link Control
MAC	Media Access Control
MAI	Multiple Access Interference
MCS	Modulation/Coding Scheme
MF	Multiplication Factor

MM	Mobility Management
MPEG	Moving Pictures Expert Group
MRR	Modified Round Robin
MS	Mobile Station
MSC	Mobile Switching Center
MSDU	Max. Service Data Unit
MTU	Maximum Transmission Unit
NACK	Negative Acknowledgement
NCCR	Network Control Cell Reselection
NRT	Non-Real Time
OVSF	Orthogonal Variable Spreading Factor
PABAC	Power Averaged Based Admission Control
PCCPCH	Primary Common Control Physical CHannel
PCPICH	Primary Common Pilot CHannel
PCE	Power Control Error
PCF	Point Coordination Function
PCU	Packet Control Unit
PDCH	Packet Data CHannel
PDCP	Packet Data Convergence Protocol
PDF	Point Decision Function
PDF	Probability Density Function
PDP	Packet Data Protocol
PDSCH	Physical Downlink Shared Channel
PDU	Protocol Data Unit
PHY	Physical Layer
PIR	Peak Information Rate
PF	Persistence Factor
PF	Proportional Fairness
PL	Path Loss
PLEBAC	Path Loss Estimation Based Admission Control
PLMN	Public Land Mobile Network
PRACH	Physical Random Access CHannel
PS	Packet Switched
PSNR	Peak Signal to Noise Ratio
QAM	Quadrature Amplitude Modulation
QBSS	Qos Basic Service Set
QCBR	Quasi-Constant Bit Rate
QCIF	Quarter Common Intermediate Format
QoS	Quality of Service
QPSK	Quadrature Phase Shift Keying
QSTA	QoS Station
QSIF	Quarter Source Intermediate Format
RAB	Radio Access Bearer
RACH	Random Access CHannel
RAN	Radio Access Network
RAT	Radio Access Technology
RB	Radio Bearer
RED	Random Early Detect
RF	Radio Frequency
RLC	Radio Link Control
RLP	Radio Link Protocol
RNC	Radio Network Controller
RR	Round Robin
RRC	Radio Resource Control
RRM	Radio Resource Management

RSCP	Received Signal Code Power
RSSI	Received Signal Strength Indicator
RT	Real Time
RTCP	Real Time Control Protocol
RTP	Real Time Protocol
RTS	Request to Send
RTSP	Real Time Streaming Protocol
RTT	Round Trip Time
SAW-ARQ	Stop and Wait Automatic Repeat reQuest.
SCFQ	Self-Clock Fair Queuing
SDP	Scene Description Protocol
SDS	Service Differentiation Scheme
SF	Spreading Factor
SGSN	Serving GPRS Support Node
SGW	Security Gateway
SHO	Soft Handover
SIF	Source Intermediate Format
SIFS	Short Inter-frame Space
SINR	Signal to Interference and Noise Ratio
SIR	Signal to Interference Ratio
SLA	Service Level Agreement
SMIL	Streaming Multimedia Integration Language
SNDCCP	Sub-network Dependent Convergence Protocol
S-PCF	Slotted PCF
SQCIF	Sub-Quarter Common Intermediate Format
SRB	Signalling Radio Bearer
STA	Station
TB	Transport Block
TBF	Temporary Block Flow
TBTT	Target Beacon Transmission Time
TCP	Transport Control Protocol
TDD	Time Division Duplex
TDMA	Time Division Multiple Access
TESPC	Traffic Specification
TF	Transport Format
TFC	Transport Format Combination
TFCS	Transport Format Combination Set
TFI	Temporary Flow Identifier
TFS	Transport Format Set
TM	Transparent Mode
TTI	Transmission Time Interval
TXOP	Transmission Opportunity
UDP	User Datagram Protocol
UE	User Equipment
UL	Uplink
UM	Unacknowledged Mode
UMTS	Universal Mobile Telecommunication System
USF	User State Flag
UTRA	UMTS Terrestrial Radio Access
UTRAN	Universal Terrestrial Radio Access Network
WCDMA	Wideband Code Division Multiple Access
WLAN	Wireless Local Area Network
WQB	Wireless QoS Broker
WWW	World Wide Web

