ADVANCES IN WIRELESS COMMUNICATIONS

edited by

Jack M. Holtzman
Michele Zorzi

Kluwer Academic Publishers

ADVANCES IN WIRELESS COMMUNICATIONS

THE KLUWER INTERNATIONAL SERIES IN ENGINEERING AND COMPUTER SCIENCE

ADVANCES IN WIRELESS COMMUNICATIONS

edited by

Jack M. Holtzman
Rutgers University

Michele Zorzi
University of California at San Diego

KLUWER ACADEMIC PUBLISHERS
NEW YORK, BOSTON, DORDRECHT, LONDON, MOSCOW

eBook ISBN 0-306-47041-1
Print ISBN 0-792-38126-2

Created in the United States of America

Visit Kluwer Online at: http://www.kluweronline.com
and Kluwer's eBookstore at: http://www.ebooks.kluweronline.com

CONTENTS

This page intentionally left blank.

PREFACE

In the last few years, Rutgers University's WINLAB has hosted periodic workshops on third generation wireless information networks. The first Workshop, held in 1990, was a pioneering effort in trying to define the needs and technical challenges stimulated by this idea of integrated wireless networking. Since that first one, the WINLAB Workshops have been a forum for discussion of the issues arising in this research area, reaching out to diverse fields and addressing scientific as well as business, regulatory and sociological factors, which are also very important aspects to the development of the "wireless age".

Another characterizing factor of these meetings has been the effort to mix industry and university, in a continuous effort to define research areas that can be of long-standing impact as well as of immediate interest for commercial application. The presence of authors from both companies and universities, and papers jointly authored by industrial and academic partners, testify to the success of this viewpoint.

The sixth Workshop was held in New Brunswick, NJ, on March 20-21, 1997. Although, by now, the number of people talking about third generation networks has dramatically increased and other conferences have sessions on this topic, experts from all over the world still recognize the importance of this event as a fruitful exchange of ideas, both during the presentations and during breaks and social events, in which state-of-the-art technologies are discussed and new solutions are sought. The papers presented at the Workshop, which are collected in this book, cover a broad range of topics. Papers describing state-of-the-art solutions are collected along with basic theoretical studies in information and communications theory, to offer what we think is a far-reaching panorama of this exciting field.

The 25 contributions have been grouped in 6 areas, which form as many chapters in this book: Characterizing and Serving Mobile Users, ATM and Broadband, Power and Energy Management, Capacity and Performance, Propogation and Traffic Modeling, and Locating Mobile Users. It is seen that the topics cut across all the protocol layers. In fact, as challenging as the more standard communication theory related problems are, it is the multifaceted and multilayer system problem of wireless and mobile communications that offer the most significant opportunities for breakthroughs.

The first topic starts with a paper by James Katz and Philip Aspden of Bellcore analyzing a survey of the demographics, attitudes, and social situation of wireless communication users. It is a step towards giving an empirical anchoring to speculations about wireless communications. The next paper by Wierlemann of DeTeMobil describes results from the *On the Move* Project. The objective is to provide personalized mobile-aware multimedia services and applications available regardless of mobile hardware, user location, and wireless network characteristics. The next paper by Tanabe of NTT Data Corporation is concerned with adaptively deciding between on-demand and broadcast modes, by characterizing data as popular or unpopular.

The next topic, ATM and Broadband, is concerned with going two steps beyond the ubiquitous voice services of wireless communications. Integrating relatively low rate data is the first step. Going up to broadband and using ATM are essential ingredients of further development. A variety of issues need reexamination for ATM and broadband. services and their requirements need to be defined with a platform to support them (Leskinen and Niemi of Nokia Mobile Phones). Other issues are access control (Narasimham, et al. of NEC USA, Inc., Ebert, et al. of The Technical University of Berlin, Artale and Winkler of Fondazione Ugo Bordoni), location management (Dommety of Ohio State University and Veeraraghavan of AT&T Labs), phone number translation delay (Jain of Bellcore), and handoffs (Campbell, et al. of Columbia). The paper by Fong et al. of AT&T Labs is on bandwidth allocation in fixed broadband networks, which are of great international commercial importance.

Over the last five years much work has been done on power control. The centerpiece of this theory is a beautiful result stating the solution to a formulation of the problem in terms of an eigenvalue and eigenvector of a particular matrix. This result, for centralized control, has stimulated much further work on extending the theory and developing distributed algorithms. Two further contributions are by Mitra and Morrison of Lucent Technologies and Rosberg of IBM Israel. Motivations for optimizing power control are to guarantee the feasibility of accommodating the maximum number of users, and to save energy and extend battery life. The papers by Röhl, et al., of The Technical University of Berlin and Zorzi and Rao of UCSD directly address the energy minimization issues.

The Capacity and Performance Sections address the seemingly unending new issues arising in wireless communications. CDMA for high speed data are in Kumar and Nanda of Lucent Technologies. Information theoretic considerations are in Foschini and Gans of Lucent Technologies. Reconfigurable networks, in which there are no base stations but rather all nodes are the same, are studied in Haas of Cornell University. The interaction of source and channel coding, and coding and multiple access with a multicode, variable bit rate embedded source coding is studied in Babich and Vatta of the University of Trieste - Italy. Trading off of C/I performance of users according to their service class and

location is considered in Khan and Zeghlache of the Institut National des Télécommunications - France.

Capacity and performance may significantly depend on the models used for propagation artifacts and traffic characteristics. It is therefore very important to be able to develop more realistic models that make performance studies meaningful for real-life scenarios. The issue of estimating the radio coverage in a cellular systems is addressed by Bernardin et al. of Nortel. An accurate model for the shadow fading taking into account spatial dependence of signal attenuation is proposed in Kumaran and Borst of Bell Laboratories. Characterization of new traffic sources, based on data communications patterns of Internet sessions based on the TCP/IP protocol is described in Anderlin and Zander of the Royal Institute of Technology, Sweden.

Tracking and paging users are the subject of the Locating Mobile User Section. Optimal paging using search theory is in Verkama of Nokia. Location tracking usually refers to location within a cell or location area. More precise position location (needed, e.g., for 911) when line-of-sight paths may be blocked is discussed in Wylie and Holtzman of WINLAB.

The contributions here collected have stimulated good discussion among the Workshop participants, both during the technical sessions and informally during the breaks. We believe it is worth offering this material to a wider audience, who can find in this book some stimulating ideas and presentation of state-of-the-art technologies relevant to wireless communications. We hope that with this effort we can give a small yet significant contribution to the understanding of this exciting and fast-growing field, and that the material here collected will be useful to students and researchers in their own search for new and better solutions towards the realization of the wireless information age.

Jack Holtzman
WINLAB
Rutgers University

Michele Zorzi
Center for Wireless Communications
University of California at San Diego

Acknowledgment

We are very grateful to Noreen DeCarlo for invaluable help with both the Workshop and in preparation of this volume. We also thank other members of the Program Committee, which included John Villasenor of UCLA and members of WINLAB.

This page intentionally left blank.

MOBILE COMMUNICATIONS: THEORIES, DATA, AND POTENTIAL IMPACTS A LONGITUDINAL ANALYSIS OF U.S. NATIONAL SURVEYS

James E. Katz
Bellcore, Morristown, NJ

Philip Aspden
Center for Research on the Information Society, Pennington, NJ

ABSTRACT
This paper examines the extent to which ownership of mobile communications is related to demographic variables and/or functionality needs. The study draws on data from seven national mail or telephone random surveys carried out during the period 1993-5 and totaling over 10,000 respondents. We found that the key determinants of mobile communications ownership were household income, race/ethnic background, need to be in touch and social/work mobility. Further, we found that those owning a cell phone plus a pager, those owning only a pager, and those owning only a cell phone had quite different demographic characteristics. While no longer a "rich man's toy", ownership of the cellular telephone is, nevertheless, still associated with the more affluent. Two important variables, gender and feelings of overload, did not appear to have any explanatory power. We speculate that ownership of mobile communications is determined more strongly by location effects, for example, need to be in touch or being highly mobile, than by gender. Regarding feelings of overload, we found no evidence that ownership of mobile communications generates feelings of overload, controlling for other variables.

INTRODUCTION
By 1999, it is expected that there will be more than 120 million cellular phone customers worldwide. (Common Carrier Week 1994). In mid-1996, there were already more than thirty-eight million subscribers in the US alone, or about 14.5% of the entire U.S. population. (Cellular Telecommunications Industry Association, 1996) (This contrasts with about 1% in the mid-1980s. Mayer, 1994.) A cheaper but more limited personal wireless system – the pager – had in 1996 about 8 percent penetration

overall but among older teens it was 17% (USA Today, August 27, 1996: D-1). Yet this adoption rate is small compared to Singapore, where one out of three adults uses a pager. (There are about 90 million pagers in the Asia-Pacific region.) (Associated Press 9-6-96). Paging technology is becoming two-way, and predictions are for a five-fold increase in world-wide subscribership by the year 2000. (Szaniawski, 1995)

What has been the impact wrought by this technology in people's personal and business lives? The mass media have presented sundry items ranging from a car-jacked man being rescued from his car trunk to British royalty being eavesdropped. But compared to media attention, the intellectual community has hardly probed the uses and implications of mobile communication.

It appears to us that, from a social analytical viewpoint, wireless personal communication has been overshadowed first by the proliferation of personal computers then by the Internet. Like its intensively scrutinized, socially transformative cousin, the personal computer, wireless personal communication has experienced a revolution since 1983. And like its socially transformative forebear – the telephone – personal mobile wireless technology has been largely ignored by scholars who claim to study communication modalities and social processes.

Our purpose is to examine wireless personal communication in terms of the demographics, attitudes and social situation of its users. We do this with an eye towards giving an empirical anchoring to speculations about this technology's impact. The consequences and power of this technology have contemporary ramifications for how individuals and groups communicate to pursue their needs. And, in light of the emerging wireless potential of the Internet, these technologies and their extensions are likely to have important consequences in a variety of areas from personal happiness to social equity, and from economic success to personal safety. (On a technical note, we use pager and beeper interchangeably, and exclude for our purposes dispatch radio. We are also aware of the interaction between technology and social change, i.e., that the causal arrow is not one way. But the bi-directional relationship, or mutual structuration of technology and society, will be dealt with in a separate article.)

RESEARCH PERSPECTIVE

As discussed above, there are numerous questions about mobile communications technology, regarding both how they affect society and human behavior as well as what use reveals about social theory and policy. While by no means definitive, it would for example be helpful to have data on how ownership patterns change over time, particularly in terms of user's educational attainment, gender, income, and age. It would also be helpful to know how this technology affects people's lives and relationships. Does it improve them, or make them more difficult? Are there gender differences in this regard?

Given concern about the commercial uses of mobile communications, there is much riding on whether they are actually a useful tool for economic success. Assuming for the moment that they are, questions of equitable distribution by social class or race assume a great deal of importance. This would especially be the case if their possession (or lack thereof) exacerbates equity problems, such as access to jobs or

information necessary to an informed citizenry. (This is related to the so-called digital divide.)

There is also the interpersonal dimension. Here it remains an open question as to whether ownership helps bind families, allowing "parallel mothering" (a term used in Rakow & Navarro, 1993), or drives them further apart by yielding less "quality time" or "face time."

Another important question is whether wireless communication technologies are fundamentally liberating or enslaving. While the answer is likely to be some combination of both phenomena, it would be helpful to have data which could actually illuminate the question. For instance, while by no means alone, Giddens (1990; 1991; Bryant & Jary, 1991) has spoken of the subtle controls over individual movements that technology might give, and Gary Marx (1985a; 1985b), Gandy (1994), and Katz (1988; 1990) have spoken of the ways in which these technologies can be abused to remove anonymity and freedom,

In light of our work and that of other scholars, we have reviewed some of the numerous research questions that are pending concerning wireless communication's social impact. In an attempt to address them, we have conducted national opinion surveys over a three-year span. The analysis that follows probes these relationships, and presents data that suggest preliminary answers to some of the above questions.

METHOD

Our main source of data is a 2,500 person telephone survey (identified as **Survey 95**). These data were taken from an October 1995 national random telephone sample, surveyed by a commercial firm under contract from Bellcore. The survey sample has a close match on socio-economic variables compared with the U.S. population as a whole. Based on comparisons with 1990/91 U.S. Census data, respondents in our sample are similar to the national average in ethnic mix, age composition and household income, but slightly more female and better educated.

Where appropriate we have also sought to bring out longitudinal trends by drawing on the results of several recent surveys we have carried out:

- a mail survey in early 1993 (identified as **Survey early 93**). A total of 1,870 questionnaires were completed and returned, resulting in a response rate of 35%.
- two mail surveys carried out simultaneously in late 1993 (**Survey late 93.a** and **Survey late 93.b**). A total of 1038 questionnaires were completed and returned for Survey a (response rate equals 24%) and 912 questionnaires were completed and returned for Survey b (response rate equals 21%).
- three mail surveys carried out simultaneously in 1994 (**Survey 94.a, Survey 94.b** and **Survey 94.c**). The response rates for these surveys were 36% for Survey a (1,380 completed questionnaires), 34% for Survey b (1,345 completed questionnaires), and 36% for Survey c (1,228 completed questionnaires).

Even careful surveys of users can be biased if they exclude those who at one time adopted wireless communication, but then discarded it because of unsatisfactory

results. Further, early adopters must be distinguished from later adopters, and results cannot be reliably extrapolated to non-adopters since their needs and effects might be quite different from those of adopters. Finally, the data are from public opinion surveys and have all the limitations that affect studies based on this technique. Perhaps most problematical in this regard are the possibilities that our indicator variables were not valid in the first place, or they were not reliably answered by respondents. These caveats apply to all our comments about findings, but shall not be repeated at each juncture their specter is raised.

RESULTS

Cell-Only And Pager+Cell Phone Usage Growing Faster Than Pager-Only Usage

Our approach has been to divide the sample population into four groups - those who report owning or using:

1. Neither a pager nor a cellular telephone (the “neither” group).
2. A pager only (the “pager-only” group).
3. A cellular phone only (“the cell-only” group).
4. Both a pager and a cellular phone (the “both” group).

In our 1995 survey, 63 percent of respondents reported not owning a pager or a cellular phone, 10 percent reported owning only a pager, and 16 percent reported owning a cellular phone, while 11 percent reported owning both a pager and a cellular phone (see Table 1). On the basis of historical data it would appear that over the past few years when there has been significant growth in both cellular and pager usage, pager-only growth has been slow if not static and certainly much slower than growth in cellular-only usage and combined cellular and pager usage. Across our seven surveys spanning more than two years, 6 to 10 percent of respondents report being pager-only owners, while cell-only owners grew from 9 percent to 16 percent of respondents, and owners of both grew even more rapidly, from 4 percent to 11 percent of respondents.

Table 1: Ownership rates by ownership groups

(in percent)	Neither	Pager-only	Cell-only	Both
Early 93	79	8	9	4
Late 93.a	78	7	9	6
Late 93.b	79	8	8	4
94.a	70	9	14	8
94.b	73	6	14	7
94.c	72	9	14	6
95	63	10	16	11

The ownership rates in the 1995 survey for cell phones and pagers are higher than the national subscriber rates quoted earlier. This difference can probably be ascribed to the fact that individual pagers and cell phones can have multiple owners, for example, one cell phone could be used by more than one member of a household. In addition, an unknown proportion of owners are non-subscribers because they no longer use their cell phones or pagers. Further, despite our best efforts we have a somewhat biased sample regarding mobile communications ownership rates, and to the fact that. The effect of this bias is reduced by seeking to explain ownership rates in terms of intra-sample demographic and attitudinal variables. We elected not to weight our sample due to its relatively small bias and the problems inherent in weighting (Katz, Aspden, & Reich, 1997). However, we recognize we must be circumspect about longitudinal (inter-sample) comparisons.

Cell-Only Usage - No Gender Difference; Pager Users More Likely To Be Male
In the 1995 survey 17 percent of male respondents and 16 percent of female respondents reported being cell phone-only users (see Table 2). Our earlier surveys suggest a growing convergence of male and female ownership rates for cell phones only.

The gender ownership pattern of cell phones only contrasts with ownership of pagers only and both cell phones and pagers, where proportionally more men than women reported owning them. In the 1995 survey 13 percent of male respondents and 8 percent of female respondents reported owning only a pager. Most of the earlier surveys also indicated that proportionally more men than women owned only pagers.

Similarly in the 1995 survey for the cell phone plus pager group, 13 percent of males and 9 percent of females reported owning both a pager and a cell phone. Again, the earlier surveys indicate that proportionally more men than women owned both pagers and cell phones.

Table 2: Ownership rates by gender

(in percent)	Pager-only		Cell-only		Both	
	Male	Female	Male	Female	Male	Female
Early 93	7	8	11	7	5	3
Late 93.a	7	7	10	9	9	3
Late 93.b	11	5	9	8	5	4
94.a	10	8	16	12	10	7
94.b	8	5	14	14	8	6
94.c	10	8	16	13	8	5
95	13	8	17	16	13	9

Our surveys show that the gender mix of mobile communications users has changed from 1989 when Rakow and Navarro (1993) reported that "more than 90 percent of subscribers were men." Indeed our surveys suggest that the gender gap for cell phone-only usage is on the verge of disappearing.

Declining Age Of Cell Phone Users; Pagers Mainly Owned By Young People

In the 1995 survey respondents who reported only owning a pager were predominantly under fifty years old (see Table 3); moreover, ownership rates were approximately the same for all the five-year age categories below 50. Over 50 years old ownership rates declined significantly down to 1 percent for the over 65 category. Earlier surveys had similar patterns.

Table 3: Pager-only: percent ownership rates per five-year category

	18-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65+
Ea 93	14	14	7	14	8	8	8	5	6	2
L 93.a	8	10	12	10	13	5	3	8	6	1
L 93.b	16	5	10	6	11	15	9	13	6	1
94.a	13	13	17	9	11	10	6	8	5	2
94.b	7	10	7	8	8	11	6	1	5	1
94.c	19	18	7	8	12	10	12	7	4	2
95	18	13	10	16	9	10	7	4	3	1

For those in the 1995 survey who reported owning only a cell phone (see Table 4), the age distribution of ownership was somewhat different from the 1995 pager-only age distribution. Ownership of cell phones was spread fairly evenly over the age range 18 to 64 with ownership rates varying from 15 to 22 percent per five year age category. This appears to be a change from the results of our earlier surveys where reported ownership rates tended to be highest in the age range 35-55.

Table 4: Cell phone-only: percent ownership rates per five-year category

	18-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65+
Ea 93	3	7	7	13	14	17	13	11	5	2
L 93.a	5	8	13	8	10	12	14	5	11	5
L 93.b	11	5	9	9	12	9	13	8	6	5
94.a	13	11	8	18	15	19	18	18	9	8
94.b	7	13	20	14	18	17	18	16	14	7
94.c	13	11	10	17	16	23	24	14	13	6
95	15	16	16	15	19	20	22	16	19	12

For the group reporting owning both a pager and a cell phone, the 1995 survey suggests a gradual decline in ownership rates from a 16 percent ownership rate for the youngest age category (18-24) to a 9 percent for the 55-59 category (see Table 5). For the categories 60 years old and more, the ownership rates are about 3 percent. The earlier surveys show a slightly different pattern; the 1994 and 1993 surveys indicate peak ownership rates over the age range 35-49.

Table 5: Pager + cell phone: percent ownership rates per five-year category

	18-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65+
Ea 93	5	4	6	8	5	5	2	3	1	0
L 93.a	5	8	7	8	8	9	11	3	3	0
L 93.b	0	5	7	9	7	8	3	5	0	1
94.a	11	13	8	11	9	11	8	4	4	3
94.b	13	8	9	10	11	11	3	11	4	0
94 .c	4	6	10	11	9	7	7	4	1	0
95	16	15	14	13	9	10	9	9	3	4

The results of our surveys provide further proof that cell phones are no longer the preserve of "power elites". Our earlier surveys do indicate highest ownership rates across the age range 35-50; the most recent survey shows that the highest ownership rates are spread over the age range 18-64, a much broader age range.

For the pager-only group, our surveys show that over the period 1993-95 ownership has continued to be by predominantly younger people, that is people aged less than 45 or 50. There is some suggestion in the results of the 1995 survey that the highest ownership rates are at the younger end of the 18-45 age range.

The Affluent/Better Educated Respondents More Likely To Own Cell Phones

For the cell phone-only group in the 1995 survey, ownership rates increase as household income increases, from 6 percent for the under $15,000 category to 38 percent for the $100,000 or more category (see Table 6). Similarly for the group which owns both a pager and a cell phone, ownership rates increase as household income increases, from 4 percent for the under $15,000 category to 26 percent for the $100,000 or more category. The earlier surveys showed a similar pattern of increasing ownership rates as reported household income increases.

For the pager-only group, however, the results for the 1995 survey are somewhat different. Here we see ownership rates independent of household income at around 12 percent of respondents. This is a change from earlier surveys which show a slight increase in ownership rates as reported household income increases.

Table 6: Ownership rates by household income (1995 survey)

(in percent)	Under $15 K	$15-24 K	$25-34K	$35-49 K	$50-74 K	$75-99 K	$100 K plus
Pager only	10	12	11	12	11	14	12
Cell phone only	6	10	14	18	24	31	38
Both	4	6	7	10	19	23	26

We also examined how ownership rates varied with the respondent's highest achieved education level. In the 1995 survey (see Table 7), ownership rates for the cell phone-only group increase with higher educational levels, from 10 percent for the group who left school without gaining a high school diploma (or GED) to 28 percent for the group who gained a Ph D. The earlier surveys showed a similar pattern.

In the 1995 survey, for the pager-only and the pager plus cell phone groups the relationship between ownership rates and highest education level achieved was less clear and it could be hypothesized that these data indicate no relationship between ownership and highest achieved educational levels. The earlier surveys do suggest a weak trend for both groups toward higher ownership rates for those reporting higher educational levels.

Table 7: Ownership rates by educational achievement (1995 survey)

(in percent)	Less than HS dip	HS grad	Some college	Tech sch	College grad	Some grad sch	Masters grad	PhD
Pager only	16	9	12	14	6	8	9	15
Cell ph only	10	13	17	17	24	20	21	28
Both	10	7	12	5	15	11	14	20

Our survey results suggest that the pager on its own has become a "classless" tool, since ownership rates appear to be independent of income and highest educational level achieved. On the other hand, higher ownership rates of cell phone, either alone or in conjunction with ownership of pagers continue to be associated with higher income and educational levels. It is possible there are independent "income" and "educational" effects, but since income and highest achieved educational levels are highly correlated, it is also possible we are seeing a purely income effect or a purely education effect. Later analysis will probe these issues.

Differences in ownership rates across ethnic groups

Analyzing ownership rates by reported ethnic group shows significant differences between ethnic groups. Blacks, Hispanics and Asians (with ownership rates in the range 44-47 percent) are more likely than whites (ownership rate 36 percent) to own mobile communications (see Table 8).

For the pager-only usage, blacks (ownership rate 19 percent), and Hispanics (ownership rate 17 percent) have much higher ownership rates than whites (ownership rate 9 percent) and Asians (ownership rate 5 percent) have. Asian and whites (ownership rates 23 percent and 17 percent, respectively) are more likely to own only cell phones than Hispanics and blacks (ownership rates 12 percent and 11 percent, respectively). Finally, Asians, blacks and Hispanics (ownership rates 19 percent, 17 percent and 16 percent, respectively) are more likely to own cell phones pagers than whites (ownership rate 9 percent).

Table 8: Ownership rates by race/ethnic group (1995 survey)

(in percent)	White	Black	Asian	Hispanic	Others
Neither	64	53	53	56	67
Pager-only	9	19	5	17	5
Cell phone-only	17	11	23	12	11
Both	9	17	19	16	17
No of respondents	100%= 2030	100%= 232	100%= 43	100%= 113	100%= 96

Those Mobile At Work/Socially More Likely To Own Mobile Communications

In the 1995 survey we asked some questions about the extent of mobility at work and socially, since *a priori* highly mobile people should have a disposition to own mobile communications. Respondents were asked the extent they agreed to the statement, "Your job requires you to be frequently away from your place of work." Non-owners of mobile telecommunications systems were less likely to agree to this statement than owners (see Table 9). For the group owning neither a pager nor a cell phone, only 5 percent strongly agreed and 10 percent agreed to the statement, whereas for the pager-only group, 13 percent strongly agreed and 21 percent agreed, for the cell phone-only group, 10 percent strongly agreed and 18 percent agreed, and for the group owning both a pager and a cell phone, 14 percent strongly agreed and 24 percent agreed.

Table 9: Job mobility (1995 survey)

"Job requires . . . frequently away from work-place" (in percent)	Strongly agree	Agree	Neutral	Dis-agree	Strongly disagree	No of resp
Neither	5	15	5	56	19	542
Pager-only	13	21	6	46	14	104
Cell phone-only	10	18	6	51	15	167
Both	14	24	7	38	18	123

Respondents were also asked in the 1995 survey the extent they agreed to the statement "In your social life you are frequently away from home." Again, non-owners of mobile telecommunications systems were less likely to agree to this statement than owners (see Table 10). For the group owning neither a pager nor a cell phone, 8 percent strongly agreed and 27 percent agreed to the statement, whereas for the pager-only group, 12 percent strongly agreed and 38 percent agreed, for the cell phone-only group, 12 percent strongly agreed and 34 percent agreed, and for the group owning both a pager and a cell phone, 14 percent strongly agreed and 36 percent agreed.

Table 10: Social mobility (1995 survey)

"In your social life you are frequently away from home" (in percent)	Strongly agree	Agree	Neutral	Disagree	Strongly disagree	No of resp
Neither	8	27	12	42	11	808
Pager-only	12	38	8	37	4	134
Cell phone-only	12	34	8	41	5	220
Both	14	36	10	35	5	143

We also used the number of children as a proxy measure for ***daily*** mobility and examined whether the number of children in the household related to reported ownership of mobile telecommunications. In particular, given Rakow and Navarro's work, we might have thought that the need to do parallel social and work activities would result in those individuals whose households contained children would have a greater need for communications, with the result they would be heavier wireless communication users. For the 1995 survey the results suggest there may be a weak

relationship between number of children and ownership of mobile communications. Households with no children were less likely to own mobile communications system than those with children. For households with no children, 33 percent reported owning either a pager or a cell phone or both. For households with one child, the ownership proportion was 42 percent, with two children, 44 percent, and for those with three or more children, 39 percent.

The results of our surveys regarding mobility support the idea proposed by Davis (1993) that the use of mobile communications provides "a sense of personal control over space and time." In regard to control over space, it would appear that those with greater mobility at work or in their social life are more likely to own mobile communications.

Those Needing To Be In Touch More Likely To Own Mobile Communications

In the 1995 survey we asked respondents the extent they agreed to the statement, "There are often times when you urgently need to get though to another person." Non-owners of mobile telecommunications systems were less likely to agree to this statement than owners (see Table 11). For the group owning neither a pager nor a cell phone, 10 percent strongly agreed and 36 percent agreed to the statement, whereas for the pager-only group, 13 percent strongly agreed and 47 percent agreed, for the cell phone-only group, 9 percent strongly agreed and 43 percent agreed, and for the group owning both a pager and a cell phone, 20 percent strongly agreed and 45 percent agreed.

Table 11: Ownership rates and the need to keep in touch (1995 survey)

"There are times when you urgently need to get though . . ." (in percent)	Strongly agree	Agree	Neutral	Disagree	Strongly disagree	No of resp
Neither	10	36	15	35	5	808
Pager-only	13	47	13	21	5	134
Cell phone-only	9	43	10	32	7	220
Both	20	45	10	21	4	143

In the only other survey where we asked this question, there was a very similar result. In the late 1993.b survey, for the group owning neither a pager nor a cell phone, 8 percent strongly agreed and 36 percent agreed to the statement, whereas for the pager-only group, 13 percent strongly agreed and 47 percent agreed, for the cell phone-only group, 13 percent strongly agreed and 38 percent agreed, and for the group owning both a pager and a cell phone, 8 percent strongly agreed and 53 percent agreed

Again in the context of controlling time, our results support the idea proposed by Davis (1993) that the use of mobile communications provides "a sense of personal control over space and time." Our surveys indicate that those with a greater need to keep in touch are more likely to own mobile communications.

Stressed Respondents Are More Likely To Own A Pager

We mentioned earlier the ongoing debate about whether mobile communications add to or ease the stress of modern living. To see if we could throw light on this debate we asked, in the 1995 survey, the extent respondents agreed to the statement, "You feel that you have more to do than you can comfortably handle." Those owning a pager reported more agreement (see Table 12) with this statement - in the pager-only group 17 percent agreed very strongly and 37 percent agreed strongly, and in the pager plus cell phone group 22 percent agreed very strongly and 26 percent agreed strongly. For the group without mobile communications and the cell phone-only group, the reported response rates were very similar - 17 percent agreed very strongly and 26-28 percent agreed strongly.

Table 12: Ownership rates and stress (1995 survey)

"You have more to do than you can comfortably handle" (in percent)	Strongly agree	Agree	Neutral	Disagree	Strongly disagree	No of resp
Neither	17	26	16	36	5	1575
Pager-only	17	37	13	28	5	263
Cell phone-only	17	28	12	36	6	411
Both	22	26	15	29	8	265

In three earlier surveys we also asked this question about the extent respondents have more than they can handle. Taking the four surveys together, the pager-only group reported most agreement with the statement. In each of the surveys, fifty percent or more respondents either agreed or strongly agreed with the statement.

Earlier we reported that the pager-only group was predominantly under fifty years old. Generally, we have found that younger people are more likely to report that they have more than they can handle, so we investigated whether the fact that the pager-only group was the most likely group to report having too much to handle was just an age effect. For the 1995 survey this proved not to be the case (see Table 13). We divided the sample set into those up to 44 years old and those 45 and over. Anxiety levels for the pager-only group did not decrease with age. For the younger half, 17 percent strongly agreed and 35 percent agreed with the statement, while for the older half, 19

percent strongly agreed and 43 percent agreed with the statement. For the other ownership groups anxiety levels decreased with age.

Table 13: Ownership rates by stress level and age (1995 survey)

"You feel that you have more to do than you can comfortably handle" (in percent)	Young-Strongly agree	Young - Agree	Old - Strongly agree	Old - Agree	No of respon dents
Neither	21	30	12	22	1575
Pager-only	17	35	19	43	263
Cell phone-only	18	32	15	24	411
Both	21	29	23	19	265

Similarly for the early 1993 survey, anxiety levels for the pager-only group did not decrease with age, while anxiety levels for the other three ownership groups did decline with age. For the younger half of the pager-only group, 15 percent strongly agreed and 38 percent agreed with the statement, while for the older half, 11 percent strongly agreed and 43 percent agreed with the statement. As in the 1995 survey, the anxiety levels decreased with age for the other ownership groups.

Although our surveys did not explore changes in feelings of overload before and after owning mobile communications, our surveys do indicate that the pager-only group is particularly likely to express feelings of overload. Whether this is a group inherently subject to stress or the ownership of pagers generates stress, we are not able to deduce.

Those Owning Mobile Communications More Likely To Own PCs

In the 1995 survey, those owning a cell phone were more likely to own a PC than those without a pager and a cell phone and those owning only a pager (see Table 14). Sixty-one percent of those only owning a cell phone and 59 percent of those owning both a pager and a cell phone also owned a PC. The percentage of PC ownership in the other two groups was much less - 40 percent for the group without mobile communications and 43 percent for the pager-only group.

In our earlier surveys we observed a somewhat similar pattern with the group without a pager and a cell phone having the lowest PC-ownership rates, about 30 percent. PC-ownership rates for the pager-only group were significantly higher than the neither group and were generally in the range 40-55 percent. In the 1993 and 1994 surveys, PC-ownership rates for the cell phone-only and both groups were significantly higher than the pager-only group. Ownership rates for the cell phone-only group were generally in the range 55-65 percent and for the both group in the range 60-80

percent. In some cases the ownership rates for the both group were significantly higher than for the cell phone-only group.

Table 14: PC ownership rates

(in percent)	Neither	Pager-only	Cell-only	Both
Early 93	28	37	57	58
Late 93.a	30	57	55	68
Late 93.b	30	53	59	72
94.a	33	56	65	64
94.b	30	50	63	84
94.c	32	48	70	79
95	40	43	61	59

There are various plausible explanations for the relationship between mobile communications ownership and PC ownership. (A similar relationship exists with the ownership of answering machines, see Katz, Aspden & Reich, 1997). Those owning mobile communications have lifestyles requiring the use of electronic tools such as PCs and answering machines. Alternatively, there are people with a pre-disposition to want to own new technological devices such as mobile phones, PCs and answering machines.

MODELING PAGER AND CELL PHONE OWNERSHIP

To some degree, reported ownership of mobile telecommunications equipment correlates with all the above demographic, mobility and attitudinal variables. To determine which variables were relatively more important and to examine their independent contribution to mobile communications ownership, we created three logit models using the 1995 survey data. We chose to use logit models (SAS, Software Release 6.07.02) since our data is predominantly categorical, and the dependent variable was dichotomous.

The logit models had the following dependent variables:

1. Pager ownership in the combined group without pager and cell phone and the pager-only group.
2. Cell phone ownership in the combined group without pager and cell phone and cell phone-only group.
3. Pager and cell phone ownership in the combined group without pager and cell phone and group owning both a pager and a cell phone.

The independent variables for these models are defined in Table 15.

The significance levels of the parameter estimates for the three models are given in Table 16.

Table 15: Definition of independent variables used in the logit models

Variables	Level = 1	Level = 2
Highest educational level completed	Less than HS dip, HS dip (or GED), some college or tech school	College degree, some grad work, master's degree, Ph D
Household income	up to $49,000	$50,000 and above
"Your job requires you to be frequently away from home"	Strongly disagree, disagree, neutral	Agree, strongly agree
"In social life, frequently away from home"	Strongly disagree, disagree, neutral	Agree, strongly agree
"There are often times when you need to get through"	Strongly disagree, disagree, neutral	Agree, strongly agree
"You feel that you have more to do than you can handle"	Strongly disagree, disagree, neutral	Agree, strongly agree
Use of PC at home	No	Yes
Age	Up to 49	50 and above
Gender	Female	Male
Number of children in the household	None	One or more
Race	White	Black/Hispanic

Table 16: Significance levels of independent variables

Variable	Neither v pager only	Neither v cell ph only	Neither v both
Gender	ns	ns	ns
Age	.05	ns	ns
Household income	.01	.0001	.0001
Educational level	.03	ns	ns
Ethnic background	.06	ns	.0001
Children in household	ns	ns	ns
Work mobility	ns	ns	.04
Social mobility	.05	.09	ns
Need to be in touch	ns	.08	.0005
Feelings of overload	ns	ns	ns
PC ownership	ns	ns	.06

Interpretation of the statistical results

By considering those variables with the highest significance levels in the three models we have developed the following model of mobile communications ownership, the variables are listed in approximately decreasing order of importance:

1. *Ability to pay* – household income is the most highly significant variable in our three models, particularly for ownership of cell phones either on their own or in combination with a pager. This could suggest that for many people ownership of a pager/cell phone is still perceived as a luxury item.

2. *Racial/ethnic background* – our results show that ethnic background (white as compared to blacks/Hispanics - we left Asians and "Others" out of this analysis). This possibly reflects family structure, cultural patterns, or symbolic aspects of services. (These questions will be explored later in the article.)

3. *The need to be in touch* – this is a statistically important variable for ownership of cell phones particularly in combination with a pager.

4. *Work mobility* – this is statistically important variable for pager and cell phone ownership, but not important for the other two models.

5. *Social mobility* – this is statistically important in the pager-only and cell phone-only models. Social mobility appears to be at least as important as mobility-at-work in predicting ownership of mobile communications. The number of children in the household, a variable which we thought might be a proxy for daily mobility or need to be in touch, did not appear to be statistically important in predicting mobile communications ownership for the 1995 survey.

Again, based on the size of the significance levels of the parameter estimates, there are three variables (educational level, age and PC ownership) with weaker explanatory powers than the above five variables:

1. *Educational level* – this appears to be only statistically important for the pager-only model. Earlier, we speculated on the separate existence of "income" and "educational effects". The results of our analysis suggest there are two separate effects, but that the "income" effect has more explanatory power than the "education" effect, overlapping and exhausting education's independent contribution.

2. *Age* – again, this appears to be only statistically important for the pager-only model, with younger respondents more likely to own pagers, perhaps suggesting that there could be a "fashion" element in owning pagers.

3. *PC ownership* – only statistically important for the explaining ownership of both pager and a cell phone. We speculate that someone who likes to own new technologies might be more likely to own both a pager and a cell phone.

Alternatively, those owning mobile communications have lifestyles requiring the use of electronic tools such as PCs and answering machines.

Two important variables, gender and feelings of overload, do not appear to have any statistical explanatory power when the other variables are kept constant. These are potentially important observations. Mobile communications have been held by many to be a male power tool, yet our models do not suggest any statistical gender effect. Our earlier analyses showed that ownership of mobile communications was related to gender. We speculate that ownership of mobile communications is determined more strongly by location effects, for example, having a highly mobile job or needing to keep in touch, than by gender.

Feelings of overload do not appear to have any statistical explanatory power when the other variables are kept constant. We speculate that the people who feel overloaded do not perceive that pagers and cell phones can help them reduce overload anxieties. By contrast our model supports the view that those who are highly mobile or need to keep in touch perceive pagers and cell phones as useful to them.

CONCLUSION

Our data suggest many interesting relationships upon which theoretical interpretations could be built. Space limitations preclude a full elaboration of these possibilities but, guided by our discussions at the outset of the paper, we can highlight a few intriguing possibilities.

Our investigation reveals, that while no longer "a rich man's toy" cellular telephone ownership is associated with income. And in contrast to some of the more pessimistic speculations about cell phone ownership, including our own (Aspden & Katz, 1994), there is as yet little evidence to suggest that cell phone ownership has a pernicious impact on the quality of life. However, as we indicated earlier, we cannot be sure of the antecedent situation in a cross-sectional study, so any claims about impact must be extremely circumscribed.

In general ownership seems guided by what we might call "social location" variables, that is a combination of socio-economic, demographic, and life style conditions which influence decisions perhaps more powerfully than individual personality characteristics. People may not personally wish to have wireless communication, but due to conditions of their life – such as job exigencies, work and personal mobility – they find they need to have wireless communication. Thus while we found no relationship between certain personality measures, such as extroversion or phone liking, perceived "needs" can be an important predictor of service utilization, in this case a perceived need "to keep in touch." This finding is somewhat at odds with an initial analysis of some of the early data reported here, which purported not to find a relationship between needs and telecommunications technology use. (Steinfield, Dudley, Kraut & Katz, 1992) Certainly the relationship between needs and gratifications is one that has been explored at length in the mass media literature; that they should be connected in the telecommunications area as well should not come as a surprise. Yet we anticipate that "social location" will eventually become recognized as equal in importance, if not paramount to, the "needs"-based model.

Interestingly, racial/ethnic self-identification is an important variable. The importance of this variable was somewhat surprising to us. While there are several possible explanations, for example cultural patterns and geographic location of respondents, the importance of this variable also fits with notions we have working with, which may be characterized as "affordable luxury." This concept means that certain (apparently) high status items are available for purchase at relatively low cost. This means that groups or individuals who might not have a high status, as defined by the dominant society, might seek to enhance their status in various ways. Certainly the more affordable a luxury is, the easier it would be for people to buy it. Explanations along this path might help us to understand not only wireless communication ownership, but a host of other behaviors as well. Currently we are exploring this area, and hope to report on it soon. (See Katz, Aspden & Fussell, 1997)

In terms of policy issues, such as information rich-information poor, or the so-called digital divide, some of our findings are troubling. Specifically the income dimension of wireless communication ownership suggests that there is a possibility that those who cannot afford may be shut out of many of society's benefits, with severe personal and political ramifications. (Aufderheide, 1987; Sawhney, 1994).

Summing up, we have seen some surprising data shedding light on wireless communication relative to income, age, education, ethnicity/race, household ties, social activity and job activity mobility and attitudes. These data reflect on important theories of equity, innovation, gender relations, and quality of life issues. While we have only scratched the surface, we believe we have shown that this much-neglected area of wireless communication can have both substantive and theoretical import.

The authors acknowledge and thank Charles Steinfield, Kate Dudley, and Robert Kraut for their role in collecting and analyzing the "early 1993" data set reported here.

REFERENCES

Aspden, Philip and James Katz. *Mobility and Communications: Analytical Trends and Conceptual Models.* Report for the U.S. Congress, Office of Technology Assessment, OTA N3-16040.0, November, 1994.

Associated Press. "Singapore Celebrates 1 Millionth Pager Customer" September 6, 1996 (0179)

Aufderheide, Patricia. 1987. Universal service: telephone policy in the public interest. *Journal of Communication.* 37 (Winter): 81-96

Cellular Telecommunications Industry Association, 1996. "Wireless growth sets new annual records." Mimeo. September 19. Washington.

Common Carrier Week, June 6, 1994.

Davis, Dinch M. 1993. "Social impact of cellular telephone usage in Hawaii." Pages 641-49 in *Pacific Telecommunications Council Fifteenth Annual Conference Proceedings,* Session 3.1.1. to 4.4.1 Edited by James G. Savage and Dan J. Wedemeyer, volume 2. January 17 - 20, 1993.

Gandy, Oscar H. *The panoptic sort*: A political economy of personal information. Boulder, CO, Westview, 1993.

Giddens, Anthony. 1990. *The consequences of modernity.* Cambridge: Polity Press in association with Basil Blackwell, Oxford, UK.

Giddens, Anthony. 1991. *Modernity and self-identity.* Palo Alto, CA, Stanford University Press.

Katz, James E. 1988. "Public Policy Origins of Privacy and the Emerging Issues," *Information Age,* 10 (3), 1988: 47-63.

Katz, James E. 1990. "Social Aspects of Telecommunications Security Policy," *IEEE Technology and Society,* 9 (2) (Summer): 16-24.

Katz, James E., Philip Aspden and Warren Reich. 1997. "Public Attitudes Toward Voice-Based Electronic Messaging Technologies in the United States," *Behavior & Information Technology*, Volume 16, Number 3, 125-144.
Katz, James, Philip Aspden, and Susan Fussell, 1996. "Affordable luxury and telephone servcies: Perceptions and behavior among a random sample of consumers." Mimeo in press.
Marx, Gary T. 1985a. The surveillance society: the threat of 1984-style techniques. *The Futurist.* v. 19 (June): 21-26.
Marx, Gary T. 1985b. I'll be watching you: reflections on the new surveillance. *Dissent* v. 32 (Winter): 26-34
Mayer, William G. "The rise of the new media." *Public Opinion Quarterly*, 58: 124-46.
Rakow, Lana F. and Vija Navarro. 1993. "Remote mothering and the parallel shift: Women meet the cellular telephone." *Critical Studies in Mass Communication*. 10 (2): 144-57.
Sawhey, Harmeet. 1994. Universal service: prosaic motives and great ideals. *Journal of Broadcasting & Electronic Media.* 38 (Fall): 375-95.
Steinfield, Dudley, Kraut & Katz, 1993. Rethinking Household Telecommunications. Paper presented at the International Combinations Association annual meeting.
Szaniawski, Kris. 1995. "Operators push low-cost advantage." *Financial Times*, November 27, 1995. p. 5.
USA Today, August 27, 1996: D-1,

This page intentionally left blank.

THE *On The Move* PROJECT: DESCRIPTION OF MOBILE MIDDLEWARE AND EXPERIMENTAL RESULTS

T. Wierlemann[1], T. Kassing[1] , Julie Harmer[2]
[1] DeTeMobil GmbH, Deutsche Telekom MobilNet GmbH,
Box 8865, D-48047 Münster, Germany
[2] BT Laboratories, Martlesham Heath, IPSWICH, England.

[1]The On The Move project is sponsored by the European Commission in the ACTS programme, Bonnier Information Systems AB (S), British Telecommunications plc (UK), Burda New Media GmbH (D), Deutsche Telekom MobilNet GmbH (D), Ericsson Eurolab Deutschland GmbH (D), Ericsson Radio AB (S), IBM (F), Iona Ltd. (UK), Royal Institute of Technology (S), RWTH Aachen (D), Siemens AG (D), Sony (D), Swedish Institute of Computer Science (S), Tecsi (F), and University of Singapore Centre for Wireless Communications (S).

Abstract

The ACTS [1] OnTheMove[1] project [2] is investigating mobile multimedia services for the Universal Mobile Telecommunication System (UMTS) [3]. A Mobile Application Support Environment (MASE) is being designed and implemented to answer the challenges of providing personalised mobile-aware multimedia services.

The MASE system will provide personalised access to multimedia applications. These applications will be available to users irrespective of the mobile hardware, user location, and wireless network characteristics.

A Mobile Application Programming Interface (Mobile-API) is being developed which will provide a standardised means of accessing the facilities of the MASE.

The functionality of the system will be demonstrated to users and service providers through the implementation of a multimedia business information system which forms the basis of a series of field trials.

The project began in September 1995 and lasts for three years. This paper introduces the background and rationale behind the OnTheMove project. An overview of MASE functionality is given and the need for a distributed MASE is explained. Experimental results from the first field trial with the business application "Infomotion" are presented. Finally, the potential impact of the project is assessed.

Introduction

Mobile Office the buzz-word of todays portable computing is a rapidly growing market sector. Recently the swedish company PC Card Distribution Scandinavian AB has launched a PC card with a entire GSM phone on it including a modem for sending and receiving of data and fax. Other mobility supporting devices like GPS receivers, which are the key for location awareness, are estimated to have a market potential of $10 milliard by 2000 following a market research done by "Forward Concepts". The value of the remote computing hardware market in 1995 was estimated at $4.7 billion and is predicted to increase to $12.3 billion by 1997 [4]. It is estimated that 48 million employees spend more than 20% of their working day away from their desks [5, 6]. Portable computers are becoming increasingly powerful, and Personal Digital Assistants (PDAs) now provide pocket-sized computing [7].

On fixed networks the widespread use and popularity of the Word Wide Web (WWW) has dramatically increased the volume, richness, and availability of information. With its explosive growth and innovations, the Internet is changing the nature of communications through the rapid adoption of Internet standards and providing universal connectivity. The estimated total 300-500 million world-wide Internet users by the year 2001, coupled with the continued growth in wireless network penetration rates, will create the potential for a mass market for mobile

communications. It has been suggested that there may be 40 million mobile connections in the European Union by the year 2020 [8].

Second generation GSM cellular mobile radio systems will evolve to third generation wireless networks, such as the Universal Mobile Telecommunications Service (UMTS). UMTS will offer greater access flexibility and bandwidth and provide a communications platform for a wide range of mobile multimedia services and applications for mobile devices. The IP-based traffic of the Internet and its diverse content will have a major influence on these next generation mobile communication networks. Much of this content makes use of video and audio in addition to text and graphics.

Today's wide area wireless networks cannot deliver the bandwidth necessary to adequately support full multimedia applications and the mobile environment imposes a number of challenges which impair the robustness of client/server operation: adaptation to varying quality of service, robustness in the face of disconnected links, roaming between different operators and network types, movement between different geographical locations, reconfigurable real-time multi-party connections, and personalised information filtering. Mobile terminals have varying capabilities, ranging from a GSM phone or personal communicator, PDAs with small screens to fully multimedia capable laptops. To maximise the benefit of mobile multimedia solutions, an architecture is necessary which can answer the challenges of mobility, and provide the user with robust, personalised applications irrespective of the terminal or wireless network available.

The OnTheMove project is developing mobile middleware - a Mobile Application Support Environment (MASE) and a Mobile-API which will enable a new generation of "mobile-aware" applications and also provide additional functionality to "legacy" applications. The prototypes are based on commercially available products and near term research results. OnTheMove's architecture will help to facilitate and promote the development of a wide spectrum of mobile multimedia applications.

Mobile Middleware

The MASE middleware [9] will hide the complexity of underlying networks and provide a complete support environment for mobile applications. A diverse range of mobile networks will appear as a seamless, homogeneous communications medium. Differences in these networks appear to applications - and therefore the user - as changes in Quality of Service (QoS). User profiles allow the user to influence the behaviour of the MASE by specifying their communication preferences which include cost, quality, time, and security. Terminal profiles provide a means of determining the characteristics of the mobile device which has an important impact on the way that information is presented to the application. For example, video content should not be sent to a terminal device which cannot support video, even if there is sufficient network bandwidth available. MASE functionality includes: network adaptation, disconnected operation, multimedia conversion, Quality of

Service adaptation, location awareness, and user profile management. The MASE and its relation to wireless networks and applications is illustrated in Figure 1.

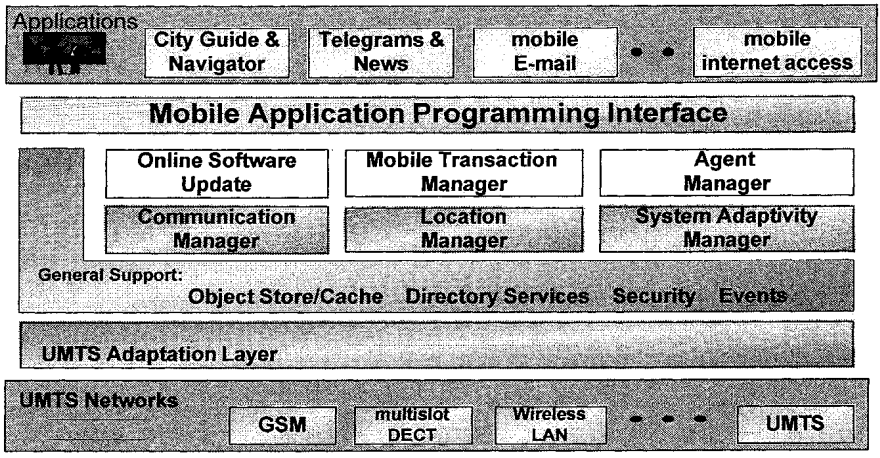

Figure 1 Mobile Application Support Environment

Mobile-API

Mobile-API specifications allow “mobile-aware” multimedia applications to communicate QoS requirements to the computing and communication subsystem below them and receive information about currently available services. The API will be a superset of a recognised API, or a number of APIs, in order to provide backwards compatibility for legacy applications. The Mobile-API will be a key factor in enabling the exploitation of MASE functionality and will be input to standards bodies.

The Distributed MASE

It is intended that the MASE will support the widest possible range of mobile terminals, from cellular phones, to PDAs, to fully functional multimedia laptops. As a result of this diversity of equipment, in many cases it will not be possible for the MASE to get all the necessary information and resources from the mobile terminal itself. For this reason, a distributed approach will be taken using mobility gateways. These systems are attached to the wired network and are aware of the current status of mobile terminals and the underlying communications networks. Based on user and application profiles, the MASE is then able to gather information and resources from the mobile terminal and the mobility gateway in order to fulfill the needs of the application.

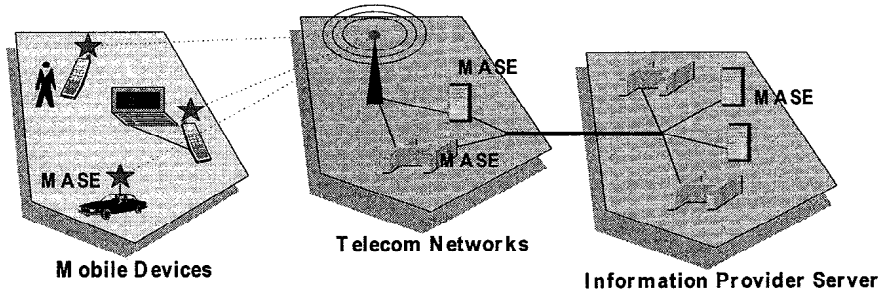

Figure 2 The distributed MASE

Examples of functionality to be provided by the mobility gateway include: multimedia conversion - adapting the information representation to match both the capabilities of the mobile terminal and the QoS offered by the communications network in use; disconnected operation - ensuring that when an unplanned disconnection occurs, the application can continue to operate without loss of data or unnecessary retransmission of data; transaction management - setting up distributed transactions on behalf of the mobile terminal. Some caching mechanisms may also be implemented in the mobility gateway to improve performance. Figure 2 illustrates the distribution of the MASE on mobile terminals and mobility gateways.

Experiments and Field Trials

The development process of the MASE is accompanied by a series of field trials [10]. The objective of these field trials is to collect feedback and user reaction at any stage of the development process and to address the user requirements by key features of the application and supporting layer software. The first of these field trials, “Experiment 0", has been already conducted in 1996 distributed over three European cities in Sweden (Stockholm) and Germany (Munster, Munich). Two further field trials are planned in at least 5 European cities and one test-site in Singapore during 1997 and 1998.

Experiment 0

The actual execution of Experiment 0 was divided into two parts complementing each other: a user part (application trial) focusing on user reaction and feedback about the acceptance of a mobile multimedia business system called Infomotion and a network part (technical trial) evaluating the performance of the underlying communications infrastructure with respect to roaming, routing, throughput, response time and jitter.

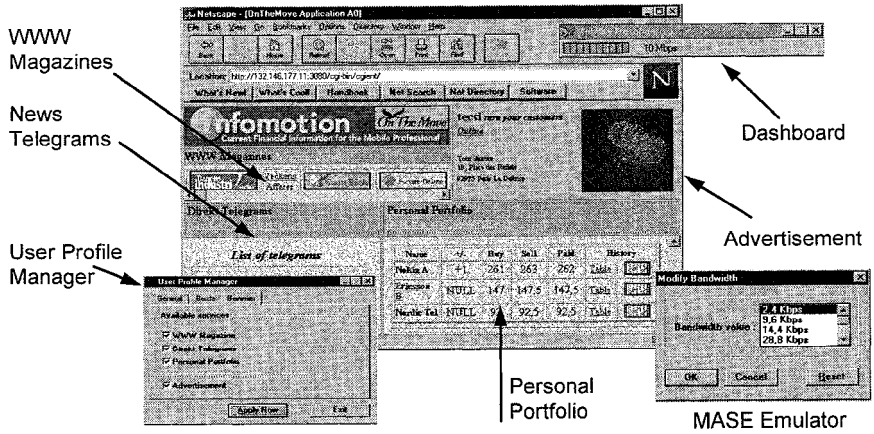

Figure 3 Multimedia Business Information System “Infomotion”

Experiment 0 - User Part

The user part of Experiment 0 has been conducted on the basis of commercially available infrastructure and equipment. The business application lnfomotion offered in experiment 0 was based on existing WWW technologies and provided user feedback on a reduced set of mobile environment functions such as location awareness and bandwidth management. The services of *Infomotion* consisted of a portfolio based on Swedish stock exchange data, news services with data filtering capabilities, a location aware travel service and personalised WWW-magazines. The services layout is sketched in Figure 3.

Early MASE functionality could be demonstrated to the end-users by emulation (MASE Emulator, Dashboard, User Profile Manager). The experimental testbed (see Fig. 4) comprised two WWW information servers to run the services of the Infomotion system. Both content servers were connected to the Internet via proxies which allowed anonymous logging. The log files were part of the material used for the evaluation of results. The networks have been carefully selected to encompass important UMTS bearer services and mobility pattern. GSM thus represented a circuit-switched bearer supporting full mobility, MODACOM[1] was a representative for a packet-switched bearer supporting full mobility and the local Wave-LAN[2]

[1] MODACOM is a packet-oriented cellular radio network based on the DataTAC technology defined by Motorola. The gross transmission speed amounts to 9600 bit/s per cell. Over the air interface, DataTAC uses the Radio Data Link Access Protocol (RD-LAP) with the Slotted Digital Sense Multiple Access (S-DSMA) protocol for media access.

[2] WaveLAN is a product from Lucent technologies (formerly NCR, formerly AT&T) which operates over the 2.4 GHz band. The radio unit consists of a PC-Card and an antenna. The power is obtained from the computer. Connection to the fixed network is realised through a bridge. WaveLAN has an operating capacity of 2 - 4 Mbps, depending on the number of users and the radio environment.

served as representative for an office environment allowing for high bit rate data communication.

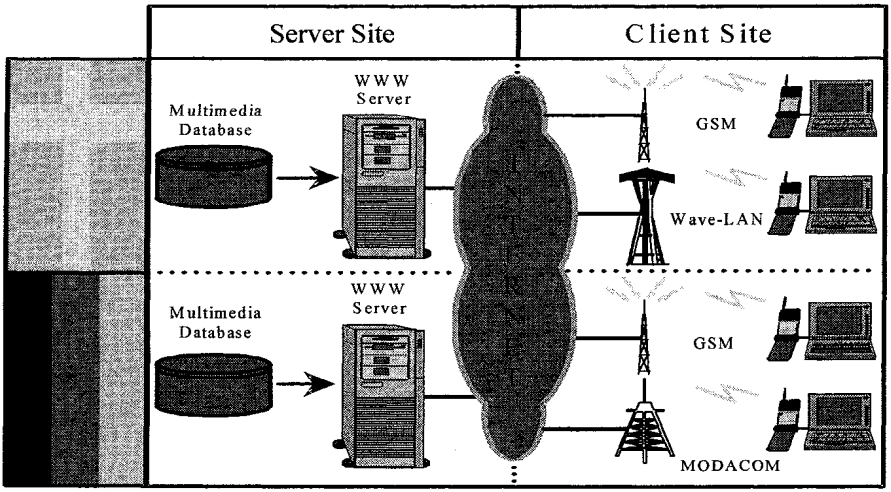

Figure 4 Experiment 0 Infrastructure (User Part)

The multimedia information servers were globally accessible from any wireline network and all radio network specific implementation details were securely hidden in the IBM ARTour Gateway[3] . Additionally, the GSM access to the Internet was realised by PPP dial-up connections via analogue modems. Experiment 0 involved 48 long-term users equipped with laptops, GSM handhelds, MODACOM terminals (only in Germany) and access to wireless LAN (only in Sweden).

Open questionnaires were performed with each test-user after a two weeks test period to receive direct and subjective answers on how to improve the services of *Infomotion.* The questionnaire itself was divided into 5 main parts: a network part to judge the performance of the underlying communications infrastructure; a user interface part for evaluation of personal profiling capabilities, user interface realisation and location management; a content part to collect feedback on the usefulness of the offered contents; a layout part to collect user feedback on the general presentation and layout (colours, buttons, text styles, etc.); and a future part to collect thoughts on what mobile communications should provide in the future, as well as personal expectations on the future development of the multimedia-industry. Altogether, about 60 questions had to be answered.

The analysis of the questionnaires gave indications about the attractiveness of the Infomotion system and it's performance as experienced by the test-users. In the

[3] ARTour is IBM's Open System communication platform for mobile computing. It gives access to computer and network resources from portable computers via radio networks with the standard protocols included in the TCP/IP Internet Protocol Suite. The focus of ARTour is on the communication between a mobile unit and a stationary unit. The ARTour Gateway acts as IP router between radio and wireline networks. Thus, transparent IP addressing can be applied between mobile and stationary units.

following, most meaningful results of the experiment are presented for the different parts of the questionnaire.

Network part: The main problems with respect to the underlying mobile communications infrastructure occurred during the connection set-up phase and were not caused by the network itself, but by the client and server communications software. In average, 30% of the times the test-users failed to establish a connection with the information server (see Fig. 5).

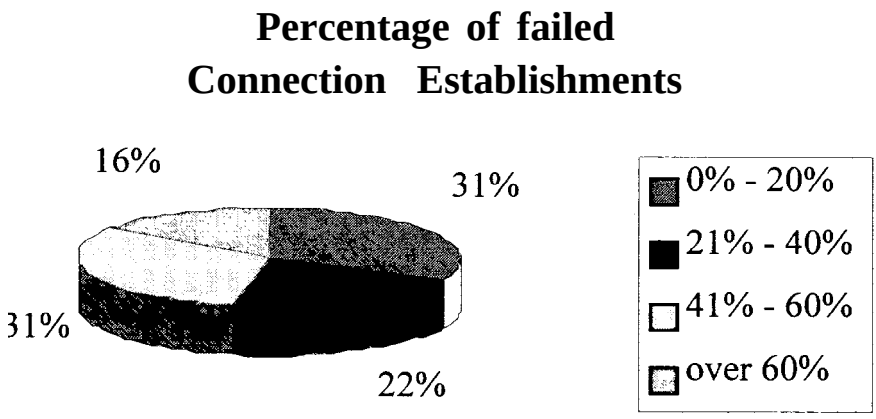

Figure 5 Percentage of failed connection establishments over the whole test period

Independently of the nature of the failure, the test-users became easily frustrated when their attempt had failed several times. Once a connection was found, it proved relatively stable. Anyhow, sudden disconnection has always a severe effect on a test-use! Those who experienced a lot of network problems in the first days of the test soon lost interest. The time factor is probably also important. Re-establishing a connection must be easy and swift in future applications. Transmission speed is important. It is difficult to even imagine under what circumstances low data rates would be acceptable to a professional user without additional service support from the underlying network. The information value must then be of extreme height. The network is a commodity and of as much interest to users as petrol chemistry is to drivers. Capacity or price is the option one has, little else is of interest for the user.

User interface part: For the great majority of people participated in the experiment, the user interface was practical and easy to handle, however, nearly all test-users got lost when trying to move around, either by trying to change network, reconfigure a setting or going backward to check e.g. a chosen keyword. Future applications must reduce the number of clicks/choices considerably. Furthermore, the design of the user interface (buttons, scroll bars, etc.) should be comparable to legacy applications like Windows. The majority of test-users participated in experiment 0 was familiar with interactivity and hyperlinking. They could not be impressed by what is perceived as an ordinary web site. The user requires a simple feedback mechanism showing the current status of the network: Which network am I currently using? What are the costs (per unit/minute)? Am I

connected/disconnected? How much bandwidth is currently available? Which service can I execute on top of this bandwidth?

The content part: Some users saw *Infomotion* as a well done Internet-Site, others considered it as a real service. Especially the location aware travel service, intelligent information filtering and personalisation are seen to make the step from a well designed web site to a new service (see Fig. 6).

As time/bandwidth and money are scarce sources, the quality of the content is of highest importance. Consequently, users should be enabled to profile the content according to their individual needs. Furthermore, the requested information should be close to the top layer and easy to find. Alerting and truly real-time feed could help to make information more easily accessible. Personalisation, interactivity, relevance and in-depth information are the main criteria for future content!

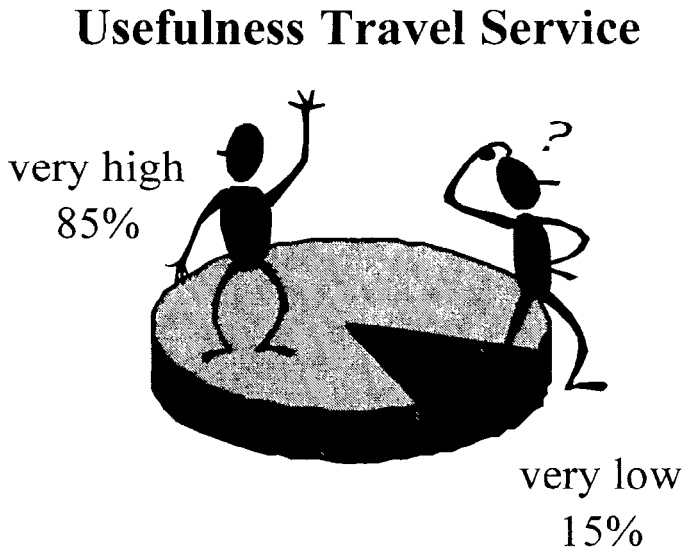

Figure 6 Usefulness of travel service

The layout part: This was probably the area where user trials of this kind can be of best use and give the clearest answers. Complex screen design with too many options given at the same time must be avoided. It is questionable whether a frame design should be used for services intended to be viewed on laptop screens. One common and recognisable layout for the whole service (simple, clear and self-explaining) was requested, however, the layout was valued as less important, content comes first, the layout only underlines it.

The future part: The network performance should be stable and reliable! This was the most desired future condition to support mobile applications, of course heavily affected by the fact that the test-users could not distinguish between the network access software and the network itself. End-user devices should be small,

lightweight and integrated, capable of voice and data communication (smart-phones), with improved battery standby times. For future developments, robustness and user-friendliness are perceived as golden goals.

User Part Results Summary: The majority of the users involved in the user part of experiment 0 fell into the category of ”knowledge workers” as defined by the Electronic Publishing report [11]. This category is not cost sensitive, but time sensitive. They also used the PC as an information terminal. Building on the conclusions of this report, it is evident that this user category is a probable user market for future broadband mobile services which the OnTheMove project is designing, and that services should be designed with them in mind.

Data transmission speed offered by current cellular networks is too slow, to overcome this bottleneck, enhanced data compression and media conversion technology is needed. Wireless connections in general are too unreliable, different modes of operation (on-line, off-line, disconnected), packet or asynchronous message based, are required to handle these problems. Alerting mechanisms are required to make information more easily accessible and mobile-aware applications that can adapt to different networks and make best use of available network resources.

To summarise, there is an emerging need for mobile multimedia applications and services. However, the feedback given upon the *Infomotion* system clearly showed that present available cellular networks are not very well suited to cover the demands of multimedia communications. Current cellular networks lack of mobility support environments! The MASE represents such an environment and can alleviate most of the problems specific for mobile multimedia communications (limited bandwidth, disconnection, etc.). The Mobile-API allows applications to make use of this environment and to become mobile-aware.

Experiment 0 - Network Part

Future cellular networks like UMTS have to be designed to support ubiquitous communications, regardless of media types (voice, image, video, etc.), making best use of available network resources by means of adaptation to bearer types. The UMTS Adaptation Layer (UAL) as a lower layer part of the developed MASE will encounter this. The UAL will select and configure the appropriate networks, bearer services and protocol stacks transparently to the user, according to the requested QoS of the involved applications.

The general functional requirements of the UAL are threefold: to select the appropriate transport- and network protocols (e.g. UDP for unreliable datagram service or XTP for multipoint connections; e.g. IPv6 in order offer network layer security features) and configure it for efficient use (e.g. adjusting packet size and timers to the bearer service parameters), to select and configure an appropriate bearer service if more than one is available, to measure and monitor bearer service and protocol stack activities to respond to changes in the QoS and network state.

In the network part of experiment 0 the behaviour of TCP/IP traffic over wireless links was evaluated to identify basic design constraints for the UAL in terms of throughput, delay and jitter.

The measurements involved mobile hosts with access to the GSM network connected to a fixed host on a high speed LAN (see Fig. 7). Preliminary tests with different applications showed that interactive network applications such as TELNET, do not cause performance problems for such kind of heterogeneous networks. On the other hand the bulk data transfer is the hardest performance test as the optimum utilisation of the available bandwidth is evident to guarantee a good QoS. Therefore, the FTP application was chosen for these measurements.

The experiments were conducted under three conditions: static mobiles with high quality of the radio signal, static mobile with poor quality of the radio signal, and moving mobiles to obtain results under representative real life situations. The data transfer was performed in both directions, up- and down-link. For the measurements files of 250 Kbytes were used to be able to observe possible retransmissions due to transmission errors, and to measure the influence of TCP slow start and congestion recovery mechanisms. For the experiment data transfer performance, and dial-up connection establishment characteristics were measured.

For this environment it was observed that the down-link transmission speed was almost 2 times slower than on the up-link for all three test conditions (see Fig. 8). For further measurements the cellular network was substituted by a serial-line cable. With this set-up it could be proven that the reason for this difference in up- and down-link transmission speed is the slow start mechanism of the TCP/IP protocol due to a buffer congestion, as within this experiment the throughput could be increased proportionally with changing window sizes.

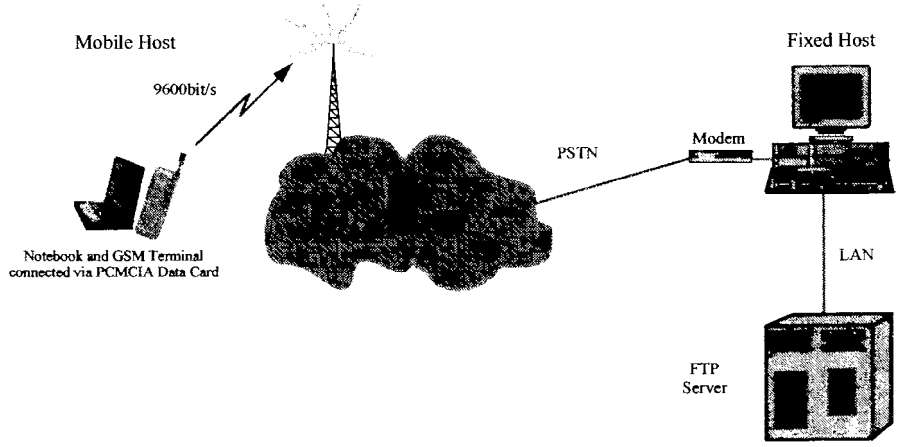

Figure 7 Experiment 0 Infrastructure (Network Part)

The maximum transfer rate (868 bytes/s) was achieved under static conditions with high quality of the radio signal on the up-link. Whereas the minimum transfer

rate (214 bytes/s) was measured on the down-link for static mobiles with poor quality of the radio signal.

On the up-link all data transfers could be completed successfully under all three test conditions. This high value of reliability decreased dramatically for moving mobiles to 17% successful transfers on the down-link. This shows that the buffer congestion problem on the down-link has a severe impact on the reliability of the data transmission in general.

Connection establishment time (time between call command and receiving "connect" message) for static conditions and good quality of the radio signal was at a maximum 35.1 seconds and the minimum time was 30 seconds. All attempts were successful for this condition. For the poor quality condition the connection establishment time was between 52.1 seconds and 30.5 seconds. Normally the successful connection establishment took 30 - 40 seconds and 22% of all dial-up attempts were unsuccessful. For moving mobiles the connection establishment time increased to 83.3 seconds maximum and the minimum was 29.5 seconds. With 33 % the percentage of unsuccessful attempts was much higher as for the static conditions. But it was almost always possible to establish a connection on the second try despite the difficult conditions (handover, changing quality of the radio signal) which can be experienced by moving mobiles.

In a multi dimensional approach the influence of IP parameter settings for Maximum Transmission Unit (MTU) and Window Size (WS) was evaluated. With a careful selection of parameter values the transfer rate could be increased by almost 100% (see Fig. 9). For static conditions with good quality of the radio signal the transfer rate could be increased form 3.6 Kbit/s to almost 7 Kbit/s for a MTU of 808 byte and a WS of 2304 byte.

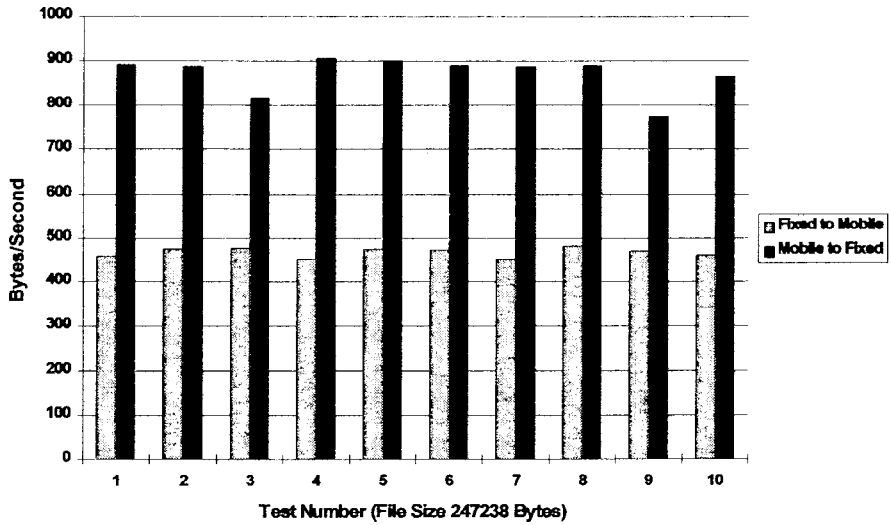

Figure 8 Transfer rate for static mobiles with good quality of the radio signal

Network Part Results Summary: Optimum selection of types and adaptive configuration of transport protocols according to application requirements is very important to guarantee satisfying QoS.

The UAL developed as part of the MASE enables users to receive mobile multimedia services with satisfying QoS over heterogeneous networks.

Especially, for moving mobiles the measurement and monitoring of the bearer services and protocol stack activities is necessary to improve the reliability of the involved connections.

The design of the UAL concerning GSM data bearers should take the following points into consideration: for optimum throughput small TCP window sizes (2-3 segments) should be used. For the non-transparent GSM data bearer the maximum TCP segment size should be used as the GSM Radio Link Protocol guarantees error free transmission on the radio path. For the transparent GSM data bearer the minimum TCP segment size should be used as the transmission on the radio path is not error free.

Future measurements will reveal which parameters can be optimised for an adapted TCP/IP protocol stack..

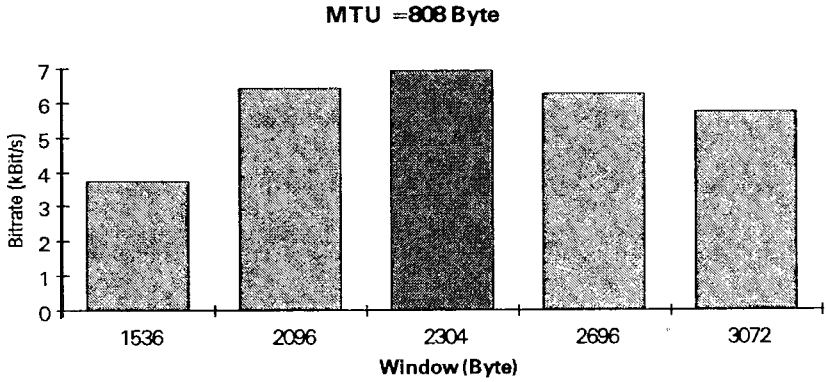

Figure 9 Transfer rate for MTU of 808 byte and varying WSs

Conclusion

In this paper we have presented the architecture of a mobile middleware (MASE) capable to support legacy and mobile aware applications. This MASE is a direct answer to the changing business requirements of mobile professionals towards the end of this millennium. Through the Mobile-API of this new MASE application programmers and end users have access to a set of new services (disconnected operation, scalability of services, location awareness, and personalisation) which give optimum support for ubiquitous mobile computing.

To fully exploit the services of this architecture a new multimedia business information system Infomotion has been developed. This system was provided to end users during an application trial. The results from this first Field trial can be exploited manifold (MASE-, application development, and future field testing).

Most remarkable was the feedback given by the end-users which clearly reflects the need and usefulness of applications like Infomotion. The end users, which were faced for the first time with this new kind of services, were extremely keen on personalisation, and location awareness. This services are seen as the step from a new well designed WWW site towards a real new value added service.

Further results from both the user and network part clearly showed that present available cellular networks are not very well suited to cover the demands of multimedia communications. Especially the QoS in terms of reliability and throughput requires the services provided by the MASE. More specifically, the services of the UAL are required to enable users to receive mobile multimedia services with satisfying QoS over heterogeneous networks.

Outlook

For future field trials, the business information system Infomotion will be continually upgraded to encounter user requirements identified from experiment 0. A MASE at different level of development will be available for the next trials and the infomotion system will be adapted accordingly to make efficient use of the services provided by the MASE through the Mobile-API.

Acknowledgments

This work has been partially funded by the Commission of the European Communities in the ACTS program under the AC034 OnTheMove Project. The authors would like to acknowledge the contributions of their colleagues from Bonnier Information Services AB, BT, Burda Com GmbH Media Solutions, Deutsche Telekom MobilNet GmbH, Ericcson Eurolab Deutschland GmbH, Ericsson Radio Systems AB, Tecsi, IBM, Royal Institute for Technology, RWTH Aachen, Siemens AG, Swedish Institute of Computer Science, Sony Deutschland GmbH, IONA Ltd., National University of Singapore, although the views expressed are those of the authors and do not necessarily represent those of the project as a whole.

References

[1] *ACTS homepage*: http://www.uni-stuttgart.de/SONAH/Acts/PRtit.html
[2] *OnTheMove WWW homepage: http*://www.sics.se/~onthemove
[3] R.S Swain, *RACE UMTS Vision*, Brussels, 1996
Abstract: http://www.uni-stuttgart.de/SONAH/ Acts/mobility/umts_vi.htm
http://www.uni-stuttgart.de/SONAH/Acts/ mobility/doc/umts_vi.doc (Word document)
[4] *BIS Strategic Decisions Inc Study*, 1995
[5] DeBelina, J. (1995). *The Wireless Data Market*, IEEE Communications Society. New York Chapter, 89th Seminar Proceedings, May 18

[9] Robert K. Heldman: Future Telecommunications (Information, Applications, Services, & Infrastructure), McGraw-Hill, Inc., 1992
[7] P. J. Thomas: Personal Information System: Business Applications, Stanley Thornes Ltd., 1995
[8] Ovum Ltd: Data over GSM: Market development, London 1996
[9] Park, Meggers & Ludwig, *Mobile Middleware: Additional functionality to cover wireless terminals*, MoMuC-3, 1996
[10] T. Wierlemann, David, K, et. al. *Field Trials of the European Project: OnTheMove (Testbeds and Results)*, MoMuC-3, 1996.
[11] European Commission DG XIII/E: *Strategic Developments for the European Publishing Industry towards the Year 2000*, Luxembourg 1996.

This page intentionally left blank.

INFORMATION-PROVIDING MECHANISM COMBINING BROADCAST AND ON-DEMAND MODES IN MOBILE COMPUTING ENVIRONMENTS

Masanori Tanabe, Satoshi Hakomori, and Ushio Inoue

Laboratory for Information Technology
NTT Data Corporation
Kawasaki city,
Kanagawa 210 Japan

tanbe@lit.rd.nttdata.co.jp

Abstract: People with portable computers are to be able to retrieve a variety of information at any time and at any location by using wireless communication systems. Since the bandwidth of wireless communication systems is narrower than that of wired communication systems, however, it is difficult for servers to use wireless systems to send much information to many clients. This paper proposes an information-providing mechanism combining broadcast and on-demand modes of data delivery. The broadcast mode enables a server to support many clients via narrow bandwidth and the on-demand modes enables each client to access different data. An algorithm deciding which data should be broadcasted is presented. It minimizes expected length of waiting time of clients to get needed data from the server. Evaluation results show the effectiveness of this algorithm.

1 INTRODUCTION

Recent advances in wireless communication technologies have made it possible for portable computer users to access databases in remote servers at any time and at any location. A variety of information services for many users will be provided soon, but the limited bandwidth available in wireless communication systems causes following problems[1] [2]:

(a) When many users access databases at the same time, the communication channel becomes busy and some clients have to wait until others finish their access to the databases.

(b) Since the data transfer rate is generally lower than those of wired communication systems, users need to wait a long time whenever they access large amounts of data, such as graphic, audio and video data.

To relax the problems caused by the bandwidth limitation, broadcasting should be utilized in nature of wireless communication. However, it will take a long time to broadcast a large amount of data due to their low data transfer rates. This results in very long waiting time for users to get data they need. We are developing an information system with a mechanism combining broadcast and on-demand modes of data delivery. Our assumptions on the environments where our system will be used are the following:

(a) Data which depends on each location is provided in the fixed area such as museums and exhibition halls.

(b) Each user with a portable computer wants to get the information while in the fixed area.

(c) A wireless communication system with relatively narrow bandwidth and low speed is used to deliver data from servers to users.

(d) There may be many users in each area, and they may want to get data from the server at the same time.

This paper is organized as follows: Section 2 describes two kinds of conventional information-providing mechanisms using a wireless communication system. Section 3 presents the structure of our information-providing system with a new mechanism combining broadcast and on-demand modes. It also presents how to decide which data is to be broadcasted. Section 4 shows evaluation results of the new mechanism. Section 5 refers a related work by Imielinski et al. Section 6 concludes by summarizing this paper and mentioning our plan for future work.

2 INFORMATION-PROVIDING MECHANISMS USING A WIRELESS COMMUNICATION SYSTEM

There are two modes of information-providing mechanisms using a wireless communication system: on-demand mode and broadcast mode.

(A) On-demand mode

In this mode a client sends a request to a server which then provides the requested data. Although the client can get exact data from the server, the server needs to provide a data for each request. If the number of clients increased, the response time would therefore increase and the network would be busy. Dynamic Documents [3] is a mechanism that enables servers to send data in qualities fitted to each client's resource limitations (such as display size and resolution), but the amount of data increases when many users access the servers at the same time.

(B) Broadcast mode

In this mode servers periodically broadcast a fixed set of data to clients, and the response time is constant regardless of the number of clients. Since the data is sent periodically, however, there may be a long waiting time until the necessary data is sent next time. The Broadcast Disks [4][5] is a mechanism for broadcasting data that many users access more frequently than data that few users access, but the waiting time is still very long when a lot of data is being provided.

3 COMBINING BROADCAST AND ON-DEMAND MODES

3.1 Information-providing system

When data is sent in the on-demand mode, users can get the exact data they need. And when data is sent in the broadcast mode, many users can get the data at the same time. So we propose an information-providing system combining broadcast with on-demand modes so that many users can get the exact data they need. Data that servers provide is divided into popular data, which is frequently accessed by many users, and unpopular data, which is rarely accessed because it is not of a common interest. The proposed mechanism broadcasts only the popular data. This means that users do not have to send a request to the server in order to get the popular data. When users want to access the unpopular data, on the other hand, they have to send a request to the server.

In an exhibition hall, for example, users may frequently access a guide map. If such data is sent in the broadcast mode, the bandwidth of the wireless communication system can be used efficiently because clients do not need to send the same requests to the server. A user may access data of a product displayed in the exhibition hall. If the data in which he is interested is not common interest among the users, it is not broadcasted. But our system allows each user to access any data of individual interest in the on-demand mode. The new mechanism combining broadcast and on-demand modes can thus provide a lot of data to many users.

3.2 System Structure

Figure 3.1 shows the structure of the proposed information-providing system, which we call MobiCaster. MobiCaster consists of servers and clients, and their roles are as follows:

(A) Servers

The servers select the way to provide data to clients and sends data to the clients in a fixed area that server is in charge of. They determine which data is popular by counting users accesses, and they send popular data in the broadcast mode. They also send a time table of broadcasting, and they update this time table and send it again whenever the popular data changes. When unpopular data is requested, the server sends it to the requesting client in the on-demand mode.

(B) Clients

The clients recognize whether data requested from users is popular or unpopular. The clients use a time table to recognize the popular data. If the data is popular data, clients get it when it has just been sent in the broadcast mode. If the data is unpopular data, clients send requests to servers and receive the data sent by the server.

Figure 3.2 illustrates the roles of servers and clients, and data flow how each client gets data from a server in MobiCaster.

3.3 Mechanism

Single-downlink broadcast mechanism

If there are many channels to receive data, clients have to select a channel that the data they need is transmitted. As the result, a procedure of the channel selection is necessary for clients. However, if servers send data in

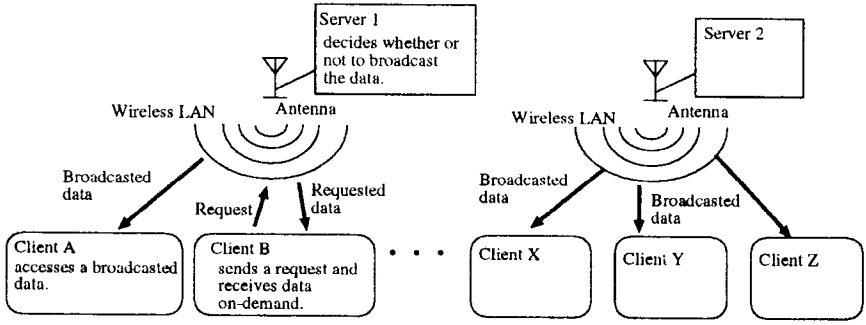

Figure 3.1 Structure of our information-providing system.

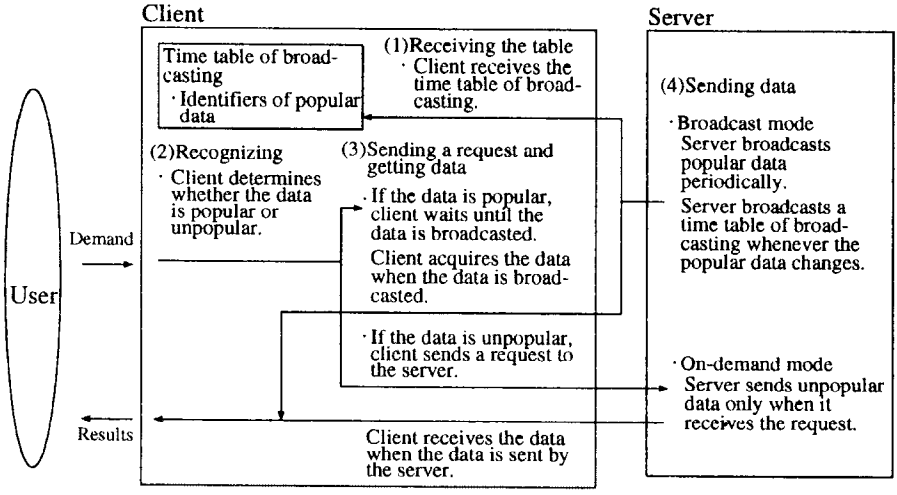

Figure 3.2 Data flow in MobiCaster.

only one channel, this procedure is not necessary for them. We propose a new mechanism using the same channel to deliver both popular and unpopular data. This Single-downlink broadcast mechanism (SDB) has one downlink channel and one uplink channel:

(a) The downlink channel carries popular data periodically and carries unpopular data after the server received a request from a client.

(b) The uplink channel carries users' requests to the server in the on-demand mode.

Each server has its own SDB mechanism and can send data without being influenced by other servers. Therefore, if the requests for unpopular data increase, the server can decrease the rate of sending popular data and increase the rate of sending unpopular data. And if the requests for unpopular data decrease, the server can increase the rate of sending popular data. Like this case, because the server change the amount of popular data, the server can efficiently use bandwidth in accordance with the conditions.

Assignment of the data to the downlink channel

The SDB mechanism uses time slots (Fig. 3.3) to send data in the downlink channel. Some of the slots in one period are used for popular data in the broadcast mode, and others are used for unpopular data in the on-demand mode. Popular data appears in the same order again in the next period, but unpopular data may not appear again in the next period because clients may not request it.

The ratio of the number of slots for popular and unpopular data in a period is decided according to how frequently the data is accessed by users. Popular data is assigned to the slots in order from the beginning of the period so that we can easily calculate the expected length of time a client waits for the requested data.

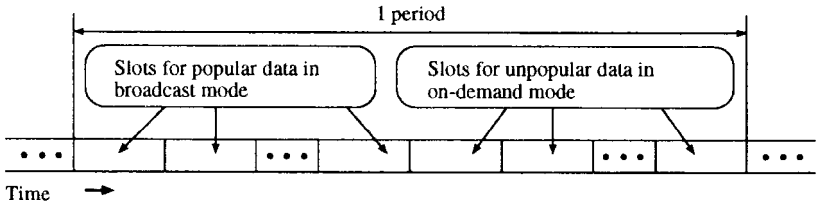

Figure 3.3 Time slot assignment.

3.4 Deciding data which is to be broadcasted

To implement the proposed information-providing mechanism, we have to answer the following two questions:

1. How does the system distinguish popular data from unpopular data ?
2. When the server broadcasts popular data, how much does the server send per unit time ?

In this section, we describe the assumptions used to decide which data is popular and describe ways to calculate the expected length of time clients wait until they get data. We also describe the way to decide which data should be sent in the broadcast mode.

An algorithm to calculate an expected length of the waiting time

The assumptions under which the expected waiting time is calculated are the following:

1. Size of each data provided by servers is fixed.
2. The requests from each client occur according to the M/M/1 queuing model.
3. There are N pieces of data, k of which are broadcasted and the rest of which are sent when requested by clients.

Equations to calculate the expected length of the waiting time until clients get the data are following.

(a) Expected length of waiting time in the on-demand mode

Clients who want to get unpopular data have to wait while popular data is being sent in the broadcast mode because the SDB mechanism sends unpopular data in the on-demand mode after it sends popular data in the broadcast mode. The waiting time while data is being sent in the broadcast mode added to the waiting time in the sending queue of the server gives the expected length of the waiting time until clients get data in the on-demand mode:

$$D_y = \frac{\rho}{1-\rho} \times \frac{1}{S - S_b} + W_b \tag{3.1}$$

D_y Waiting time when clients get unpopular data in the on-demand mode.
ρ Ratio of slots which unpopular data uses.
S_b Number of slots used to send popular data (per unit time).
S Number of slots per unit time.
W_b Waiting time until popular data is being sent.

ρ is the ratio of slots used to send data in the on-demand mode The number of all requests from all clients multiplied by the ratio of the requests in the on-demand mode is the number of slots used in the on-demand mode. And the number of the data requested from clients divided by the number of the slots in the on-demand mode is ρ.

$$\rho = \frac{Q \sum_{y=k+1}^{N} \lambda_y}{S - S_b} \tag{3.2}$$

λ_y Rate at which data y is accessed per unit time.
Q Total number of requests per unit time.
N Total slots in a period.
k Slots used in the broadcast mode.

Then we calculate the average waiting length W_b until unpopular data is sent. The average waiting time while the popular data is sent in the broadcast mode added to the time until a slot used in the on-demand mode begins is the average waiting time. k slots are used to send the popular data. So the average waiting time in the broadcast mode is the time taken up by half of k slots. And the probability of data being sent in the broadcast mode is k/T. The probability with which data is sent in the on-demand mode is (T-k)/T and the average waiting time in the on-demand mode is the half-time of one slot. The half-time of one slot is 1/2 because we suppose the length of a slot to be 1 unit time. So the average waiting time until unpopular data is sent is given by

$$\begin{aligned} W_b &= \left\{ \frac{k}{T} \times \frac{k}{2} + \frac{T-k}{T} \times \frac{1}{2} \right\} \times \frac{1}{S} \\ &= \frac{k^2 + (T-k)}{2T} \times \frac{1}{S} \end{aligned} \tag{3.3}$$

T Number of slots until the same data which is sent in the broadcast mode is sent again in the next period.

(b) Expected length of the waiting time in the broadcast mode

Because the SDB mechanism broadcasts the popular data periodically, a period is divided by both the half of a period and the number of slots per unit time gives the expected length of the waiting time until clients get the popular data.

$$B_x = \frac{T}{2} \times \frac{1}{S} \tag{3.4}$$

B_x Waiting time when clients get popular data in the broadcast mode.
S Number of slots per unit time.

(c) Expected length of the waiting time until a client gets a data which is either popular or unpopular

The expected length of time the clients wait until they get popular data multiplied by probability with which each item of popular data is received by clients, added to the expected length of time the clients wait until they get unpopular data multiplied by probability with which each item of unpopular data is received by clients, is the expected length of a waiting time until clients get the data which is either popular or unpopular:

$$W = \sum_{i=1}^{k} P_i B_i + \sum_{i=k+1}^{N} P_i D_i \tag{3.5}$$

P_i Probability with which users get data *i.*
D_i Expected length of a waiting time in on-demand mode.
B_i Expected length of a waiting time in broadcast mode.
N Total number of data items.
k Number of popular data items.

The SDB mechanism decides k so that W is the smallest and sends k data items as popular data in the broadcast mode.

3.5 Deciding which data to send in the broadcast mode

MobiCaster determines popular data according to the probability of its being accessed:

1. MobiCaster calculates the expected length of the waiting time (Equation 3.5). k is beginning with 1 and k is increased by 1 each to N.

Namely, it calculates the expected length of the waiting time in each case that one data item is broadcasted, that two data items are broadcasted, ..., N data items are broadcasted. Data is sorted in order of its probability of being accessed.

2. MobiCaster broadcasts *k* data items in order, starting with data that has the highest probability of being accessed.

For example, suppose the following conditions:

1. Server has seven data items.
2. Eight requests per unit time are sent from clients.
3. Server can send seven data items per unit time.

The probabilities of accessing each data item are listed in Table 3.1. When one item (Data A) is sent in the broadcast mode; that is, when $k = 1$, the expected length of the waiting time is 9.31. When two items (Data A and B) are sent in the broadcast mode, (that is, $k = 2$), the expected length of the waiting time is 0.656. The expected length of the waiting time is 0.475, which is the shortest, when three items (Data A, B and C) are sent in the broadcast mode ($k = 3$).

Table 3.1 Sample of data access probability and expected length.

Data number	A	B	C	D	E	F	G
Access probability	0.26	0.25	0.23	0.20	0.03	0.02	0.01
Pieces of data (*k*)	1	2	3	4	5	6	7
Expected length	9.31	0.656	0.475	0.485	0.495	0.500	0.500

Therefore, in the case where data are accessed according to the probabilities listed in Table 3.1, the server sends three data ($k = 3$) in the broadcast mode.

4 EVALUATION

In the first half of this section we evaluate the relation between the expected length of time clients wait until they receive a needed item of popular data in the broadcast mode. And in the second half of this section we evaluate the relation between transmission speed and the number of data which servers can send.

Suppose that there are many requests from users and the server cannot send all data to the clients. The server thus has to send some data in the broadcast

mode and other data in the on-demand mode. We let the length of the time unity to normalize the length of the time. In the evaluation environment the parameters of the equations in Sec. 3.4 are the following:

1. Length of the time slot is one unit of time.
2. There are seven slots per unit of time.
3. There are eight user requests per unit of time.

Figure 3.4 shows the relation between the ratio of access to the popular data and the expected length of time clients wait until they get the data. It shows three cases differing by the number of slots used in the broadcast mode. In Case 1, one slot is used in the broadcast mode and six slots are used in the on-demand mode. In Case 2, two slots are used in the broadcast mode. In Case 3, three slots are used in the broadcast mode. The X-axis is the ratio of requests to access popular data to total number of requests per unit time. For example, when the ratio of the accesses is 0.65, if 100 data requests are received from clients, the number requests for popular data is 65 and the number requests for unpopular data is 35.

If one popular data were sent in the broadcast mode and the ratio of access was 0.65, the expected length of the waiting time would be about 0.401 unit time. If the ratio of access to one data item were 0.5 and the ratio of access to another data item were 0.2, the expected length of the waiting time would be about 0.433 when both items were sent as popular data in the broadcast mode. However, when only the item for which the ratio was 0.5 was sent as popular data in the broadcast mode, the expected waiting time would be about 0.452. So, in this case, the expected time is longer when the server sends two data items in the broadcast mode. This shows that the expected length of the waiting time depends both on the number of the data items sent in the broadcast mode and on the ratio which the data is accessed by clients.

Figure 3.5 shows the speed of data transmission:

$$Speed = \sum_{i=1}^{k} P_i \frac{1}{D_i} + \left\{ \sum_{i=k+1}^{k} P_i Q \right\} \left\{ \sum_{i=k+1}^{k} P_i \frac{1}{B_i} \right\} \tag{3.6}$$

If some of the data were sent as popular data in the broadcast mode, clients which requested that data could get it at the same time. So the expected speed in the broadcast mode multiplied by the number of requests for popular data is the transmission speed for the broadcast mode. If a data item were accessed by 75% of the clients and only that data was sent as popular data in the broadcast mode, the maximum number of clients the server could send it would be 27. The SDB mechanism limits the number of requests from users

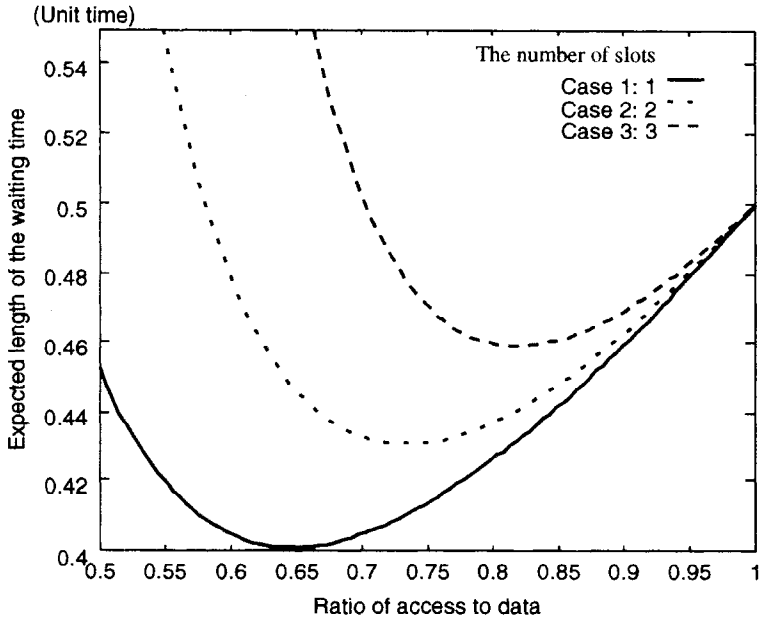

Figure 3.4 Expecting length of the waiting time.

because the number of slots used in the on-demand mode is limited. That is, D_i is not less than 0. If the probability of accessing a data item is 0.75 and the number of requests is 27, D_i is equal to 0. When the number of request is more 27, D_i is less than 0. So the maximum number of clients to which the server can send is 27 when the probability of accessing a data is 0.75. When MobiCaster provides data in a suitable way, it can provide data to more clients than it could if it used only on-demand mode.

If only the on-demand mode were used, whenever there were more than seven requests, the SDB mechanism would pile the data the server has to send into the sending queue of the server because the transmission speed of SDB is low. For example, when the single downlink can send seven data items per unit time, if there were eight requests per unit time, the data for all requests could not be sent.

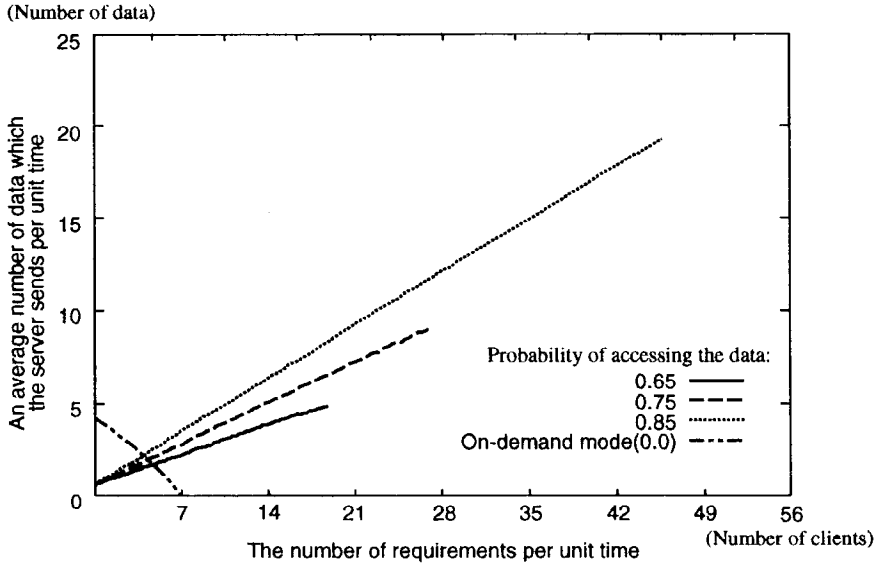

Figure 3.5 Number of data which a server can send.

5 RELATED WORK

Imielinski et al. proposed a way of combining broadcast and on-demand modes to reduce energy consumption [6][7]. Their proposal has the following features:

(a) The system has several channels (called multicast channels) that can be used to carry data. One of these channels is used for sending data required on demand, and the others are used in the broadcast mode.

(b) Each channel carries single data periodically. That is, there is only one slot in each channel.

This mechanism can be called a Multiple Downlink Broadcast (MDB) mechanism in contrast to our SDB. MDB requires as many channels to broadcast as kinds of data, but the number of channels available is usually limited in commercial wireless communication systems. For example, there is only one channel in a wireless LAN using TCP/IP. Our mechanism fits better to such a wireless communication system.

When the system has many channels, data are sent through one. Thus, clients select the channel in which the data is sent, and then they wait until the data is sent. It is more difficult to make a structure that selects a channel and waits for the data than to make a structure that only waits on a single channel until data is sent.

6 CONCLUSION

The bandwidth of wireless communication systems is so limited that it is difficult to provide much information to many users. Broadcasting information is a partial solution to this problem but it cannot provide the exact data each user wants. We have therefore proposed an information-providing system, called MobiCaster, that combines the broadcast and on-demand modes. MobiCaster broadcasts the popular data accessed by many clients, but it sends seldom accessed data only when it receives requests from clients.

We also proposed a single-downlink broadcast (SDB) mechanism in which popular and unpopular data are sent using one channel. Some slots are assigned to popular data and other slots are assigned to unpopular data, and the length of time clients wait for data is shorter than it is when only the broadcast mode or only the on-demand mode is used. We have shown a formula for calculating the waiting time and have shown a criteria for selecting data to be sent in the broadcast mode in order to minimize the waiting time. We have also shown results of evaluating the proposed mechanism.

The SDB mechanism needs to monitor the probability of data access. We are developing a prototype MobiCaster system that has a mechanism for this requirement and are planning to evaluate the actual waiting time in a wireless communication system.

References

[1] T. Imielinsky and B. R. Badrinath: “Data Management for Mobile Computing,” SIGMOD Record, Vol. 22, No. 1, Mar. 1993.

[2] G. H. Forman and J. Zahorjan: “The Challenges of Mobile Computing,” IEEE Computer, Apr. 1994.

[3] M. F. Kaashoek, T. Pinckney, and J. A. Tauber: “Dynamic Documents: Mobile Wireless Access to the WWW,” Proc. of IEEE the 1st Workshop on Mobile Computing Systems and Applications, Santa Cruz, CA USA, Dec. 1994.

[4] S. Acharya, M. Franklin, and S. Zdonik : “Dissemination-based Data Delivery Using Broadcast Disks,” IEEE Personal Communication, Vol. 2, No. 6, Dec. 1995.

[5] S. Acharya, M. Franklin and Stanley Zdonik: “Prefetching from a Broadcast Disks,” 12th International Conference on DATA ENGINEERING, pp. 276-285, Feb. 1996.

[6] T. Imielinski and S. Viswanathan: “Adaptive Wireless Information Systems,” Proc. of SIGDBS workshop, pp. 19-41, Tokyo, Japan, Oct. 1994.

[7] T. Imielinski, S. Viswanathan and B. R. Badrinath: “Data on Air: Organization and Access,” Technical Report WINLAB-TR-95, Mar. 1995.

This page intentionally left blank.

Wireless ATM - Multimedia Service Platform

Tom Leskinen, Markku Niemi

Nokia Mobile Phones, Wireless Data, P.O.Box 68, FIN-33721 Tampere, Finland
email: tom.leskinen@nmp.nokia.com, markku.niemi@nmp.nokia.com

Abstract

The need to provide multimedia services in the wireless environment has increased remarkably. This is being driven by the rapid service development in the computer environment as well as by the increasing penetration of wireless communication systems. In this paper we describe the development of wireless multimedia services and their requirements. A gradual introduction of multimedia capabilities to the users is discussed. For each phase of this process a corresponding wireless system is identified. Finally, Wireless ATM is presented as a potential technology to be able to provide enough bandwidth along with a good Quality of Service (QoS) support for new multimedia services in a wireless environment.

1. Introduction

Communications services have evolved from pure speech, text or computer data towards more rich and easily interpretable forms commonly know as multimedia. As the services have become more attractive to users, people have started to utilise them extensively. An outstanding example is the Internet and especially the World Wide Web (WWW). So far these services are available mainly for wired users. However, there is an increasing need to use these multimedia services also in wireless manner. This need evolves from the wired multimedia service environment; users want to have similar environment, i.e., applications and services, in wireless use as in wired use. Most of the computer environment multimedia services as well as some niche services could benefit remarkably from wireless means of communication. In addition, for some services, such as messaging, wireless means of communication is essential. Furthermore, wireless communication opens up possibilities for service developers and providers to introduce

new, innovative services taking full advantage of the wireless communication.

Wide range of wireless systems are being standardised in different foras. A common nominator for all of these efforts is that they aim towards more attractive services, higher bit rates for the end user and try to achieve similar capabilities than available in wired environment, especially this concerns multimedia capabilities. Existing digital cellular mobile systems, like GSM and PCS 1900, are being developed to contain higher bit rate data services at the same time as 3rd generation mobile systems are being developed. Most recently started activity is wireless ATM standardisation work carried out in ETSI STC RES10 and in ATM Forum wireless ATM group. [2]

In this paper we focus on the multimedia services in the wireless environment. These services are introduced and their characteristics are analysed. We consider wireless as a transmission media for these services and discuss some of the special characteristics of the wireless platforms. Furthermore, a few existing wireless systems are briefly analysed and their properties are discussed. Special attention is paid on Wireless ATM. It is investigated as a solution for wireless short range multimedia communication and it's capabilities as a wireless multimedia platform are analysed. Special attention is paid on the justification for the wireless ATM based applications and services. Chapter 2 describes the multimedia services. Introduction to the multimedia service characteristics in the wireless environment is given in Chapter3. Chapter 4 discusses the development of different platform, wired and wireless. In Chapter 5 we briefly investigate the different wireless systems relevant for the wireless multimedia service provision. Wireless ATM is investigate more thoroughly in Chapter 6.

2. Multimedia Services

Multimedia services are emerging rapidly in the computer environment as well as in telecommunications environment. Such services as WWW information access, multimedia data base access, entertainment, news services, video telephony and video conferencing are existing examples, many others are expected in the near future. The possibilities for new multimedia services are unlimited and new services can be expected to evolve with time.

Multimedia services consists of several simultaneous media components, e.g. speech, audio, data, images and video and they can be categorised into several different classes each having different requirements. Here we

classify services according to ITU service classification [5]. The most relevant classes from the wireless point of view are retrieval services (non-real-time) and conversational services (real-time) [1].

Compared to traditional services, like speech and data transmission, multimedia has stringent requirements. Due to the existence of different components, especially video, low error rates are a necessity. Combination of all media components tend to increase the bandwidth requirement and the burstiness of the transmission. In addition, conversational multimedia services calls for short delays to fulfil the real time requirement. Furthermore, different media components are synchronised between each other and have to be transferred between two communicating ends maintaining the timing synchronisation.[3] It should also be noted that the quality of multimedia services is subjective and depends on the personal preferences and expectations of the user. These subjective aspects are still very uncertain and unknown due to the low amount of available experience in usage of multimedia services.

2.1 Multimedia services in wireless environment

Due to the limited capacity available in the current wireless systems, the introduction of multimedia services should be gradual. The introduction of new wireless transmission systems and enhancements of existing systems will increase system capabilities stepwise. The introduction is going to be started in near future when the capacity and required QoS support will be available. The driving forces for the introduction of new services are:

- user expectations based on computer environment,
- wireless systems development,
- development of the coding techniques (speech coding, video coding),
- development on the terminal implementation technology, and
- development of the regulatory environment.

Table 1 below describes envisaged phases of a gradual introduction of wireless multimedia services.

Table 1. *Phases of wireless multimedia service introduction.*

Envisaged Phase	*Description*	*Estimate schedule*
Phase 0	Current situation, i.e., speech or low bit rate data	Available now

Phase 1	Combination of two media components with low QoS	1998/1999
Phase 2	Combination of several media components with flexibility	2001/2002
Phase 3	High Quality wireless multimedia services	2003/2004

Phase 0 - Current situation, provides currently available *single media services.* No real multimedia services available in wireless systems. Good quality speech or low bit rate data is available in most of the digital mobile systems.

Phase 1 - Combination of two media components with low QoS is a continuation of phase 0. Combination of existing components, speech and data and possibly addition of new components, e.g. still pictures or low bit rate video. Development is mainly based on the enhancements of current mobile communications systems to provide higher bit rates. Bit rates between 28.8 kbit/s and 64 kbit/s will be available.

Phase 2 - Combination of several media components with flexibility will be available with 3rd generation mobile communications systems. System have to be optimised for multimedia in order to support multimedia services in an efficient way. Multimedia services envisaged to be available in Phase 2 consist of conversational services like good quality video telephony, audio graphics and video conferencing. Flexibility of conversational services allows dynamic managing, i.e. adding, modifying and dropping, of media components during a call. Retrieval services available are, e.g., WWW information access containing integrated audio-visual information, entertainment, electronic (3-D) shopping and traffic information delivery (maps and traffic information). Phase 2 introduces a significant step from Phase 1, in terms of quality. However, the offered services stay mostly similar.

Phase 3 - High quality wireless multimedia services. All relevant multimedia services could be provided in a wireless manner. Phase 3 is a major step towards full multimedia capabilities in a wireless environment. Same multimedia services could be available as in wired environment. This phase sets the most stringent requirements to the underlying network. This phase, as well as phase 2, are strongly dependent of the overall multimedia development in the near future. Without a large scale penetration and acceptance of high quality multimedia for everyday routines there will not be strong demand for the wireless equivalent. Furthermore, a development

of transmission capabilities in the backbone networks of these multimedia applications has to be remarkable.

3. Multimedia Service Characteristics

Multimedia data bases and multimedia applications of the future are envisaged to contain huge amounts of high quality information, audio, video as well as data. Access to this information could be interactive and require highly flexible transmission capabilities in order to be able to transmit all required multimedia information, e.g. images, 3D-graphics, high quality video, data base information, high quality audio and speech. The importance of accessing the information in real time is rising in the business community as well as in the leisure time access. Real time access to information requires possibility for location independent access. This requirement can be best fulfilled by wireless systems.

Wireless real time access to information is crucial in order to be able to make successful business decisions. Leisure time access requirements are also increasing as applications evolve. Electronic shopping being a good example to introduce an everyday requirement to access multimedia information. Electronic newspaper is another similar application concerning vast amount of users. In addition to these services there will be services that are targeted to wireless environment and benefit of being accessible in wireless manner. All location dependent information such as maps, city information, hotel and restaurant information and traffic information, are extremely valuable as being available in wireless environment.

Table 2. *Required bandwidth for different multimedia applications [11].*

Application	**Data rate**	
	Uncompressed	**Compressed**
Voice, *8ksamples/s, 8bits/sample*	64 kbit/s	2-4 kbit/s
Video telephony (10fps), *frame size 176x120, 8bits/pixel*	5.07 Mbit/s	8-16 kbit/s
Audio conference, *8 ksamples/s, 8bits/sample*	64 kbit/s	16-64 kbit/s
Video conference (15fps), *frame size 352x240, 8bits/pixel*	30.41 Mbit/s	64-768 kbit/s
Digital audio (stereo), *44.1 ksamples/s, 16bits/sample*	1.5 Mbit/s	128 kbit/s - 1.5 Mbit/s
Video file transfer (15fps), *frame size 352x240, 8bits/pixel*	30.41 Mbit/s	384 kbit/s
Digital video on CD-ROM (30fps),	60.83 Mbit/s	1.5 - 4 Mbit/s

Application	Data rate	
	Uncompressed	**Compressed**
frame size 352x240, 8bits/pixel		
Broadcast video (30fps), *frame size 720x480, 8bits/pixel*	248.83 Mbit/s	3-8 Mbit/s
HDTV (59.94fps), *frame size 1280x720, 8bits/pixel*	1.33 Gbit/s	20 Mbit/s

The throughput requirement of multimedia cannot be explicitly defined. Each multimedia application and service have slightly different throughput requirement due to the varying content and available user interface. Table 2 lists some example throughput values for encoded multimedia applications. Values vary from very low bit rate, i.e., several kbit/s, to very high bit rates, i.e., several Mbit/s. The flexible nature of multimedia adds complexity to throughput definition, because of dynamically varying amount of involved media components and their parameters. For high q multimedia (phase 3) a throughput of several Mbit/s is required, respectively.

In wireless environment transmission errors have a strong affect to the transmitted multimedia information and especially to video information. Video coding is basically based on transmitting the changes between the subsequent pictures, i.e., the temporal redundancy being eliminated. A much higher compression ratio can be achieved compared to transmission of actual subsequent pictures. This however introduces an drawback of being much more sensitive to the transmission errors. An error on the transmitted information equals to an visual artefact in the encoded video stream. Furthermore, due to the transmission of only change information the artefact is not eliminated from the picture, until the whole picture is transmitted again. Transmission of a whole picture, introduces huge bit rate increase to be handled by the channel. Erroneous environment could have drastic effects to the QoS offered to users. Therefore a wireless multimedia system should be able to guarantee the negotiated transmission , e.g., BERs from 10^{-6} or even up to 10^{-8}.

Encoding video sequences equals to encoding of consecutive still pictures each containing temporal variations, i.e. different amount of information. The encoder output bit rate is therefore of varying nature (see figure 2). Video encoders normally include an output buffer which adapts the output of the whole encoder to the available channel rate. A video sequence containing a lot of movement at one time produces large amount of information, which cannot be transmitted on the channel or even stored into the buffer. To cope with this kind of varying situations the encoder is

controlled to produce less bits, i.e. buffer signals the encoder to use lower quality for each consecutive picture.

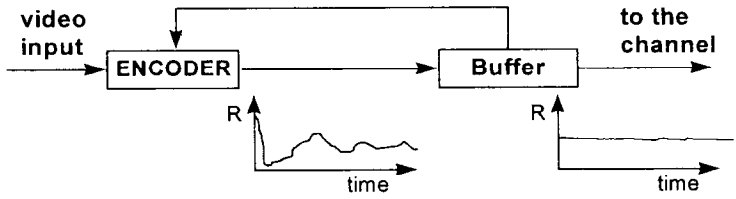

Figure 1. *Video encoding structure and behaviour of produced bit rate.*

In order to provide a constant quality encoded video sequence the varying output of the encoder should not be smoothed by controlling the quality of consecutive pictures. The variable output should be transmitted as such to the transmission channel, capable of variable rate behaviour. The required rate of variability depends of the actual sequence, but can be extremely demanding. A very efficient packet based transmission system or a flexible circuit switched system would be optimal transmission media to support variable rate transmission. [10]

4. Service platforms

4.1 Wired platforms

In fixed networks, multimedia services are transmitted over various platforms. Recent advances in video coding techniques enable relatively good quality video conferencing over ISDN. The rapid development of PSTN modem technology suggest that this platform could be used for video transmission as well. The phenomenal success of Internet has had an enormous impact for the technical development of these systems and is a good example of service driven technology development. Also in local area networks, development has been rapid. Especially the evolvement of Ethernet is worth noting. These systems have their indisputable merits in the tasks they were designed for. None of these systems, however, are especially suited for the transmission of different kinds of multimedia information. ATM, on the other hand, promises to fulfil these requirements. ATM provides an easily scaleable, high-speed switching technology for different types of services. The ability of ATM to offer guaranteed Quality of Service (QoS) to individual connections according to their needs provides good support for large scale of different service types. This

property along with highly efficient transmission enables ATM to support the demanding requirements of new phase 3 multimedia applications.

4.2 Wireless platforms

Portable terminals, including laptops and Personal Digital Assistants (PDAs), are rapidly gaining prominence and the sales expectations are high. Due to the mobility of these devices there is a natural need to be able to perform information transfer in different places and situations. Wireless communication is a natural solution to fulfil these needs. Keeping this in mind, it is not surprising that wireless communication features are becoming an integrated part of portable devices. A good example of integrated sophisticated computer and communication features is Nokia Communicator 9000, which is a cellular phone integrated with rich PDA features. The development of these equipment towards more user friendly interfaces is an inevitable trend. High resolution colour displays are the obvious target also for portable small terminals.

Another terminal category with different kind of properties are laptop computers. So far, laptop computers can be equipped with adapters providing wireless communication features. These adapters can be either Wireless LAN (W-LAN) cards enabling short range wireless access to the wired LAN or short range terminal to terminal communication. On the other hand, cellular data cards are also available for laptop users. These enable wider coverage but the transmission capacity is rather low. The recent high growth in penetration of laptop computers and trend to include multimedia capabilities will also increase the requirement for wireless multimedia.

However, both of these system categories have their own shortcomings. Mobile telephone systems are primarily designed for circuit switched transmission of speech and data, one information component at a time. The W-LAN systems, on the other hand, are designed for a wireless extension of a LAN. In this context, it is natural that either of these communication systems, although they have their own merits, do not meet the demanding requirements of high quality multimedia information transmission. Further development steps of both these systems are described more thoroughly in the Section 5.

To overcome these problems and in order to be able to better fulfil user needs, QoS guarantees should be provided to the end user in a wireless environment. Wireless ATM will make new types of services possible for the end users by offering ATM QoS, higher capacity and efficient interworking to ATM backbone networks. Wireless ATM is targeted mainly.

for indoor office-like usage (domestic premises and business premises). However it can be deployed also in public places like airports, shopping malls and other densely populated places. [4]

5. Wireless systems

Wireless communications systems developed have been targeted to different environments due to the characteristics of these environments. Outdoor systems normally contain efficient mobility management and support moving users. On the other hand indoor systems can be targeted to more stationary users moving occasionally. Figure 2 describes the available bit rates in different environments and in different types of systems.

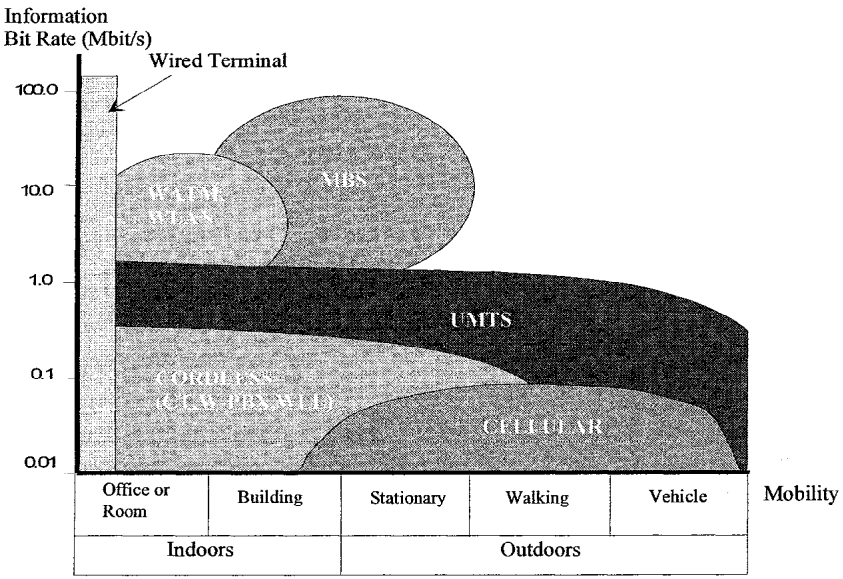

Figure 2. *Wireless communications environments vs. available bit rates.*

5.1 Cellular systems

Mobile communications systems, i.e., cellular systems, are the most popular wireless systems available. These are mainly designed to provide speech service for the subscribers. Digital cellular systems also support low bit rate data services which corresponds to the phase 0 in service development scenario described in Chapter 2.1. These systems provide wide area coverage and support mobility, being therefore suitable for various environments and usage scenarios. The requirement for higher data rates and better support for enhanced data services, including multimedia, have

initiated future development of these systems. The development of e.g., GSM is based on HSCSD (High Speed Circuit Switched Data), GPRS (General Packet Radio Service) and 14.4 kbit/s data services. These services enhance the data rate up to 115 kbit/s for single user. The capability to support multimedia services increases remarkably with these higher data rates and most probably fuels the introduction of first multimedia services in wireless, wide are coverage environment. This corresponds to the phase 1 in service development scenario, which was described in chapter 2.

Research and standardisation work towards 3rd generation mobile communications systems has been going on for several years. UMTS (Universal Mobile Telecommunications System) and FPLMTS (Future Public Land Mobile Telephony System) are standardised by ETSI and ITU, respectively. The target is to get first phases of these systems ready somewhere between year 2000 and 2002. The objectives of 3rd generation mobile communications systems is to provide global roaming, support multimedia services, offer comparable service quality to fixed networks, provide flexible service capabilities up to 2 Mbit/s, etc. As being optimised for data and multimedia, rather than speech, 3rd generation mobile communications systems introduce distinct advantages compared to 2nd generations systems. Higher bit rates, negotiable transmission parameters (e.g. delay, BER) and flexible bearer management will make the QoS of the multimedia services much better compared to earlier mobile systems. Phase 2 in service development scenario is fulfilled technically with 3rd generation systems. [3]

Further development of the 2nd and 3rd generation mobile communications systems from the technology point of view is not more limited than the development of any other system. Limiting factors are rather different, the major limitation being the lack of adequate amount of available spectrum. Current frequency allocations are seen as inadequate for large scale provision of full range of services from mass market speech up to high quality multimedia. However, provision of these services using currently assumed allocations is viable. Availability of additional frequency spectrum for 3rd generation mobile communications purposes are being studied. [9]

5.2 Wireless short range systems - State of the Art

Wireless short range networks primarily intended for unlicensed use on the private premises. These wireless customer premises networks can be stand-alone networks or they can used as extensions to the wired networks. Typically these networks are used for wireless access to a local LAN. Ad-

hoc configurations of these networks enable wireless computer to computer communication without the need of a fixed infrastructure.

Currently, these wireless networks operate at speeds from a few 100 kbit/s to a few Mbit/s. However, bitrates up to 155 Mbit/s are being studied. These networks are privately deployed and they operate in "unlicensed spectrum" meaning they have to accept interference from other users of these frequencies.

The private deployment typically means limited geographical range. Currently, the services these network are able to provide networks are almost exclusively data oriented; most of them are targeted to Ethernet networks. This category of wireless premises data networks is known as wireless LANs. Standards for this type of networks have been developed by IEEE and ETSI.

A wireless LAN standard developed by IEEE committee P802.11 operates in the unlicensed 2.4 GHz ISM band. The 802.11 standard is primarily Ethernet oriented and provides bit rates from 1 to 2 Mbit/s. The standard includes some features which enable the provision of limited service dependent support. However, these features are not enough to be able to provide high quality wireless multimedia services to the users. Furthermore, the 2.4 GHz ISM frequency band is a very hostile environment due to many unpredictable interference sources, such as microwave ovens, utilising the same frequency band. In addition, the characteristics of the access rules for this band make the provision of predictable QoS difficult [7,8].

In Europe, ETSI RES10 is developing a family of standards for wireless broadband communications serving variety of applications, including wireless LANs, mobile wireless ATM access networks and wireless ATM infrastructure networks. These standards operate in the 5 GHz and 17 GHz frequency bands which are assigned for HIPERLAN equipment on a non-protected and a non-interference basis. Furthermore, similar band in the U.S., called Unlicensed National Information Infrastructure (U-NII) band have been designated specifically for high speed wireless premises networks on a unlicensed basis.

HIPERLAN Type 1 - Wireless 8802 LAN standard is primarily intended for providing high speed, short distance radio links between computer systems. This standard is intended to be used for local, on-premises networking. Operation is based on Carrier Sense Multiple Access (CSMA) peer-to-peer (Ethernet based) communication with decentralised architecture. The functional standard is completed. HIPERLAN Type 1 provides bit rates from 1 up to 20 Mbit/s per channel. This bandwidth is enough for most multimedia services. The problem with HIPERLAN Type 1, however, is

that under heavy load the transmission delay increases significantly and cannot be controlled. Furthermore, being an purely shared system anyone willing to use the channel is allowed to do so, i.e., the channel load cannot be controlled. This inevitably results in a somewhat unpredictable QoS. Another significant problem with the HIPERLAN Type 1 standard with regard to the multimedia services is that the system is designed for the transmission of relatively long (Ethernet) packets. The transmission efficiency of short (e.g., ATM) packets is rather low [6].

6. Wireless ATM

As the discussion above suggests, none of the existing wireless networks are able to satisfactorily fulfil the demanding requirements of true multimedia services. In order to be able to provide users new real phase 3 (Chapter 2.1) multimedia services, several Mbit/s, small delay and small BER is required. Clearly, this is not possible using existing wireless systems and even the future 3rd generation cellular systems will not be able to fulfil all these requirements. This calls for a new wireless system being able to meet these demands.

Today, the concept of wireless ATM is being studied extensively and there seems to exists agreement that wireless ATM is one of the most promising technique to meet these stringent requirements.

Wireless ATM is seen as a wireless extension of fixed ATM. As such it should be able to support the main features of fixed ATM networks including the capability to provide bandwidth on demand, and the capability to give support for different traffic types taking their individual requirements into account.

The research on wireless ATM has been going on for some time and many ambitious research projects, for example within the ACTS research programme framework in Europe, are currently underway. Recently, also two standardisation activities have been launched [2]. The wireless ATM air interface work will include Radio Physical layer (PHY), Medium Access Control (MAC), Logical Link Control (LLC) for error control and a wireless control protocol for radio resource management. Special wireless functionality is needed at MAC and PHY levels to cope with the QoS requirements of ATM. Support is needed also to the ATM network to support terminal mobility e.g. handover and location update.

Technically, the most challenging issue is to transmit short cells efficiently over the wireless interface. In addition to these technical challenges, there are many other issues that have crucial meaning to the real (commercial)

success of the concept. These issues include, e.g., the progress of the standardisation work, frequency allocation and frequency etiquette issues both in Europe and in the U.S. Other regulatory issues will also have a significant impact. Many of these questions are still open. In the following, we will focus on some of these issues, along with an general technical overview to the wireless ATM concept.

6.1 Requirements

As discussed earlier in this document, different media components have different bandwidth requirements. Furthermore, their requirements vary also in terms of delay, delay variation, bit error rate, and other Quality of Service (QoS). ATM solves this problem by dividing traffic into different service categories. These are [12]:

- Constant Bit Rate (CBR)
- Real-Time Variable Bit Rate (rt-VBR)
- Non-Real-Time Variable Bit Rate (nrt-VBR)
- Available Bit Rate (ABR)
- Unspecified Bit Rate (UBR)

These service categories together with the quality of the radio link set requirements on the radio subsystem performance. This interaction is illustrated in Figure 3 below.

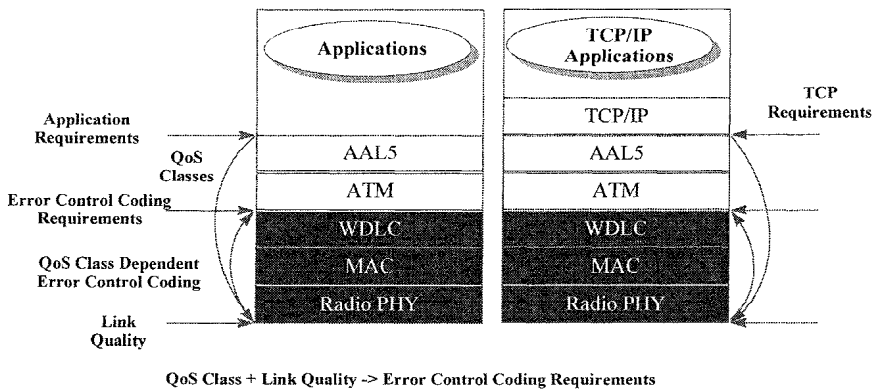

Figure 3. *Requirements for the WATM radio subsystem performance.*

Here, the application requirements for wireless ATM MAC are simplified into a need to support all of the existing ATM service classes. ATM service

classes are more stable and solid base for deriving requirements when compared to evolving applications.

Mobility is maybe the most important feature gained by wireless transmission. On the other hand, mobility introduces some additional requirements for the network. Especially handovers have an important effect to the performance of a system operating in a microcell environment where handovers occur frequently.

The wireless ATM Medium Access Control (MAC) has to connect wireless and wired worlds. The constraints imposed on the MAC are extremely stringent. ATM is a connection oriented technology guaranteeing certain QoS to the user and relying on the statistical multiplexing of ATM cells. Therefore, true wireless ATM MAC should provide service comparable to fixed network and hide the unreliable wireless link.

The radio physical layer (PHY) has to be able to transmit short ATM cells (53 bytes) efficiently. First of all, there is a need to find a suitable frequency band for wireless ATM operation in the unlicensed, unregulated basis.

6.2 Spectrum Allocations

The 5 GHz band is one of the most interesting targets for wireless ATM systems.

In Europe, the CEPT has designated the following frequencies for HIPERLANs in the 5 GHz band and these are also candidates for the Wireless ATM frequencies in the 5 GHz band (see Table 3). In addition, 200 MHz spectrum is available for HIPERLANs in the 17 GHz band (17. 1-17.3 GHz).

Table 3. *HIPERLAN allocation in the 5 GHz band in Europe.*

Band (GHz)	***Peak power density (e.i.r.p.)***
5.15-5.30	1 W

In a recently approved Technical Report, ETSI RES10 has estimated that at least 250 MHz of spectrum, in addition to existing allocation, is needed to fulfil the future application requirements. Therefore CEPT has been requested to designate more spectrum for the HIPERLAN family; this process is on-going.

In the beginning of 1997 the following frequencies were made available in U.S. by Federal Communication Commission (FCC). The allocation was

made for Unlicensed National Information Infrastructure (U-NII) devices on unlicensed basis (see Table 4).

Table 4. *U-NII allocation in the 5 GHz band in the U.S.*

Band (GHz)	***Peak power density (e.i.r.p.)***
5.15 - 5.25	200 mW
5.25 - 5.35	1 W
5.725 - 5.825 (ISM)	4 W

It should be noted that U.S. U-NII allocation and European HIPERLAN frequency band are partly similar in terms of frequency range and power limits. This harmonious development of frequency allocations in a global basis provides a strong support for high speed wireless data devices and gives a good background for further work.

6.3 Standardisation

ETSI RES10 sub-technical committee was the first standardisation body to start a work item on ATM compatible wireless multimedia standardisation. The Wireless ATM group within ETSI RES10 started working in October 1995. Shortly after this the discussions on wireless ATM started also in the ATM Forum. The work item on Wireless ATM within ATM Forum was officially approved in June 1996.

The standardisation work within ETSI RES10 is concentrating to the air interface related areas of wireless ATM system operating in the 5 GHz frequency band. The ATM Forum, on the other hand, will cover both the air interface issues and subject areas related to mobility support in the fixed ATM network (see Figure 4).

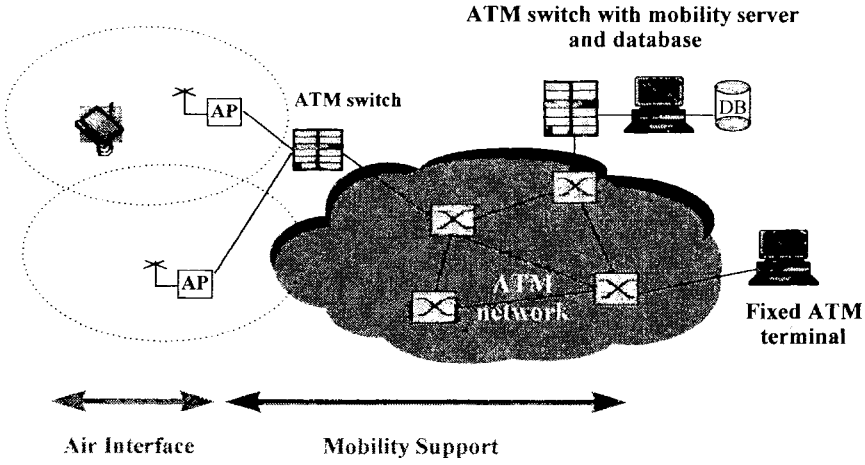

Figure 4. *WATM concept.*

ETSI RES10 has now a working assumption of altogether four different HIPERLANs. This HIPERLAN family is grouped under concept of high speed wireless access networks and customer premises networks. Short descriptions of the three ATM oriented HIPERLAN types follow:

Type 2 - Short Range Wireless Access to ATM Networks. Standardisation work has started recently and it is active work item at RES10 at the moment. The work aims to standardise wireless access to the fixed ATM networks offering a peak bit rate of at least 20 Mbit/s. The architecture will be a microcell like. Type 2 networks will operate in the 5 GHz frequency band.

Type 3 - Remote Wireless Access to ATM Networks is an outdoor application of the technology of HIPERLAN type 2 that will also use the 5 GHz band. It will be able to deliver multi-megabit ATM services over distances of up to 5 km. Different antenna configurations and protocol adaptations will be required to achieve this. HIPERLAN Type 3 is the high speed member of a family of wireless local loop networks being addressed within ETSI.

Type 4 - Wireless Interconnect to ATM Networks is intended for short range, high speed interconnection applications, e.g., point-to-point interconnection of ATM switches and wireless access points at data rates up to 155 Mbit/s. Type 4 is intended for operation in the 17 GHz band allocated to HIPERLANs. Highly directional antennas will allow high frequency re-use as well as wideband, high speed operation.

Different HIPERLAN types and their properties are briefly summarised in Figure 5.

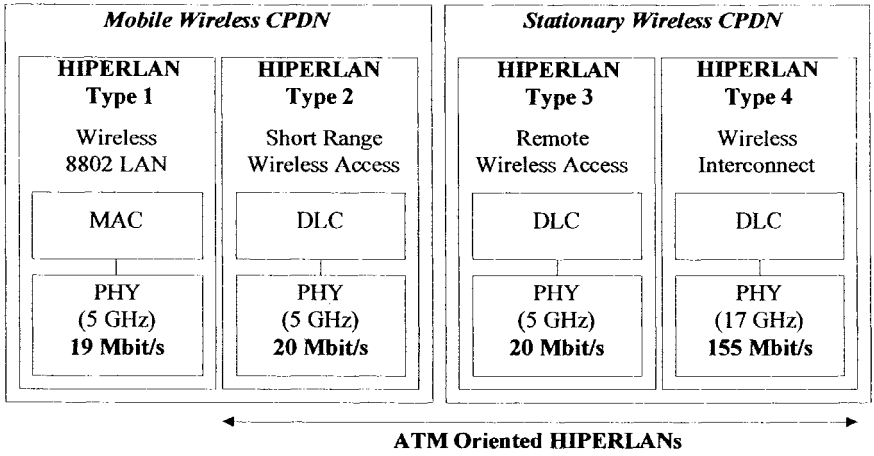

Figure 5. *Different HIPERLAN Types.*

The wireless ATM air interface work will include Radio PHY, MAC, Data Link Control (DLC) for error control and a wireless control protocol for radio resource management. Special wireless functionality is needed at Medium Access Control (MAC) and Physical layer (PHY) level to cope with the QoS requirements of ATM. The ATM Forum is expected to establish liaisons with the RES10 which has good knowledge of and experience with the high speed radio technology and standards.

The work related to mobility support in ATM networks will cover hand-off control (signalling/NNI extensions, etc.), location management for mobile terminals, Internet Protocol (IP) over ATM with mobility, routing considerations for mobile connections and Connection Admission Control (CAC) & QoS control for mobile connections. These items are being considered by the ATM Forum.

6.4 Reference Model

The overall system reference model corresponding to the WATM model presented in Figure 4 can be observed from the Figure 6.

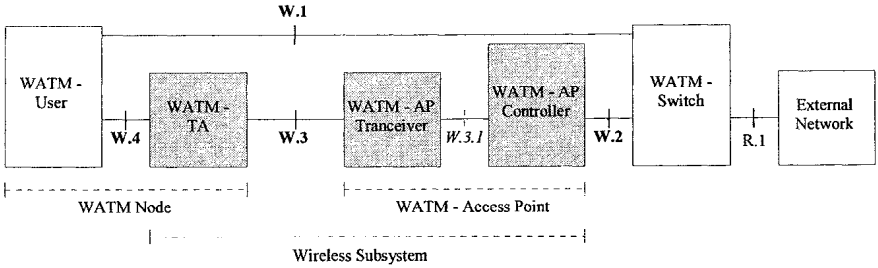

Figure 6. *WATM Reference model.*

The WATM-User includes functions for enabling and disabling of the WATM-Terminal Adapter (TA) and the selection of the wireless network to connect to.

The WATM-TA comprises functions for the use as well as the control of the wireless link. In case the WATM Node is mobile, the WATM-TA communicates with the WATM-Access Point (AP) (or optionally WATM-Switch, depending on the division of functionality between the functional elements) using a protocol that supports mobility of the Node within the environment controlled by the WATM-Switch.

Further, the WATM-TA comprises the low level functions for establishing and using connections at radio subsystem level between the WATM User Node and the WATM-AP.

The WATM-AP comprises low level functions for establishing, changing and using connections between the WATM Node and the WATM-AP as well as functions for the monitoring of the RF channel loading, for the hand-over of connections between Access Points, as well as network management functions for configuration management and other management functions.

The WATM Reference Model identifies following reference points and corresponding interfaces that are subject to standardisation.

Table 5. *Interface descriptions.*

Referenc e point	Interface description	Comments
W.1	Logical interface between the WATM Switch and WATM User or between WATM User and WATM Access Point Controller.	Includes a protocol for user registration and connection set-up and -releasing protocol as well as mobility support including handover.

W.2	Interface between the access point and the ATM Switch and its management and control functions.	A control plane interface describing the interactions between the AP and the ATM Switch for the establishment and releasing of connections, for connection handover between APs, and for capacity management purposes.
W.3	(Air) interface between the WATM-Access Point and the WATM Terminal Adapter.	Includes protocol supporting transparent ATM transport (User, Control and Management plane traffic) as well as mobility support functions such as access point acquisition and association or system functions such as channel occupancy signalling.
W.3.1	Internal interface of the WATM Access Point.	This interface is not described in this document. It is a proprietary interface.
W.4	Interface between the WATM User and the WATM Terminal Adapter.	It is defined in terms of abstract services and parameters for the user, control and management planes.

6.5 Functional Architecture

Based on the presented reference model different implementations and interoperability configurations are envisaged. In addition, these reference configurations consider both stationary and mobility specific implementations.

The main objective of the functional architecture is to define a flexible system with a wireless specific elements separated from the standard wired specific connection control elements. The separation should allow standard Connection Control (CC) functions, such as Q.2931 or UNI3.1/4.0, to exist as such in the higher protocol layers while they only have a clearly defined interface for the wireless sub-system. It should be noted that some of the mobility and authentication related information will still exist in the higher protocol layers when these are seen as a terminal level of control actions having a counter part in the network side. Figure 7 gives an overview to the envisaged WATM protocol architecture and to required interfaces.

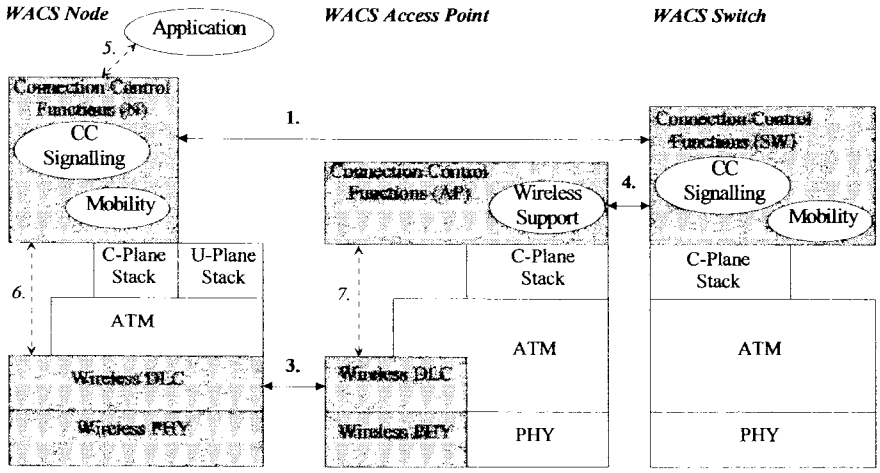

Figure 7. *WATM Protocol architecture and interfaces*

7. Conclusions

In this paper an overview to the wireless multimedia services and their characteristics was given. A phased introduction of wireless multimedia services was described. Phases will be mainly based on the development of multimedia capabilities in the wireless systems, user behaviour and development of terminal implementation technologies. The characteristics of these multimedia services set stringent requirements for the underlying transmission media. However, the wide variety of multimedia applications and services makes the definition of service requirements difficult. Throughput, as well as other requirements, can vary from one extreme to another. Fulfilment of these requirements is especially challenging in a wireless environment. Recently started activities for enhancing the capabilities of existing mobile communications systems towards multimedia will make first multimedia services available (phase 1). Also new systems for mobile communications have been developed. These so called 3rd generation mobile systems, such as UMTS and FPLMTS, are targeted for provision of multimedia services (phase 2). Wireless ATM was discussed as a most potential technology to meet these challenges of high quality multimedia and therefore enabling the introduction of new multimedia services (phase 3). Finally a brief technical overview to the wireless ATM was given. Furthermore some issues that have crucial meaning to the real (commercial) success of the concept, such as the

progress of the standardisation work, frequency allocation and frequency etiquette issues and other regulatory issues were discussed.

References

[1] J. Nieweglowski and T. Leskinen, *"Video in Mobile Networks"*, Proceedings of the European Conference on Multimedia Applications, Services and Techniques ECMAST 96, May 1996.

[2] J. Kruys and M. Niemi, *"An Overview of Wireless ATM Standardisation"*, Proceedings of the ACTS Mobile Communications Summit, Granada, Spain, November 1996.

[3] T. Leskinen, *"Wireless Multimedia Services - Developing Towards 3rd Generation Service Characteristics and Requirements"*, Proceedings of the European Conference on Multimedia Applications, Services and Techniques, ECMAST 96, May 1996.

[4] ETSI Draft ETR, *"Radio Equipment and Systems (RES); HIPERLANs; Requirements and Architectures for Wireless ATM Access and Interconnection"*, February 1997.

[5] ITU-T Recommendation I.211, *"ISDN service capabilities- B-ISDN service aspects"*, 1994.

[6] L. Nenonen and J. Mikkonen, *"Wireless ATM MAC performance evaluation, case study: HIPERLAN vs. Modified MDR"*, Proceedings of the MOBICOM96, November 1996.

[7] ETS 300 328, *"Radio Equipment and Systems (RES). Wideband data transmission systems. Technical characteristics and test conditions for data transmission equipment operating in the 2,4 GHz band and using spread spectrum modulation techniques"*, European Telecommunications Standards Institute, February 1996.

[8] FCC, *"Code of federal regulations"*, Parts 0 to 19, §15.245, October 1994.

[9] European Radiocommunications Office, "UMTS", September 1996.

[10] Naohisa Ohta, *"Packet Video, Modelling and Signal Processing"*, Artech House, 1994.

[11] V. Bhaskaran, K. Konstantinides, "Image and Video Compression Standards", Kluwer Academic Publishers, 1995.

[12] The ATM Forum, Technical Committee, *"Traffic Management Specification, Version 4.0"*, af-tm-0056.000, April 1996.

This page intentionally left blank.

DESIGN AND PERFORMANCE OF RADIO ACCESS PROTOCOLS IN WATMNET, A PROTOTYPE WIRELESS ATM NETWORK

Parthasarathy Narasimhan, Subir K. Biswas,
Cesar A. Johnston, Robert J. Siracusa,
and Heechang Kim

C & C Research Laboratories,
NEC USA, Inc.
4 Independence Way,
Princeton, NJ 08540, USA
{partha,skb,cesar,rjs,hckim}@ccrl.nj.nec.com

Abstract: In this paper, we provide an experimental view of radio access protocol design in wireless ATM networks. A proof-of-concept prototype, *WATMnet*, was recently developed at NEC USA C & C Research Labs, Princeton, NJ. *WATMnet* operates in the 2.4 GHz ISM band at a channel rate of 8 Mbps and provides an environment supporting video, audio, and data applications via a standard ATM API with both packet mode (ABR, UBR) and stream mode (CBR, VBR) services. The prototype consists of notebook computers (NEC Versa) and base stations both equipped with wireless Network Interface Cards (WNIC). This prototype system is used to discuss the design of proposed Medium Access Control (MAC), and Data Link Control (DLC) protocols. Experimental results from the performance evaluation of these protocol layers, and their impact on transport layer services in the prototype system are presented.

INTRODUCTION

Wireless ATM, as proposed in [Raychaudhuri and Wilson, 1992, Raychaudhuri and Wilson, 1994], provides a framework for next-generation wireless communication networks supporting Quality-of-Service (QoS) based multimedia services. One of the key features in wireless ATM is to provide seamless support of qualitatively similar multimedia services on both fixed and mobile terminals. In recent years, wireless ATM has become an active topic of research and development in many organizations worldwide [Eng et al., 1995, Porter and Hopper, 1995, Hyden et al., 1995, Umehira et al., 1995] and is now under standardization within applicable bodies such as the ATM Forum [Raychaudhuri et al., 1996].

Figure 1.1 shows a simplified version of a formal reference architecture for wireless ATM, now under consideration within the ATM forum [Raychaudhuri et al., 1996]. To support the above reference architecture, the standard ATM transport protocol stack is augmented with radio access channels and mobility related functions. This enables the support of ATM network layer and control plane services on an end-to-end basis.

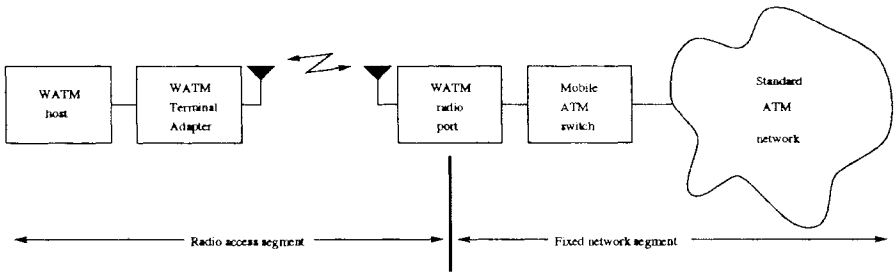

Figure 1.1 Wireless ATM reference architecture

The requirements for wireless ATM specifications can be clearly separated into two categories, namely the radio access protocol layers to handle the wireless channel specific functions, and the mobile ATM protocol extensions for mobility management functions to handle personal terminal mobility [Acharya et al., 1996]. The radio access protocol layers consist of

- a radio physical layer, capable of high-speed physical level transmission and reception,
- a Medium Access Control (MAC) layer for efficient sharing of the available bandwidth among multiple users, along with QoS management,
- a Data Link Control (DLC) layer to overcome radio channel impairments, and

- a wireless control layer for radio resource management.

In this paper, we describe the design and implementation of the radio access protocol layers using our proof-of-concept prototype wireless ATM system, *WATMnet* [Raychaudhuri et al., 1997], as a platform for experimental evaluation. We present the performance, based on measurements obtained from the *WATMnet* prototype, of the radio access protocol layers and their impact on the performance of the transport layer services. For example, the measurements show that the presence of the DLC layer significantly improves the performance of TCP applications. This is because cell level error recovery means retransmission of just lost cells rather than whole TCP segments, which is the case if the errors are recovered by the TCP layer.

This paper is organized as follows. The key issues in the design of the radio access protocol layers are discussed in Section 1.1. Section 1.5 describes the implementation of the radio access protocol layers in the *WATMnet* prototype and includes a description of the hardware and the software architectures of the prototype. Results of performance evaluation of the MAC, DLC, and the transport layer protocols is presented in Section 1.7. Finally, conclusions are provided in Section 1.1.

RADIO ACCESS PROTOCOLS

The basic idea in wireless ATM is to provide support for ATM virtual connections (VC) with QoS control on an end-to-end basis. Network level functions are handled with standard ATM cells which are augmented with a wireless-specific header/trailer on the radio link to support wireless-specific protocols such as MAC, DLC, and wireless control. Standard ATM signaling functions are terminated at the mobile terminal. Extensions to the ATM signaling protocols have been proposed to handle terminal mobility related functions such as handoff and location management [Acharya et al., 1996]. Thus, wireless ATM network specifications can be partitioned into two categories:

- Radio access protocols to handle wireless channel specific functions, and
- Mobile ATM for radio independent, mobility management functions.

The wireless ATM protocol stack corresponding to the reference architecture (Figure 1 .1) are shown in Figure 1.2. In this paper, we focus on the design and performance evaluation of the radio access protocols. For a discussion of the mobility management functions, the reader is referred to [Acharya et al., 1996, Veeraraghavan et al., 1997, Akyol and Cox, 1997].

Radio Physical Layer

Wireless ATM requires a high-speed radio modem capable of providing reasonably reliable transmission in microcells and picocells with estimated cell radius in the

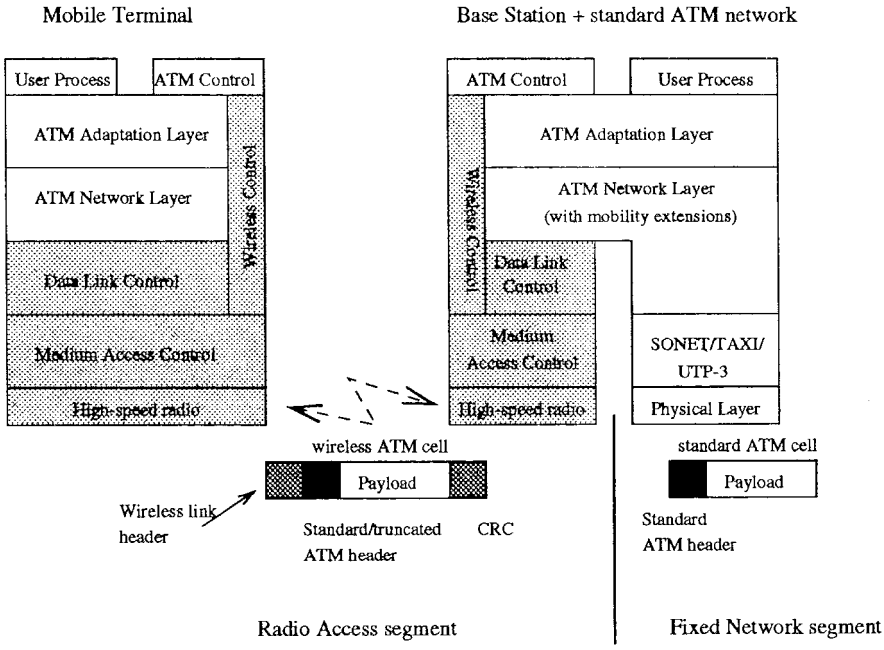

Figure 1.2 Wireless ATM protocol stack

range of 100-500m. The wireless ATM systems may operate in various frequency bands depending on national and international regulatory policies. They are usually associated with the recently allocated 5 GHz U-NII band in the US and the HIPERLAN band in Europe. Operation in higher frequency bands, for example 20-30 GHz or 60 GHz, may also become viable in the future. The expected operating frequency range is of the order of 20-25 MHz. Typical target bit rates for the radio physical layer are around 25 Mbps, with a goal of supporting per-VC service bit rates of around 6 Mbps sustained and 10 Mbps peak. Also, the modem must be able to support burst operation with relatively short preambles consistent with transmission of short control packets and ATM cells. In this paper, we assume that a reasonable radio physical layer would be available and concentrate on the design of the other radio access protocol layers.

Medium Access Control

A MAC layer is required to share the available bandwidth efficiently among multiple users. It must provide support for standard ATM services including Constant bit rate (CBR), variable bit rate (VBR), available bit rate (ABR), and unspecified bit rate (UBR) traffic service classes along with their associated QoS requirements. An important requirement for a suitable MAC protocol is that it should support these services (along with QoS requirements) while maintaining high utilization of the radio link.

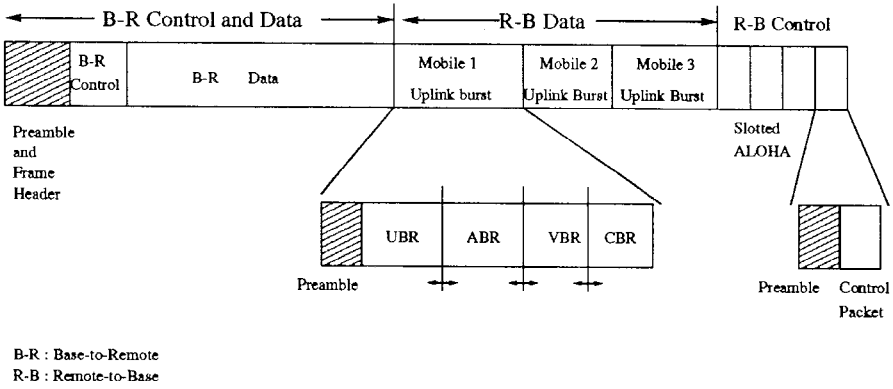

Figure 1.3 TDMA/TDD frame format for use in wireless ATM

A dynamic TDMA/TDD protocol with centralized control was proposed in [Raychaudhuri and Wilson, 1994] as a radio access protocol for wireless ATM (Figure 1.3). Downlink information from the base station (also referred to as *radio port or access point* in the literature), including control information and ATM cells, are multiplexed into a single burst and transmitted at the start of the TDMA frame (following the preamble and frame header). The base station controls the uplink bandwidth (one or more time slots, in this case) allocation for ATM cells to each mobile, taking into account the number and type of active connections and their bandwidth requirements at that mobile. Even though the base station allocates bandwidth on a per-VC basis, it is desirable that slots allocated to VCs from a single mobile, in the uplink direction, be grouped together so that the mobile terminal can transmit it in one burst. This reduces the overhead due to uplink preambles. Uplink control information, which includes bandwidth allocation requests, uplink wireless control packets, mobility-related control packets, etc., is sent in a slotted ALOHA contention mode.

There are two necessary components to a MAC protocol suitable for wireless ATM – a supervisory MAC (S-MAC) and a core MAC (C-MAC). The S-MAC, as its name implies, performs higher-level, supervisory functions, while the C-MAC performs the lower-level functions like multiplexing and demultiplexing data from the buffers associated with active VCs according to the allocation schedule determined by the S-MAC.

Since the bandwidth allocation is centralized, the S-MAC at the base station receives requests for bandwidth allocation from the mobile terminals for uplink transmissions and from the DLC for downlink transmission. The total bandwidth is then partitioned, taking into account QoS requirements, among the active connections. The S-MAC prepares a *schedule table* for each frame containing slot allocations for each VC. This information is used by the C-MAC at the base station and is also transmitted in the downlink control slots for the benefit of the C-MAC at the mobile terminals. The S-MAC at the base station also participates in the connection admission control (CAC) functions for its coverage area. The S-MAC at the mobile terminals receive requests for bandwidth allocation from the DLC and is responsible for relaying these requests to the S-MAC at the base station. Also, the S-MAC at the mobile terminal processes the schedule table that is broadcasted at the start of each TDMA frame and prepares it for use by the C-MAC during that frame. Some of the radio resource management functions are also handled by the S-MAC.

The C-MAC is responsible for multiplexing data from different VCs (the associated data buffers) for transmission according to the schedule table supplied by the S-MAC. On the receiving side, it filters the received data and notifies the DLC of the appropriate VC of any received data. The wireless control messages and bandwidth allocation requests are passed to the S-MAC. The functions of the C-MAC make it a suitable candidate for implementation in hardware. These functions are also more or less identical at both the base station and remote terminal.

CBR VCs are allocated slots periodically according to their bit rate. CBR bit rates are allocated in integral multiples of 32 Kbps and to ensure support of bit rates that require less than one slot per TDMA frame, a group of 6 frames (12 ms) is considered as a MAC superframe. The CBR slot allocation algorithm operates over the CBR slots in a MAC superframe, which produces a recurring CBR slot assignment from one superframe to the next. We propose that CBR slot assignments be maintained relatively static since these assignments are periodic and, hence, do not have to be broadcasted periodically to the mobile terminals. The other key motivation is to enable support for low-complexity telephony devices in wireless ATM networks. An adaptive slot allocation scheme [Biswas et al., pear] that tracks the short and long term bit rate variation of each individual VC is considered for VBR VCs. Based on the instantaneous rate and MAC buffer state information, the slot requirement for each VC for the next MAC frame is determined. The S-MAC at the base station, then, uses a User Parameter Control (UPC) based proration rule for deriving appropriate allocation

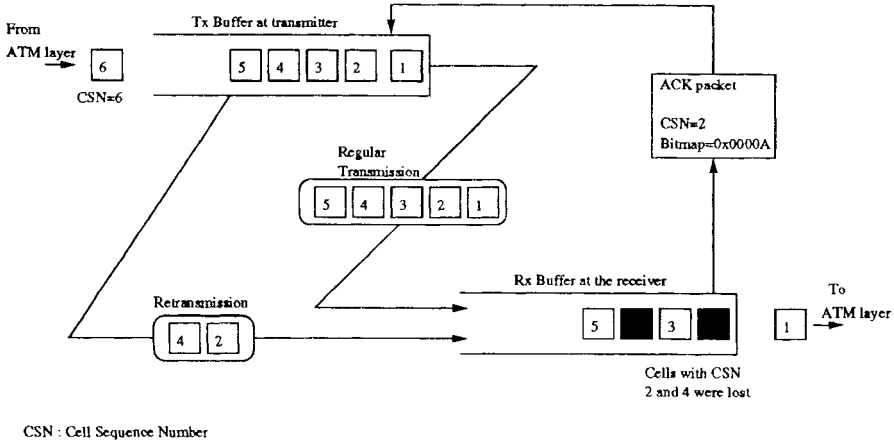

Figure 1.4 Logical representation of the Data Link Control layer

for each frame. UBR VCs are assigned slots on a burst-by-burst basis with dynamic allocation of UBR slots and unused CBR/VBR slots in each frame. A round-robin slot scheme for allocating slots among many UBR VCs is proposed to avoid one long UBR burst from delaying other shorter ones. UBR slot allocation is demand-based and in a best-effort mode, in contrast to CBR slot allocation which is periodic and guaranteed. Also, bandwidth for retransmissions of lost cells for CBR and VBR VCs are allocated in the UBR mode. A priority scheme for these retransmissions may be considered under heavy loads to ensure that the error recovery can be completed within the allowed error recovery window. The slot assignment for ABR connections is currently under study and is beyond the scope of this paper.

Data Link Control

This section presents a design outline of the data-link control (DLC) protocol [Xie et al., 1995] which was designed for transporting multiclass traffic across the radio physical links in a wireless ATM network. The primary motivation of developing such a protocol was to reduce cell error rate and to provide sequential cell delivery (over wireless), which are two major lower layer protocol requirements within the ATM framework. In addition, the DLC also provides the interface between MAC and higher layer protocols needed to support a demand driven medium-access strategy for multiservice ATM traffic.

A group acknowledgement scheme is used for error recovery where a receiver sends eight octet acknowledgement (ACK) packets, each with a bitmap vector, indicating the error status of certain number of WATM cells (Figure 1.4). Upon reception of an acknowledgement packet, the transmitter executes a selective retransmission algorithm for recovering the erroneously transmitted cells. Note that the DLC execution takes place in a per-VC mode and that requires the wireless nodes (both base and terminal) to store separate DLC state information for each individual VC.

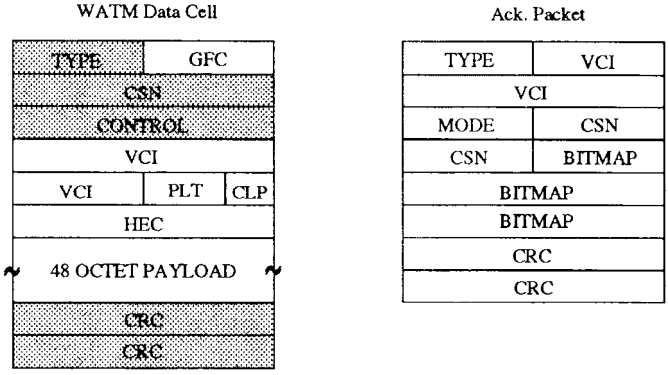

Figure 1.5 WATM data cell and acknowledgement packet format

Since this group acknowledgement and selective retransmission scheme requires a notion of Cell Sequence Number (CSN), which is otherwise absent in fixed ATM cell header, a data-link layer CSN is added into WATM cell header. Cell Sequence Number and other additional fields of WATM data cells and acknowledgement packets are illustrated in Figure 1.5. The shaded regions in figure denote the additional header and tailer information carried by wireless ATM cells. Both data cells and acknowledgement packets carry 16-bit CRC trailers for transmission error detection.

Our current DLC design supports both UBR and CBR services, and will be extended to include VBR [Biswas et al., pear] and ABR traffic classes. A demand driven medium access strategy is used for UBR where a transmit DLC needs to send medium access requests to the MAC layer before the DLC layer can actually send cells to the MAC. At the transmission side, when a burst of ATM cells arrives at DLC, an access allocation request is sent to the S-MAC and upon reception of a reply, the transmitter DLC sends the cells to the C-MAC layer for transmission. The receiver DLC sends a group acknowledgement back after receiving the whole burst and the acknowledgement information is used at the transmitter to retransmit the erroneously transmitted cells. Since a zero cell loss data-link is desired for UBR traffic, no specific time limit for error recovery is imposed. Retransmission requires the transmitter DLC to buffer cells

until a group acknowledgement indicating proper reception at the other end is received. Also note that the receiver needs to buffer the WATM cells in order to ensure sequential delivery to the ATM layer.

In the constant bit rate (CBR) mode, the DLC attempts to correct cell errors within a fixed time window, specified at the connection setup time. During call setup, the S-MAC allocates slots on a periodic basis for the CBR VCs depending on their bit-rate requirements. While this periodic allocation is used for CBR cell transmissions, additional on-demand bandwidth (supported by the MAC's UBR mode) is used for retransmission of the lost cells. Note that, unlike in the UBR mode, the constant bit rate data-link control maintains a constant end-to-end delay over the wireless link. Although the cells may arrive periodically at the transmit side of the DLC, at the receiving side some cells will be lost due to the channel error. This causes cell delay variation and out of sequence delivery over the wireless link. In order to alleviate this we use fixed length buffers both at the transmit and the receiver ends. At the receiver side, the cells are clocked out of the buffer and sent to the ATM layer at the specified data-rate. In addition to cell de-jittering, the receiver side buffer also helps to maintain a fixed recovery time window for each lost cell. The size of the buffer is determined by the bit-rate of the call and the allowable time window for error recovery. The transmit side also maintains a FIFO which is used for buffering a transmitted cell until its recovery time elapses.

Wireless Control

The wireless control sublayer is needed for support of control plane functions at the radio access layer and its integration with the ATM network. Its functions include radio resource management, and metasignaling capabilities needed for terminal migrations and handoff control. The reader is referred to [Yuan et al., 1996] for a more detailed discussion.

WATMnet PROTOTYPE IMPLEMENTATION

In this section, a brief description of the hardware/software configuration of the *WATMnet*, the wireless ATM network prototype is presented, followed by the discussion of the implementation of the radio access protocol layers in the prototype.

Hardware/Software Configuration

The *WATMnet* prototype system consisted of two identically configured base stations supporting adjacent microcells, notebook computers for portable terminals, and a mobility enhanced ATM switch connecting the rest of the ATM network to the two base stations. Each base station is a VME chassis containing standard off-the-shelf components comprising one i960 processor card with supporting RAM, one ATM

interface card, and one VME-to-PCMCIA adapter which connects to the *WATMnet* Network Interface Card (WNIC). The i960 runs a real-time embedded operating system called pSOS. The mobile terminals were NEC Versa notebook computers running the Linux operating system (kernel version 2.0.29). The WNIC plugs into the mobile terminal's PCMCIA interface.

At the base station, a software driver library has been defined to control and transfer information to the *WATMnet* network interface card, and is fully dependent on the NIC's hardware architecture. The *WATMnet* custom NIC supports multiple transmit priority queues, each assigned a traffic class such as Control, CBR, VBR, ABR, and UBR. Four receive queues are supported by the hardware, each having a circular buffer of configurable size. The WNIC driver interface library offers C callable functions for NIC configuration, buffer management, and for queue management in both the transmit and receive direction. All data flow between the base station host and the wireless NIC hardware is thorough a VME/PCMCIA interface.

Figure 1.6 shows the key components of the *WATMnet* support software running at the mobile terminal. The WNIC device driver at the mobile terminal presents two interfaces to the higher layers. One is a network interface that is exported to the IP layer and appears as just another network device to the system. To enable support for native ATM applications, a second network interface that supports AAL0 and AAL5 is provided. Upon receiving an IP packet for a new destination host, the ATM network layer uses the ARP module to resolve the destination IP address to an ATM address (if the mapping from IP address to the ATM address is not already known). Then the signaling module sends a connection request to the destination host and once the ATM connection is established, the queued IP packets are sent over that VC. Thus, one VC is used for each new destination IP address. These connections are closed after a designated period of inactivity.

The data received at the WNIC over the wireless link is demultiplexed by the core MAC on the WNIC. Once the data is validated, the core MAC generates an interrupt to the host processor. The data is read into the host in the interrupt service routine (ISR) at the host and is processed outside of the ISR.

Medium Access Control

As mentioned earlier, there are two components to the MAC layer – the C-MAC and the S-MAC. The C-MAC is implemented in hardware in the WNIC, while the S-MAC is implemented as a software module in the WNIC device driver.

Figure 1.7 presents the key components of the C-MAC as it is implemented in the WNIC [Johnston, 1996]. A NEC v53 microprocessor is used as the embedded controller to implement the *WATMnet* NIC system initialization and to perform partial WATM Cell Queue Management in firmware. WATM Cell Queue management is implemented in the WNIC to partition the local memory in different size queues, which

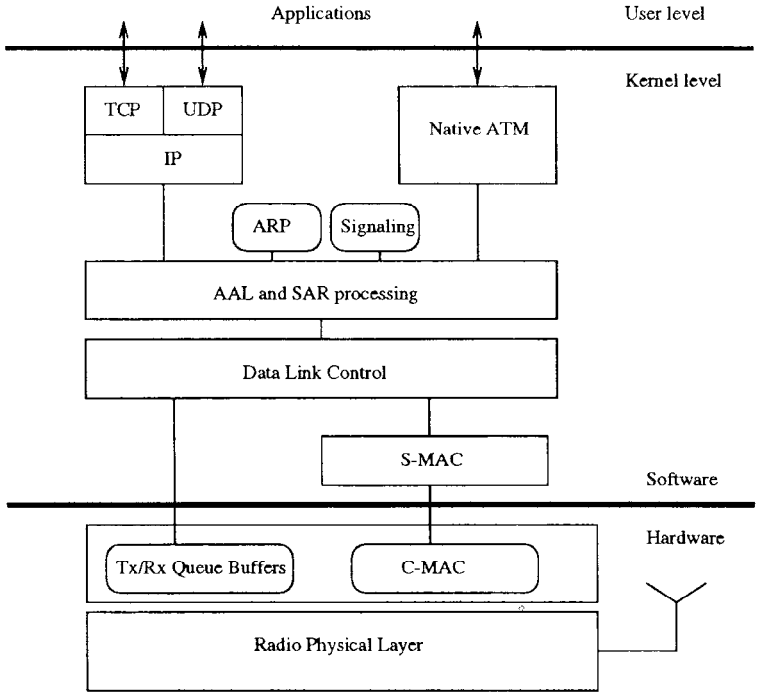

Figure 1.6 Mobile terminal software architecture

are allocated to each active connection. At the hardware, the Memory Pointer Processor module complements the v53 firmware to support WATM Cell Queue Management. The Memory Pointer Processor relieves the v53 from data movement functions. A hardware TDMA/TDD Framer controls the transmission of WATM cells into the proper TDMA slot. The requirement of the WATM layer to generate CRC-16 overhead on a per WATM cell is supported in hardware at the Cell Formatter/Deformatter block.

In the transmit direction, PDUs are processed by the ATM and SAR processing layer and cells are handed to the DLC. Once the wireless ATM header/trailer information has been formatted, the DLC, upon instructions from the S-MAC on slot allocations, moves the data to the appropriate transmit queue on the WNIC.

In the receive direction, downlink frames are synchronized and delineated at the Mobile's TDMA/TDD Framer. When a WATM control or data cell arrives, the v53 is informed, which, then, instructs the Memory Pointer Processor by indicating the Memory locations where incoming WATM cells are to be stored in a FIFO manner.

The receiver supports receive FIFO queues, one for each traffic class: i.e. Control, ABR, UBR, VBR, and CBR. While WATM cells are stored in the Memory block CRC-16 checking is performed at the Cell Formatter/Deformatter block. In the Cell Formatter/Deformatter block, a CRC syndrome is attached to each cell to be interpreted at the Host WATM Software block. Flexibility for dropping or processing cells with errors is handled in software by checking the CRC syndrome. Once the received WATM cells are stored in their respective queues, a hardware interrupt is generated to the host processor indicating the availability of information. The host implements queue control software to determine priority-based movement of data into the host memory.

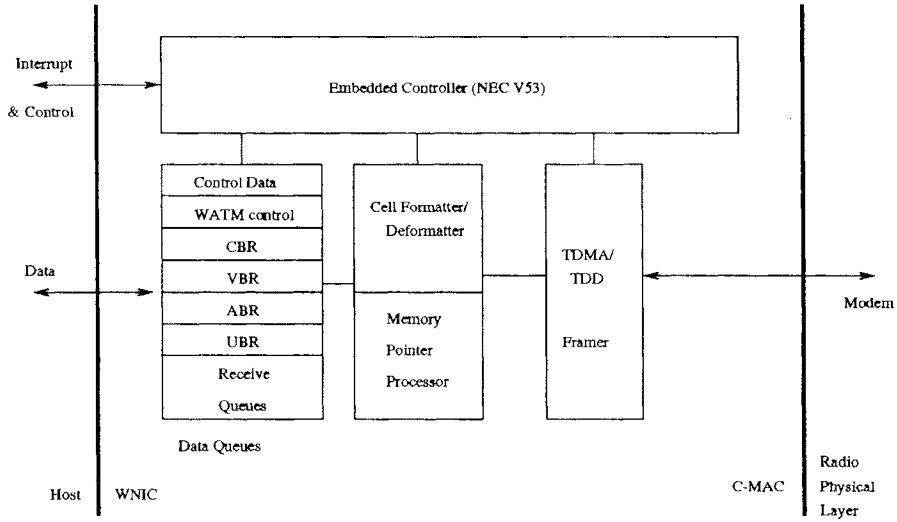

Figure 1.7 Implementation of C-MAC in the WNIC

The S-MAC is implemented in software on the host side. It maintains information on the current active connections, their bandwidth requirements, and their QoS control information. A bandwidth allocation policy based on satisfying QoS requirements is implemented, which generates a schedule table containing information on which cells are to be transmitted in which slots in the current frame. This table is passed to the C-MAC which then performs the actual act of multiplexing and formatting data so that it is ready to be processed by the modem for transmission over the radio channel.

Data Link Control

Two different implementations of the data-link protocol were necessary for the present system. At the base station (which runs a real-time OS, known as pSOS, on an i960 processor) a separate pair of transmission and reception tasks are used for the DLC processing of each virtual circuit. All the data-link tasks communicate with a fixed ATM cell relay task and a MAC task using a fairly straightforward set of inter-task communication primitives (messages queue and shared memory). Both the transmission and reception DLC tasks for a VC are created during its setup time within the base station. This per-VC DLC task allocation scheme aids the implementation of quality of service (QoS) by specifying the allocated buffer space and scheduling priority of a DLC task based on the QoS requirement of the corresponding VC. Since all these tasks run in a single address space in pSOS, the expense of frequent context switching is small enough to be outweighed by the advantages of QoS-based scheduling.

Linux, a multitasking operating system, is used at the personal terminal side where the data-link and S-MAC software are integrated within the kernel as a part of a dynamically reloadable device driver. Since Linux kernel does not support multiple tasks within the kernel, a single threaded version of the same DLC protocol is implemented where the communication between MAC and DLC is realized as a function call interface.

PERFORMANCE EVALUATION

In this section, some performance evaluation results from measurements on the *WATMnet* prototype are being presented. The objective of this study was to evaluate the performance of the radio access layers and their impact on the performance of the transport layer protocols. We have observed that a combination of video and audio multimedia applications running in the CBR mode can achieve a MAC layer throughput of around 4.4 Mbps, which is roughly a utilization of 0.75 (after accounting for the physical layer, forward error correction (FEC), and control channel overheads).

The DLC for CBR VCs operates with a fixed size buffer for error recovery and jitter removal purposes [Xie et al., 1995]. We studied the performance of the DLC in terms of its ability to recover lost cells as a function of the size of this buffer. The trade-offs here are that a longer buffer increases the ability of the DLC to recover lost cells, but adds to the end-to-end delay on the connection. A shorter delay limits the ability of the DLC to recover lost cells and hence increases the cell loss rate (CLR) faced by the connection. Figure 1.8 presents the CLR faced by the connection as a function of the DLC cell buffering time. The observed bit error rate (BER) was around $3e-5$, derived from an observed cell loss rate of 0.014. When the CBR load was high, a short DLC buffering time for the CBR cell resulted in almost no recovery. A higher rate of the CBR calls resulted in a lower bandwidth for the retransmission. Hence, the ability of the system to recover lost cells decreased rapidly with a short lifetime of the CBR cell.

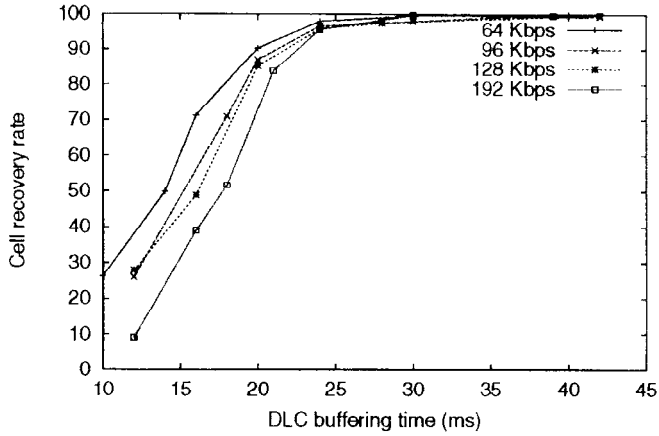

Figure 1.8 CBR cell loss as a function of the DLC buffering time

Figure 1.9 shows the net ftp throughput as a function of the maximum available UBR bandwidth, which is a function of the number of CBR and VBR VCs present in the system and also on the total bandwidth requirements of these VCs. A file of size 5984960 bytes was downloaded into the mobile terminal using ftp. The net ftp throughput was then averaged over 10 samples for each value of the maximum UBR bandwidth. There are two curves, one with the DLC layer activated and the other without the DLC layer in the wireless ATM protocol stack. This plot builds a strong case for including the DLC layer in the wireless ATM protocol stack. The presence of the DLC layer results in the doubling of the achievable net ftp throughput (as compared to without the DLC). This can be explained by the fact that cell level error recovery (at the DLC layer) is better than packet level error recovery (at the TCP layer). Also, error recovery at the DLC reduces relatively the amount of traffic that flows through the end-to-end TCP connection (this is shown in greater detail in Table 1.1). Also, we noticed that the sample variance of the download times was higher without the DLC than with the presence of the DLC. This can be explained by the fact that the amount of retransmitted data (by the TCP layer when the DLC is deactivated) depends on the position of the error and the current size of the TCP window. Even though the error recovery times at the DLC could interact with the TCP timers and cause the TCP window to shrink, this seems to happen less frequently and thus does not cause a lot of variation in the download times.

In Table 1.1, we present measurements to show the effect of the presence of the DLC in the wireless ATM protocol stack. The first set of numbers show the numbers

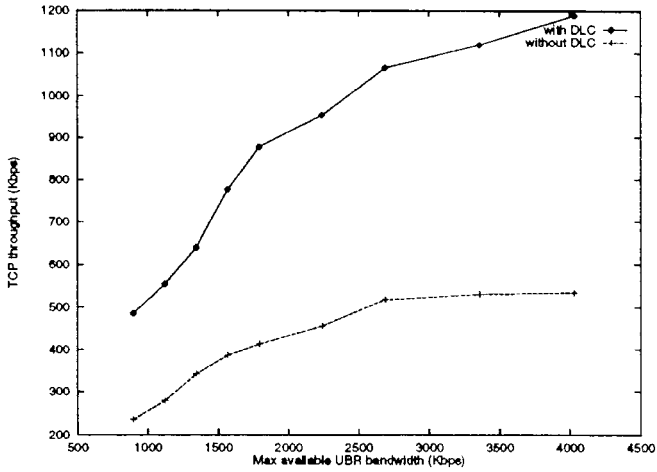

Figure 1.9 FTP throughput as a function of the maximum available UBR bandwidth

measured with the DLC layer activated. For the second set of measurements, the DLC was deactivated to study the effect the DLC has on TCP performance. A file of size 2559251 bytes was downloaded using ftp. With the presence of the DLC the download time was 27.15 seconds yielding a net ftp throughput of 754 Kbps. Without the DLC, the same file was downloaded in 59.87 seconds yielding a net ftp throughput of 342 Kbps. The offered loads at the MAC and the physical layer are also given in the table.

There are two costs of having the DLC in the protocol stack. One is the cost of transmitting acknowledgments (ACKs) in both directions to aid in error recovery of lost cells. The other cost (which can be reduced by better software design) is due to the phenomenon of retransmission of cells that were received right the first time thus resulting in waste of bandwidth. This occurs if the timers that the DLC uses to transmit ACKs are very aggressive. If the ACK transmission intervals are set longer, then the delays in error recovery interact with the timers at the TCP layer resulting in the TCP layer retransmitting packets that would have been recovered at the DLC layer.

But as the results in Table 1.1 show, the additional burden of carrying the ACKs and the duplicate transmissions is more than compensated for by the gain in the TCP throughput achieved by the presence of the DLC. Also, the amount of information transmitted over the wireless link due to input from the TCP layer is lower with the presence of the DLC. This shows that error recovery at the DLC layer is always better than error recovery at the TCP layer.

Maximum Available UBR b/w	1568 Kbps
File size (bytes)	2559251
Observed cell loss rate	0.014
observed BER	$3x10^{-5}$

Performance measure	with DLC	without DLC
Download time (sec)	27.15	59.87
ftp throughput (Kbps)	754	342
Offered Load (Kbps):		
MAC	1023	463
Physical	1263	572
ACK b/w usage (Kbps)	31.5	0
MAC statistics:		
Transmitted bytes[1]	3471012	3469488
ACK bytes[2]	109092	0
Transmitted bytes due to duplication by DLC	43624	0
Transmitted bytes due to input from TCP[3]	3318296	3469488

Table 1.1 Comparison of TCP performance with and without the DLC

CONCLUDING REMARKS

In this paper, we have presented the key issues in the design and implementation of the radio access protocols for a wireless ATM network. We have used *WATMnet,* the wireless ATM network prototype developed at NEC USA C & C Research Labs, Princeton, NJ to present an experimental view of the design, implementation, and performance of the radio access protocols. The *WATMnet* prototype, as well as the experiences of other organizations working in this field, has established the feasibility of wireless ATM as a future mobile, multimedia networking solution. The *WATMnet* prototype has shown that current implementations of MAC and DLC protocols can provide reasonable performance at throughput levels of about 0.65-0.75 (after accounting for physical layer, FEC, and control channel overheads). The presence of the DLC in the wireless ATM protocol stack is critical for the performance of transport layer protocols (such as TCP) since error recovery at the cell level over the radio link offers higher utilization and

net data rates compared to end-to-end error recovery at the transport layer, which is the case with TCP. The performance measures summarized in this paper are from the initial prototype system. We are in the process of designing the next version of the prototype (*WATMnet* 2.0), which will support a data rate of about 25 Mbps and will operate in the 5 GHz U-NII band. The experience we have gained from the work on the current prototype system is being applied in the design and optimization of the various radio access protocol layers discussed in this paper.

Notes

1. This is the total amount of data observed at the MAC layer
2. The amount of data transmitted due to ACK packets, observed at the MAC layer
3. Computed by subtracting the bytes due to acknowledgements and duplicated cells from the total transmitted bytes.

References

Acharya, A., Biswas, S. K., French, L. J., Li, J., and Raychaudhuri, D. (1996). Handoff and location management in mobile ATM networks. In *Proc. 3rd Intl Workshop on Mobile Multimedia Communications (MoMuC-3),* Princeton, NJ.

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

Biswas, S. K., Reininger, D. J., and Raychaudhuri, D. (1997 (to appear)). Bandwidth allocation for VBR video in wireless ATM links. In *Proc. ICC '97.*

Eng, K. Y., Karol, M. J., Veeraraghavan, M., Ayanoglu, E., Woodworth, C. B., Pancha, P., and Valenzuela, R. A. (1995). BAHAMA: A broadband ad-hoc woreless ATM local area network. In *Proc. ICC '95,* pages 1216–1223.

Hyden, E., Trotter, J., Krzyzanowski, P., Srivastava, M., and Agarwal, P. (1995). SWAN : An indoor wireless ATM network. In *Proc. ICUPC '95,* pages 853–857, Tokyo, Japan.

Johnston, C. A. (1995). Architecture and performance of HIPPI-ATM-SONET terminal adapters. *IEEE Communications Magazine,* 33(4).

Johnston, C. A. (1996). A network interface card for wireless ATM network. In *Proc. PIMRC '96,* Taipei, Taiwan.

Li, J. and Yuan, R. (1996). Handoff control in wireless ATM networks: An experimental study. In *Proc. ICUPC '96,* pages 387–391.

Porter, J. and Hopper, A. (1995). An overview of the ORL wireless ATM system. In *Proc IEEE ATM Workshop.*

Raychaudhuri, D., Dellaverson, L., Umehira, M., Mikkonen, J., Phipps, T., Porter, J., Lind, C., and Suzuki, H. (1996). Charter, scope and work plan for proposed Wireless ATM Working Group. In *ATM Forum/96-0530/PLEN.*

Raychaudhuri, D., French, L. J., Siracusa, R. J., Biswas, S. K., Yuan, R., Narasimhan, P., and Johnston, C. A. (1997). WATMnet : A prototype wireless ATM system for multimedia personal communication. *IEEE Journal of Selected Areas in Communication*, 15(1):83–95.

Raychaudhuri, D. and Wilson, N. D. (1992). Multimedia personal communication networks : System design issues. In *Proc. 3rd WINLAB Workshop on Third Generation Wireless Information Networks*, pages 259–288, New Brunswick, NJ. WINLAB, Rutgers University.

Raychaudhuri, D. and Wilson, N. D. (1994). ATM–based transport architecture for multiservices wireless personal communication networks. *IEEE Journal of Selected Areas in Communication*, 12(8):1401–1414.

Umehira, M., Hashimoto, A., and Matsue, H. (1995). An ATM wireless access system for thetherless multimedia services. In *Proc. ICUPC '95*, Tokyo, Japan.

Veeraraghavan, M., Karol, M. J., and Eng, K. Y. (1997). Mobility and connection management in a wireless ATM LAN. *IEEE Journal on Selected Areas in Communications*, 15(1):50–68.

Xie, H., Narasimhan, P., Yuan, R., and Raychaudhuri, D. (1995). Data link control protocols for wireless ATM access channels. In *Proc. ICUPC '95*, Tokyo, Japan.

Yuan, R., Biswas, S. K., and Raychaudhuri, D. (1996). A signaling and control architecture for mobility support in wireless ATM networks. In *Proc. ICC '96*, pages 469–477, Dallas, TX.

A DISTRIBUTED MEDIA ACCESS CONTROL FOR WIRELESS ATM ENVIRONMENTS[1]

Jean-Pierre Ebert *, Ralf Holtkamp *, Adam Wolisz *†, Louis Ramel #

* Technical University Berlin – Telecommunication Network Group
# Thomson-CSF – Communication, Navigation, Identification Division
{ebert, holtkamp, wolisz}@ee.tu-berlin.de, louis.ramel@cni.thomson.fr

Abstract: In this article we present RNET MAC - a novel MAC protocol to be used in wireless ATM (WATM) environments. The MAC protocol features a distributed control for media access. Therefore RNET MAC fits well for spontaneous network setups and frequent network configuration changes. We discuss some design options and show basic performance results of RNET MAC in different network scenarios such as fully meshed, hidden terminal and client server.

1 INTRODUCTION

Evolving WATM systems (e.g. [9, 11]) often require a media access control and data link control due to multiple access to the communication channel and the error prone medium respectively. For the sake of minimum changes of the ATM protocol functions and QoS enforcement, the MAC protocol has to provide dedicated mechanisms for bandwidth reservation, priority handling, low access delays as well as jitter. One can find several proposals of MAC protocols in the recent literature (e.g. [6, 7, 9, 11, 12, 13]), which all provide in some way QoS enforcement. However, all of them are based on the assumption that there

† also with GMD Fokus

is a dedicated entity (BS)[2] controls the access to the communication channel. Furthermore, the traffic has to be transmitted via this BS. The first assumption leads to mobiles with a small MAC functionality and therefore potentially smaller battery capacity requirements of the mobiles. On the other side, a relative expensive BS is always required and spontaneous networking events (for instance conferences or meetings), where no pre-installed infrastructure is given, are not possible. The latter assumption leads to a waste of bandwidth in case two mobiles (M) want to communicate with each other, since the data has to be transmitted twice (from M to BS and vice versa). We present in this paper RNET MAC (section II) that features a distributed MAC and point-to-point communication. In the next section (III) some performance results are given showing the capability of RNET MAC to work in ATM environments and different network scenarios. In section IV we conclude with a rough comparison to MACs with centralized control.

2 RNET MAC PROTOCOL

The basic idea of the MAC is derived from the RNET proposal (Radio Network, see [10]) as it has been proposed for HIPERLAN /2[3] systems to the ETSI[4] in 1995 by Thomson. RNET is characterized by free physically separated channels (FDMA). Two of these channels are used for signaling purposes and one is for data transmission. The signaling channels are slotted (TDMA). Data will be transmitted by an ascending ramp (see Figure 1).

2.1 RNET MAC channel structure

We consider an arbitrary amount of buffered mobiles in a wireless micro-cell. As outlined above there are two signaling (header and feedback) and one data

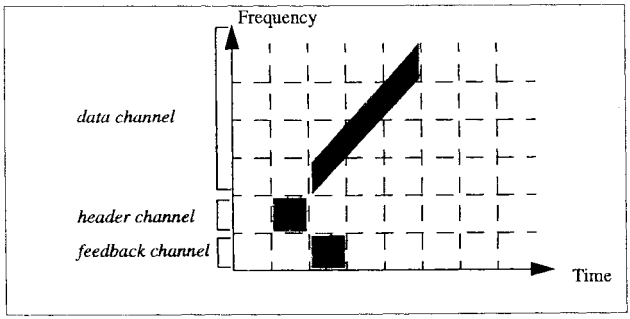

Figure 1: RNET channel structure

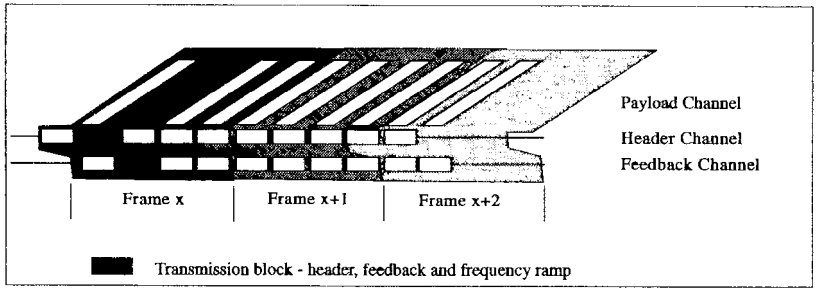

Figure 2: Channel structure with ramp spanning over three slots

(payload) channel. The payload channel consumes considerably more bandwidth than the signaling channels together. Header and feedback channels are slotted. About 56 bits fits in every slot for signaling and paging purposes. There is no fixed defined format yet. The payload channel is accessed by several ascending but independent frequency ramps at a certain point in time. One frequency ramp can carry one or more ATM cells as outlined in the following sections.

2.1.1 Framing and transmission blocks

A frame is defined as a group of consecutive header slots, associated feedback slots and payload ramps. The length of a frame is determined by the spanning width of a ramp. If a ramp spans over x (e.g. 3) slots then a frame consists of x+1 (e.g. 4) slots (see Figure 2). A header slot at time x, a feedback slot at time x+1 and a ramp, starting immediately after the header slot, are associated with each other and form a transmission block. There are guard times between the slots to encompass signal delays and clock variances.

2.1.2 Frame synchronization

Following the idea of distributed control, there is no common frame in RNET MAC. Instead, a synchronization of all mobiles with respect to frame start and frame end is not required. This simplifies the protocol (e.g. registration procedures). Rather than having a superframe to which all mobiles are synchronized, framing is locally. Every mobile maintains its own frame. Therefore, there may be offsets between the start point of frames. A local frame starts as soon as a mobile is switched on.

2.1.3 Frame size and ramp capacity

For the sake of achieving an optimal bandwidth utilization at a certain bandwidth and modulation scheme, the ascend of the ramp has to be chosen carefully. If the ascend is too low then two consecutive ramps will overlap each other. In the other case the payload channel is used only for a short time[5]. Therefore an optimum ascend m is

$$m = B_B / L_T$$

where B_B (e.g. 5 MHz)[6] is the base bandwidth and L_T is the slot time. L_T is defined as

$$L_T = L_L / T_R$$

where L_L is the slot length (e.g. 56 bit) and T_R (e.g. 5 Mbit/s) is the transmission rate. With the given values the optimum ascend m is equal to 446428.57 Hz/msec. To compute the number of slots per frame K the following formula has to be applied

$$K = 1 + R_T / (L_T + I_T)$$

where R_T is the ramp time and I_T is the intergap time (e.g. 3.8 msec). I_T is determined by signal delays and synchronization clock variances among mobiles. RT is determined by the ascend of the ramp and the bandwidth for the payload channel B_P (e.g. 190 MHz)

$$R_T = B_P / m$$

With the given values the number of slots per frame K results in 29 approximately. We can also compute the capacity of the ramp

$$C_R = R_T * T_R$$

which results in 2128 bit or about 5 ATM cells. If one considers a higher transmission rate (T_R) and/or a higher base bandwidth we will get an higher ascend of the ramp which can result in a smaller number of slots per frame and a smaller capacity of a ramp. The same is valid for a larger intergap time between the slots.

2.2 RNET MAC access control

The access control is based on a completely distributed mechanism. We assume that every mobile's MAC is associated with a buffer.

A mobile which wants to transmit data by means of radio packets must have sensed the header and feedback channel at least for one frame duration. Thus, the mobile is able to gather information about the receiving state of other mobiles and the availability of transmission blocks. The check of receiver state is necessary since a mobile can neither synchronize to two frequency ramps (see Figure 3) nor receive packets while sending.

A transmission block is considered to be available (not reserved), if the same block was not used in the last frame. The decision of availability can depend on the received header information, the feedback information or on both. This information differentiation for reservation evaluation purposes very much affects the channel efficiency in hidden terminal scenarios as discussed later.

If a transmission block is not reserved, a mobile can transmit a header on the header channel. Immediately after the header transmission follows the packet transmission on a frequency ramp (payload channel). If a mobile receives correctly a header with its own address it responds immediately with an acknowledgment in the next time slot on the feedback channel and starts packet reception. Note, the feedback is not the acknowledgment for the correct packet transmission in the payload channel. It serves as an packet sending cut-off criteria if there is no feedback. However, we assume, that a correct header reception implies the correct reception of the packet in the payload channel, because the transmission quality is considered to be stable for sufficient short time intervals.

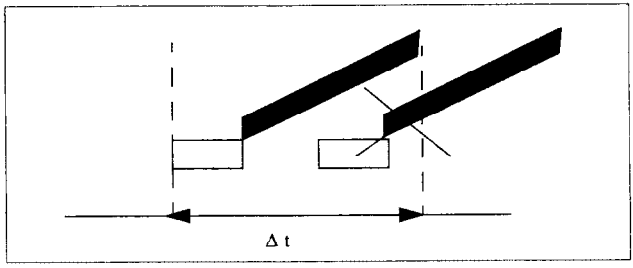

Figure 3: Single receiver station

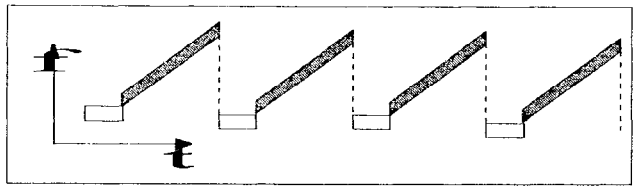

Figure 4: Temporal Channel

2.2.1 Header channels access

Header slots of non-reserved transmission blocks are accessed in a Slotted Aloha principle. In any error case (e.g. collision of header, header/feedback get lost ...) the mobile switches to a backlog state and has to wait k free slots whereby k is uniformly distributed in a certain interval. The error case is determined if the feedback is not received. The mobile will chose the first non-reserved header slot (1 persistent) for transmission.

2.2.2 Temporal channel concept

After a successful header transmission determined by the reception of the feedback a mobile has reserved automatically the same transmission block in the next frame. This extends the header channel access to an Reservation Aloha based scheme. Thus, a mobile is able to transmit constantly at a certain data rate (see Figure 4). In other words, if a mobile has gained access for one transmission block, it has contention free access to the same transmission block in all consecutive frames as long as there is data to transmit. This reservation is lost if the mobile has nothing to send and the header slot remains empty.

As an design option, an extended reservation scheme is possible. For instance, only in every second (third ...) frame the same slot is reserved. This would fit very well for low bit rate CBR traffic.

The concept of temporal channels can provide a very good bandwidth utilization in high load cases. However, an increasing load per mobiles can result in a very unfair bandwidth sharing!

2.2.3 Channel Holding Time (CHT)

Since the temporal channel concept causes unfairness under high system load conditions, a Channel Holding Time is introduced. The Channel Holding Time defines the number of packets, which may be sent consecutively by a mobile. With respect to the header channel access, the CHT defines the maximum number of transmission block reservations without disruption.

As an design option the channel holding time could be determined in a distributed dynamic fashion by every station. For instance CHT could be set with respect to QoS requirements of certain applications or due to the evaluation of network load condition. The latter could be done on the feedback channel information.

2.2.4 STOP Bit

A header slot plus the ramp in a frame can be used if no reservation exists. That is, the header slot has to be free (not used) at least in the last frame. This leads to waste of bandwidth. To eliminate this bandwidth waste, a stop or a follow bit in the header can be used to declare the next header as non-reserved. The stop/follow bit has to be switched on/off if the temporal channel holding time is over.

2.3 Access control in hidden terminal scenarios

In a scenario, where each mobile is in coverage distance of another mobile the evaluation of the slot reservation is quite simple, since all the header and feedback information is globally available. In a hidden terminal scenario it makes a difference as to which information (header, feedback or both) is taken in order to decide whether a transmission block is reserved. As known from the literature [2, 5] an evaluation based on receiver information should be preferred for the following reasons. The slot reservation evaluation is only done at the sending mobile. A receiving mobile responds in every case if the received header was correct. In case the header is not correct, as the header slot would be in use, the receiving mobile will not respond (see Figure 5 c).

2.3.1 Slot reservation evaluation based on header information

The main drawback of this method is waste of bandwidth and higher collision risk. The reason for the waste of bandwidth is, that the bandwidth is reserved around the sender. No other mobile hearing a header signal in a certain slot is allowed to use the slot although there would not be any contention at the receiver (see Figure 5 b). The higher collision risk results from the lack of information about slots already in use, if a mobile can hear the receiver but not the sender (see Figure 5 d).

2.3.2 Slot reservation evaluation based on header and feedback information

This method will reduce the problem of a higher collision risk, but the problem of wasting bandwidth still remains.

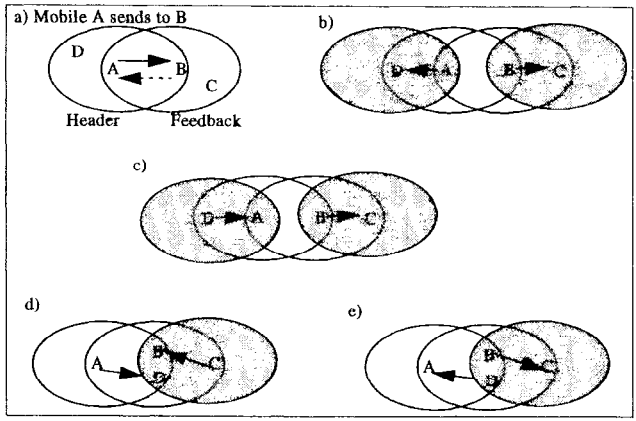

Figure 5: a) Coverage of header and feedback signals, b) c) d) hidden terminal scenarios

2.3.3 Slot reservation evaluation based on feedback information

If the feedback information is taken to evaluate the slot reservation then only a bandwidth reservation around the receiver will happen. This results in an optimal use of bandwidth since the sender does not care about contention in his coverage area if other mobiles transmit to mobiles outside its area (compare Figure 5 b, d). The problem which arises from this method is the chance of feedback contention. The sending mobile is not able to receive a correct feedback signal (see Figure 5 d, e).

Therefore, we assume the reception of any feedback signal (e.g. a burst signal) is a sufficient assumption of reception of the header signal at the receiver. In case the feedback signal contention is caused by more than 3 receivers[3], there is still a good chance of a proper header signal reception at the mobiles which justify the assumption made above. Furthermore, the corresponding header slot in the next frame is considered to be free if there is no feedback.

2.4 Time synchronization

Since RNET MAC uses a TDMA scheme, a basic working condition is the time synchronization of all mobiles. There is no explicit synchronization signal defined in RNET MAC. For synchronization purposes header and/or feedback signals are taken into account. Local timers are adjusted with the reception of every header or feedback transmission. To encompass signal delays and variances of clocks, an appropriate intergap time between two consecutive slots must be found. There are well known problems in distributed timing control

related with time shift and erroneous timing of mobiles. Even the hidden terminal scenario's may have influences on timing control because of distance and possible different synchronisation signals for a mobile. For some of these problems there are several distributed time control solutions proposed in the literature (e.g. [1, 3]). A further evaluation of that topic is necessary.

3 PERFORMANCE RESULTS

The performance of RNET MAC was evaluated by simulation. For the sake of simplicity we assume no link errors and synchronicity of mobiles. Further a frame consists of 5 slots and a message consists of a fixed number of packets, each filling up a complete ramp (about 300 bits).

The simulation scenario are shown in Figure 6. First, we took a fully meshed scenario, where all mobiles cover each other. There are 20 stations whereby mobiles 1,2 ... 10 send to mobiles 11,12 ... 20. This has the effect that the mobile compete only for channel access. Contention at the receiver is excluded (see section II B.). A sender has not to wait for the end of another ongoing transmission to the chosen receiver, since there is always only one sender for a certain receiver. We took this scenario to exploit the potential of the access mechanism.

Second, we took a hidden terminal scenario, where mobile 1 is out of reception distance of mobile 10 and vice versa, but the rest of mobiles are able to receive signals from all other mobiles including 1 and 10. We made again the assumptions of single sender/receiver pairs.

Third, we simulated a client server scenario. As outlined in section III B., a mobile can only send data to another mobile if this mobile is not receiving or sending data. Since the receiving phase is exactly one frame, the mobile can again receive data only when this phase is over. As this phase can be longer than one frame (CHT > 1) and/or other mobiles get access first, the waiting time can be very long. To evaluate such a scenario we took a fully meshed scenario, where 9 mobiles (clients) requesting access to 1 mobile (server).

3.1 Fully meshed working scenario

3.1.1 RNET MAC with/without stop bit

To investigate the influence of the stop bit, the channel holding time (CHT) was varied. Every mobile worked under full load, that is, at any point in time there was a message in the MAC buffer. A message consists of one packet.

As anticipated, the use of the stop bit has advantages (Figure A.1). The advantage is lowered as the CHT becomes larger. If the CHT is infinite, the

[3]If there is a feedback contention caused by 2 mobiles then clearly both response to different but correct header receptions. If two or more mobiles are involved in feedback contention then at least 2 mobiles are responding to a correct header reception.

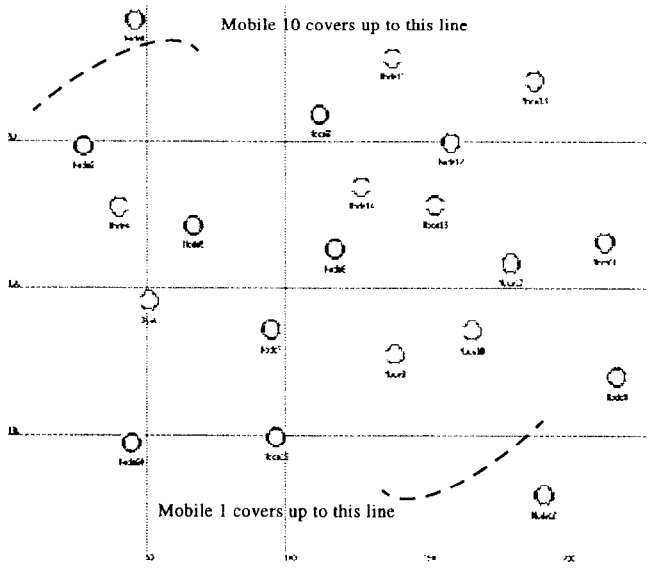

Figure 6: Simulation scenario: 20 mobiles in a 200 × 250 meter cell

throughput of both curves would be equal. But as the CHT increases the access fairness decreases. In a more realistic region of 1 to 20 CHT the gain of the stop bit is around 10%. Furthermore, at CHT 1 the performance of RNET MAC is comparable to slotted Aloha under full load. In case no stop bit is used, the performance is lower because one slot is wasted after every transmission. In all following simulations the stop bit is used and CHT is set to 5.

3.1.2 RNET MAC burst handling (handling of short and long messages)

Beside CHT, the length of the messages have to be considered. This has consequences on the throughput, queueing delay and collision rate. To evaluate the influence of the message length, two constant message sizes are chosen (1 packet per message and 5 packets per message (burst case)) and the load is varied. That is, the interfamily time between to consecutive messages is larger for messages consisting of 5 packets than for messages consisting of one packet at the same load.

At a first glimpse, the throughput curves (Figure A.2) are very similar even though the curves of collisions (Figure A.3) show a completely different behav-

ior. There are only some marginal throughput differences in the load region from 0.75 till 1.0.

The collision rate grows rapidly for short messages and stabilizes at a load of one. For long messages, the collision rate increases slower than for short messages and stabilizes at the same points as for short messages. The reason for the similar throughput curves are based on the following facts: In lower load regions there is enough capacity to resolve collisions of short (as well as long) messages. Therefore, there is no performance degradation for short messages. As the load increases, the short messages are treated by RNET MAC as long messages since the probability, that the buffer is not empty and a temporal channel is built up is higher. This effect reduces the collision rate for short messages, if the load becomes more than 0.5.

As anticipated, the queueing delays (Figure A.4) are slightly shorter for short messages in lower load regions (<0.5) since packets do not have to wait very long for transmission. For messages consisting of 5 packets, the 5th packet has to wait at least 4 slots before it is transmitted. There is a break point at a throughput of around 50%. The delay here is around 27 slots. After this point the delay for short messages will grow exponentially because of retries resulting from a large number of collisions and later on through the infinity of buffer space. Delays for long messages remains stable for loads up to 65% and also grows exponentially afterwards.

From the curves above, one can conclude that RNET MAC shows a nearly equal handling of short and long messages. Long messages only have advantages in terms of delays in higher load regions.

3.2 Hidden terminal scenario

For the sake of comparability the fully meshed is chosen. There is only one change: mobile 1 (left upper edge) and mobile 10 (right lower edge) are hidden from each other. The objective of the simulation is to show how RNET MAC behaves in a hidden terminal scenario; header information or feedback information is used to evaluate reservation of slots. As shown in [14] for the IEEE 802.11 MAC protocol[4] with a similar network configuration, hidden terminals can lead to an excessive degradation of the performance.

As anticipated in section II C, header slot reservation based only on feedback (HSRF) information outperforms header slot reservation, which is only based on header information (HSRH) (Figure A.5). The reason for that is two-fold. First, HSRH is not able to resolve collisions between the two hidden mobiles because the feedback information is not evaluated. Clearly this results in a

[4]It should be mentioned, that the IEEE 802.11 MAC protocol has an optional feature to avoid collision through a per packet bandwidth reservation mechanism. This mechanism is based on two additional signalling messages: RTS (ready to send - bandwidth reservation around sender) and CTS (clear to send - bandwidth reservation around the receiver). RTS is comparable to the header information and CTS is comparable to the feedback information in RNET.

higher collision rate (Figure A.6). Second, if two mobiles start the transmission at the same time even in the full coverage area (mobile 2-9, 11-20) they cannot recognize them. As outlined before, one can find more basic scenarios which show the superiority of HSRF.

If the performance results from section III A are compared with the results of HSRF one can only find marginal differences even though there are hidden terminals. RNET behaves nearly optimally in hidden terminal environments.

3.3 Client server scenario

As shown in the results, the performance of RNET MAC in this scenario (9 mobiles sends to one server) is relatively poor. Therefore other concepts like ramp sharing or cutting will be investigated. It should be mentioned that this effect is improved, if the server sends data to the mobiles as there is no concurrence (collision) and the clients could be served one after the other. However, data would still have to wait in the transmit queue.

A comparison with the throughput curve of the fully meshed scenario from section III A (compare Figure A.2 and A.8) shows a throughput degradation of about 70%. Less than 20% of the available bandwidth is used. But there is still a positive point; the maximum achievable throughput in this specific configuration is about 20% which is nearly achieved.

In contrast to the scenario chosen in section III A and III B, the collision rate goes up very early and to a high level (compare to Figure A.3 and A.6). Also the queueing delay (Figure A.10) becomes infinite at a low load level. This is because RNET MAC is only able to serve at maximum 1 slot per frame (about 20% throughput) in this network configuration. If the load increase above this level the queueing delay grows infinitely.

The simple idea of multiple transmitter/receiver at the server can improve the performance of RNET MAC in the client server scenario substantially. This is justified by the assumption, that a server is always more expensive than ”simple” mobiles.

4 CONCLUSIONS

We presented a novel MAC protocol, which features point-to-point communication and distributed access control. The features of this MAC as well as performance results have shown, that RNET MAC is applicable in principle in WATM environments.

The distributed access control of RNET MAC results in a good applicability for environments without pre-installed infrastructure as for instance base stations, which are the more expensive part wireless networks with centralized control. However, centralized control reduce the functionality of mobiles,

which makes them less expensive. Also, the implementation of QoS control is easier. The point-to-point communication paradigm improves the bandwidth efficiency, because communication via a BS divides the available bandwidth by the factor of two. RNET MAC is flexible with respect to packet sizes. According to the chosen parameter (for instant ascend of ramp), a ramp can carry a packet with an arbitrary number of bits. For the sake of transparency the ramp size should be a multiple of an ATM size. As well as WATM MAC based on centralized control, RNET MAC supports ATM QoS enforcement by means of reservation and the channel holding time. In particular, CBR services as well as efficient support of best effort services (UBR) are possible. RNET MAC gives also indirect means for error control. In case of collisions, a retransmission will be executed. Note, that RNET MAC bases on the assumption, that correct header reception implies a correct reception of data. That is valid for sufficient short time intervals only.

Notes

1. THIS WORK HAS BEEN SUPPORTED BY A GRANT FROM THE BMBF (GERMAN MINISTRY FOR SCIENCE AND TECHNOLOGY) WITHIN THE PRIORITY PROGRAM ATM*MOBIL*.

2. Normally, this entity is called bases station (BS) or access point (AP).

3. HIgh PERformance LAN type 2 is a Wireless LAN with ATM capabilities

4. European Telecommunications Standards Institute

5. An extreme case would be an infinite ascend (TDM) which would require cost intensive broadband transmitter/receiver. An ascend with the value zero would result in an FDM scheme.

6. The values are hypothetical and subject to change.

References

[1] Arvind, K.: ”Probabilistic Clock Synchronization in Distributed Systems”, IEEE Transactions on Parallel and Distributed Systems, Vol. 5, No. 5, pp. 474-487.

[2] Bharghavan, V et al.: ”MACAW - A Media Access Protocol for Wireless LAN’s”, Proceedings of SIGCOM’94, 1994

[3] Chuang, Justin C.-I.: ”Autonomous Time Synchronization Among Radio Ports in Wireless Personal Communications”, IEEE Transactions on vehicular technology, Vol. 43, No. 1, February 1994, pp. 27-32.

[4]For a more de tailed version see [4]

[4] Holtkamp, R.,Ebert, J.-P., Wolisz, A., Ramel, L.: ”A Distributed Media Access Control (DMAC) for Wireless ATM Networks”, accepted at 5th Intl. Conf. on Telecommunication Systems, Modeling and Analysis, Nashville, March 1997

[5] Phil Karn: ” MACA - A New Channel Access Method for Packet Radio”, AHHL/CHHL Amateur Radio 9th Computer Networking Conference, Sept. 22), 1990

[6] M. J. Karol, Z. Liu and K.Y. Eng: ”An efficient demand-assignment multiple access protocol for wireless packet (ATM) networks”, Wireless Networks, Baltzer Science Publishers, vol. 1, no. III, 1995

[7] D. Petras: ”Medium Access Control Protocol for wireless, transparent ATM access”, In IEEE Wireless Communication Systems Symposium, Long Island, NY, November 1995

[8] John Porter, Andy Hopper: ”An ATM based protocol for Wireless LANs”, ORL Technical Report 94.2, 1994

[9] John Porter et al.: ”The ORL Radio ATM System, Architecture and Implementation”, ORL Technical Report, 1996

[10] Ramel, Louis: ”Contribution to an Hiperlan (17,2 GHz)”, ETSI RES 10 Conference-Paper, Thomson / CSF-CNI, Nizza, 6.-9. June 1995

[11] Raychaudhuri, D. et al.. (1996): ”WATMnet: A Prototype Wireless ATM System for Multimedia Personal Communication”, Proc. of ICC96, pp. 469-477

[12] Smulders, P.f.M., Blondia, C.: Application of the Asynchronous transfer Mode in Indoor Radio Networks”, Proc. of Joint PIMRC and WCN conference 94, The Hague, pp. 839, Netherlands, 1994

[13] IEEE Personal Communications Magazine: Special Issue on ”Wireless ATM”, vol.3, no.4, August I996

[14] Jost Weinmiller, Hagen Woesner, Jean-Pierre Ebert, Adam Wolisz: ”Analyzing the RTS/CTS Mechanism in the DFWMAC Media Access Protocol for Wireless LAN’s”, IFIP TC6 Workshop Personal Wireless Communications (Wireless Local Access), Praque, April ‘95

Appendix: Performance figures

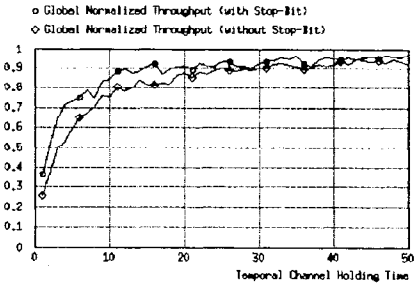

Figure A.1: Throughput vs. CHT under full load - Use of stop bit

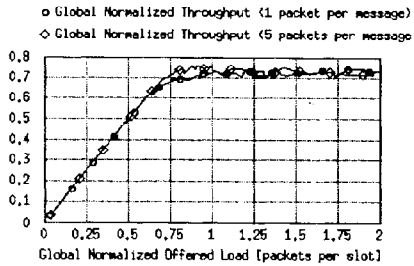

Figure A.2: Throughput vs. load for long and short messages

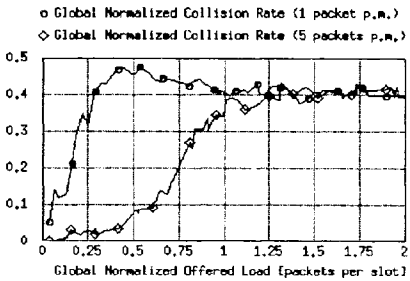

Figure A.3: Collision rate vs. load for long and short messages

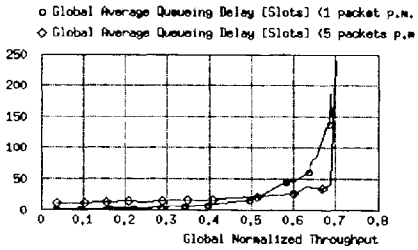

Figure A.4: Queueing delay vs. load for long and short messages

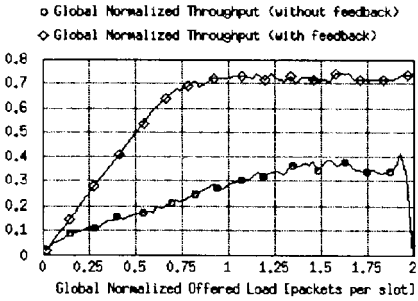

Figure A.5: Throughput vs. load - feedback or header use for slot reservation evaluation

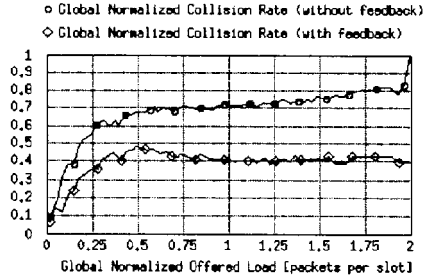

Figure A.6: Collision rate vs. load - feedback or header use for slot reservation evaluation

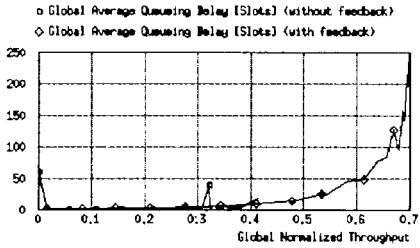

Figure A.7: Queueing delay vs. load - feedback or header use for slot reservation evaluation

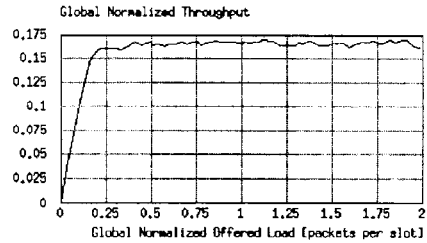

Figure A.8: Throughput vs. load in a client server scenario

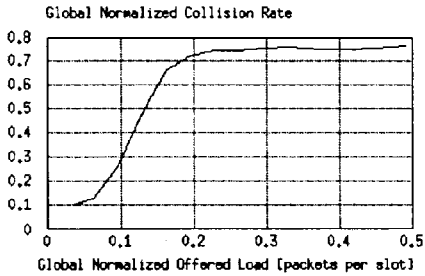

Figure A.9: Collision rate vs. load in a client server scenario

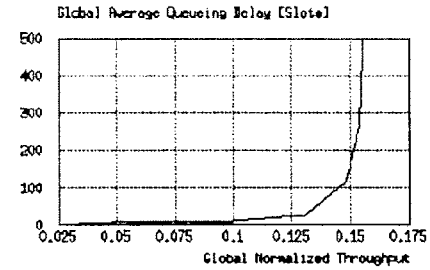

Figure A.10: Queueing delay vs. load in a client server scenario

ON DEMAND ASSIGNMENT WITH CENTRALIZED SCHEDULING: A NOVEL MAC PROTOCOL FOR WIRELESS ATM ACCESS NETWORKS

M. Artale, R. Winkler

Fondazione Ugo Bordoni
Via B. Castiglione, 59 00142 Rome Italy
e-mail: wnk@fub.it

1. Introduction

Wireless ATM has recently emerged as an appealing solution to support broadband communication services taking advantage of both the deployment flexibility of wireless networks and the information transfer capability of ATM [1]. A specific application of this new technology is in the access segment of the ATM network to provide service in rural areas and in historical city centers, where cabling has proved to be uneconomical and unfeasible, respectively.

Usually, a wireless ATM network has a star topology, in which a central device, the Base Station (BS), provides for the interworking between the access and the core networks and centralizes some protocol functions. The link between the BS and the associated Customer Equipment (CE) is organized into at least a down-channel, from the BS to the CE, and an up-channel, from the CE to the BS. The BS uses in broadcast mode the capacity of the down-channel and the CE share the capacity of the up-channel.

The last feature introduces one of the key open issues for wireless ATM: the provision of the capabilities required to coordinate the transmissions on the shared channel, so as to support efficiently the broadband communications services targeted by ATM. ATM has been designed for transmission over point to point channels and does not provide these capabilities; instead the utilization of a Medium Access Control (MAC) protocol layer is commonly considered for the resemblance of this

currently carried out by ETSI on Broadband Radio Access Networks [2] in liaison with the ATM Forum "Wireless ATM" Working Group.

This paper proposes a new MAC protocol, called On-Demand Allocation with Centralized Scheduling (DACS). DACS implements a connection-less service, that distinguishes between two levels of priority: ATM connections belonging to the CBR and the rt-VBR service categories receive high priority and are handled by means of a Time Bounded Service; the other connections are handled by means of a Best Effort Service and share the residual capacity. In a CE, each one of these two classes is associated to a MAC level buffer that is served according to the relevant DACS strategy.

DCAS is based on an on-demand assignment mechanism that either reserves or allocates dynamically the available radio resources to the established connections, provided that the CE's have requested them. The request of resources is operated according to a random channel access scheme, called Contention and Collision Resolution (CCR) strategy, that is distributed among the CE and is based on the well known 2 cell stack algorithm. The allocation of resources is operated by the Up-Channel Scheduling (UCS) mechanism, centralized in the BS, that allocates Contention-based Transmission Opportunities (CTO's) and Grant-based Transmission Opportunities (GTO's). A CTO may be used by those CE's that need to transmit information belonging to a connection that was previously inactive at the MAC level, i.e. for which there were no cells backlogged in the MAC level buffer; the concurrent utilization of a CTO by multiple connections results in a collision. A GTO is given for exclusive utilization to one CE, that uses it to transfer collision-free one MAC data unit belonging to one ATM connection for which a backlog of cells exists at the moment. The allocation of the CTO's is determined as a function of the current requirements of the CCR and of the number of active connections: a new strategy is proposed, that utilizes a threshold on the number of the established connections to determine when it is appropriate to start a new contention. The GTO allocation policy is a simple round robin with First In First Out discipline, with the optional utilization of the residual lifetime for the information units relevant to real time communications. This design is very simple and attains a very low overhead: it is therefore useful to evaluate its effectiveness with respect to various traffic profiles and levels, in order to understand the suitability of a MAC protocol that does not coordinate the medium access control to the current status of the ATM connections.

The section 2 describes the DACS protocol, with main reference to the dynamic allocation of the contention opportunities. The section 3 discusses the results of a simulation based performance evaluation of DACS.

2. ON DEMAND ALLOCATION WITH CENTRALIZED SCHEDULING

The On Demand Allocation with Centralized Scheduling (DACS) is a novel MAC protocol for wireless ATM networks.

DACS supports two connection-less information transfer services, best effort and time bounded (see the Fig. 1), and associates a logical queue to each one of these two service classes. One queue is used by the ABR, UBR and VBR-nrt ATM service categories and aims at providing a low loss ratio without stringent delay

requirements; the latter aims at delivering the information with a pre-defined maximum tolerable delay. DACS operates by taking into account that, usually, an ATM connection alternates between “on” and “off” periods and this gives rise to the activation and deactivation of the MAC level queues, where only an active queue contains ATM cells to transmit on the up-channel.

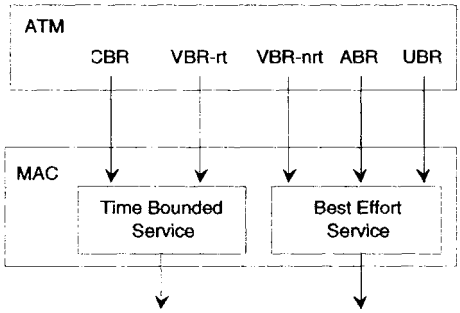

Figure 1 Relationship between DACS services and ATM service categories

Consequently, DACS shares the up-channel capacity between Contention-based Transmission Opportunities (CTO’s) and Grant-based Transmission Opportunities (GTO’s). A GTO is allocated to one CE to transfer collision-free one MAC data unit. An allocated CTO allows the CE to attempt the notification of the activation of a queue by transmitting the current head of line MAC data unit. The concurrent utilization of one CTO by multiple queues results in a collision and the transmission attempts are not successful.

The header of every MAC data unit transmitted upstream piggybacks a binary indicator to update the status of the relevant queue. This indicator is set (reset) if the queue is currently active (not active), i.e. if there is at least one backlogged cell for the relevant MAC level service class. This is a rather simple and overhead-free solution to make sure that the MAC layer does not allocate GTO’s to empty queues. This information is anyway not sufficient to handle appropriately the queues relevant to the Time Bounded Service and the additional specification of the residual lifetime of the head of line cell may be appropriate for this purpose. The implementation of this facility is left for further study and is not straightforward in the considered environment.

DACS consists of two algorithms called Up-Channel Scheduling (UCS) and Contention and Collision Resolution (CCR): the UCS allocates the CTO’s and the GTO’s; the CCR resolves the collisions and utilizes the allocated CTO’s for new contention attempts. The CCR is distributed among the CE, the UCS is centralized in the BS.

2.1. Contention and Collision Resolution

In the basic version of the CCR, called Free Access (FA), a CE that is presently not contending may use the first allocated CTO to attempt the transmission of the head of line MAC data unit on a recently activated queue, irrespectively of its instant of

activation and of the relevant traffic class. The BS notifies the outcome of the contention attempts; channel sensing by the CE is not considered for implementation because not always reliable over a wireless channel and to simplify the CE structure. A Contention Interval (CI) is the set of time slots on the up-channel that include all the consecutive CTO's that result in at least one contention attempt and the first unused CTO after them. A relatively more complex version utilizes a Gated Access (GA) to speed up the resolution of a contention. In this case the BS has the additional task of notifying the end of a CI.

Although both the GA and the FA strategy are able to support disjoint CI's for the two information transfer services, the implementation here considered relies on the undistinguished utilization of any allocated CTO by every queue that needs to advertise its activation, whether relevant to the time bounded or the best effort service.

The requirements for the fast resolution of the current contentions and for the easy identification by the BS of the end of a CI are met, among the others, by the 2 cell stack algorithm with binary feedback [3]. In this algorithm, selected also for its ease of implementation and robustness against noise and interference, the end of a CI is always and only in correspondence to the second consecutive collision-free CTO. This may be easily verified from the operation of the algorithm, whose pseudo code is shown in the Fig. 2 using the following notations:

- f_j is the binary feedback for the contention attempt carried out on the j-th CTO; f_j = 'c' (f_j = 'nc') means collision (no collision), respectively;
- $c_{i,j}$ is the binary counter of the i-th contending queue upon the allocation of the j-th CTO; $c_{i,j}$ = 1 (0) if the queue can (cannot) contend.

---

if ($c_{i,j}$ = 1) *then*
  the i-th contending queue may use the j-th CTO
  if (f_j = 'nc') *then*
    the i-th contending queue wins the contention
  else
    $c_{i,j+1}$=0 *with probability 0.5 and* $c_{i,j+1}$ = 1 *with probability 0.5*
else
  if (f_j = 'nc') *then*
    $c_{i,j+1}$ = 1
  else
    $c_{i,j+1}$ = 0

Figure 2 Two cell stack collision resolution with binary feedback

After a contention attempt, a CE waits for the feedback from the BS: in the case of collision free transmission the queue is activated and waits for the allocated GTO's to send the following MAC data units; in the case of collision the CE draws at random the value of a binary counter: the queue may use the next allocated CTO if the counter is equal to 1, otherwise it is forced to miss it and the next value assumed by the counter is determined by the outcome of the next contention attempt.

Since the up-channel is dynamically shared between CTO's and GTO's it is not easy to dimension a finite and non null window to admit in contention the activated queues; this is an important difference with respect to [3]; where the contentions are supposed to be carried out on a dedicated channel.

The proposed CCR relies on a slot based contention, in which one CTO results in the attempt to transmit at least one ATM cell per contending queue, depending on the slot length. Instead, other papers [4, 5, 6] have proposed the utilization of short MAC data units (minislots) to carry out a contention attempt; in this case a successful contention attempt does not result in the transmission of an ATM cell. The reason for this choice is that a BS will likely cover a relatively small area and serve a limited number of CE; accordingly, there is a reasonable small number of queues needing concurrently to contend and the complexity to handle the minislots is usually not counterbalanced by a performance improvement.

2.2. Up-Channel Scheduling

The UCS allocates a GTO whenever there is at least one urgent (i.e. with expiring lifetime) MAC data unit belonging to a queue relevant to the Time Bounded Service, otherwise it may either allocate a CTO or a GTO to a queue relevant to the Best Effort Service and this decision is taken on the basis of two new criteria, specifically proposed for DACS.

The first one allocates the CTO's as fast as possible when there is a contention in progress, to resolve it in the shortest possible time. The rationale is to contain as much as possible the additional random delay due to contention and collision resolution, with main reference to time bounded communications. This is constrained by the fact that the BS cannot allocate a CTO if it has not yet determined the outcome of the previous contention attempt. In addition, if the outcome of a contention attempt and the allocation of CTO's and GTO's are notified by means of specific MAC data units, their transmission on the down-channel has a delay that depends on the current backlog of data units relevant to high priority queues and on the operation of the down-channel scheduling mechanism. The minimum delay between two consecutive CTO's is $T_{down} + T_{up} + 2 * T_p$, wherein T_p is the one way propagation delay and T_{down} and T_{up} are the slot duration on the down-channel and the up-channel, respectively, and results when the feedback to the previous contention attempt and the allocation of the next CTO are notified by the same data unit, transmitted immediately after the reception by the BS of the MAC data units involved in the previous contention.

The second policy utilizes a threshold to control in a dynamic way the allocation of CTO's when there is not a contention in progress. If the number *A* of queues established and currently active is higher than or equal to a given threshold, set as a percentage of the total number *E* of established queues, than the UCS allocates a GTO to one active queue, otherwise it allocates a CTO to give a transmission opportunity to the recently activated queues (if any exists). The threshold does not take into account the traffic contract of the established queues and the status of the system and gives a qualitative picture of the current need for contention.

The pseudo code of the UCS is shown in the Fig. 3, without any details about the adopted criteria to select the active queue to be granted, if multiple choices exist at one instant. These criteria may be easily derived from the previously introduced choices to utilize round robin with FIFO for the queues relevant to the Best Effort Service and to regulate the priority of the queues relevant to the Time Bounded Service with respect to the residual lifetime of their head of line units.

```
if (there are no active queues) then
    if (the BS has not notified the outcome of the last contention) then
        the next slot on the up-channel is empty
    else
        allocate a CTO
else
    if (there is an urgent MAC data unit to be transmitted upstream) then
        allocate a GTO to the relevant queue
    else
        if ((there is a CI in progress) and (the BS has notified the outcome of
        the last contention)) then
            allocate a CTO
        else
            if (there is not a CI in progress) then
                if (the number of active queues is below the threshold) then
                    if (the BS has notified the outcome of the last contention) then
                        allocate a CTO
                    else
                        if (the active connections have some backlogged MAC data
                        units) then
                            allocate a GTO to one of these queues
                        else
                            the next slot on the up-channel is empty
                else
                    allocate a GTO to one of these queues
```

Figure 3 Pseudo code of the Up-Channel Scheduling

The UCS may be applied to realize either a slot by slot scheduling, in which a single run of the algorithm determines the utilization of the next time slot on the up-channel, or a cycle-based scheduling, in which the algorithm is iterated to determine the utilization of multiple consecutive slots on the up-channel and all the decisions are notified in a single instance. It is relevant to point out that the receiver to/from transmitter switching delay preclude the implementation of the slot by slot scheduling if TDMA / TDD is used.

3. PERFORMANCE EVALUATION

The performance evaluation of the CCR and the UCS has been carried out by means of computer simulation, focusing on the quality of service experienced by

communications supported by the Best Effort service. The UCS is applied to realize a slot by slot scheduling mechanism, less appropriate that the cycle-based scheduling from the point of view of the actual implementation but more appropriate to assess the performance bounds of the proposed strategies in the ideal case that any decision is taken exactly when needed, according to the operation of the proposed strategies. The outcome of a contention attempt and the allocation of a transmission opportunity are notified by means of specific fields in the header of the MAC data units transmitted on the down-channel; with this assumption the UCS is made independent from the scheduling mechanism for the down-channel and every MAC data unit transmitted downstream may notify any decision relevant to the UCS and the CCR. This assumption is realistic only for the assumed slot by slot scheduling, as a cycle based scheduling requires the utilization of specific MAC data units to notify the information relevant to a whole cycle.

The two channels are ideal, free from transmission errors and have equal capacity; the latter assumption does not affect the performance results as long as only one MAC data unit out of every R transmitted downstream notifies the allocation of a single transmission opportunity, wherein $R > 1$ is the ratio between the capacity of the down- and the up-channels. The propagation delay is set to zero, a reasonable assumption for most access networks. The CE are all in visibility range from the BS and each of them is involved in a single call. Mobility issues are not considered.

The user traffic is generated by E homogeneous queues, each one associated to a single ATM connection that offers cells according to an on-off process with peak bit rate $F_{p,n}$ and mean bit rate $F_{m,n}$ normalized with respect to the up-channel capacity C; the burst length follows a truncated negative exponential distribution with mean burst length $L_{b,mean}$ (cells) and maximum burst length $L_{b,max}$ (cells). The number E of established queues is kept constant during each simulation and determines the offered load, whose values are shown in the Table 1 normalized with respect to the up-channel capacity. This model is not able to show the effects of a MAC level multiplexing of ATM connections at a CE, but highlights the implications of the proposed CCR and UCS mechanisms on the quality of service of the established ATM connections. The three queue profiles described in the Table 2 are considered to evaluate the impact of the traffic model on the performance of DACS.

Table 1. Offered load

Normalized offered load	0.25	0.37	0.5	0.62	0.75	0.92
Number of established queues	6	9	12	15	18	22

Table 2. Traffic models

	$F_{p,n}$	$F_{m,n}$	$L_{b,\ mean}$ (cells)	$L_{b,\ max}$ (cells)
Profile 1 (P1)	1/5	1/24	20	70
Profile 2 (P2)	1/10	1/24	20	70
Profile 3 (P3)	1/20	1/24	20	70

Each MAC data unit carries a single ATM cell. The size B of the MAC level buffer is either 30 or 70 cells.

The mean value and the variance of these performance parameters are evaluated:

- the *Cell Loss Ratio* is the ratio between the number of ATM cells lost for MAC transmission buffer overflow and the number of generated ATM cells;
- the *Scheduling Inefficiency* is the ratio between the number of wasted GTO's and the total number of allocated GTO's for each queue; a GTO is wasted if allocated to an active queue that currently has not a cell completely available for upstream transmission;
- the *Admission Delay* is the delay from the instant of complete generation of the first cell after the activation of a queue and the time this queue contends for the first time;
- the *Contention Delay* is number of slots required by a cell to contend successfully.

In all the plots, the value 1E-9 of the Cell Loss Ratio means that no loss has been experienced during a simulation and is used to draw the plots with a logarithmic scale. The confidence interval is not shown to simplify the figures, it is always smaller than 5% of the average value for any value of the delay and for Cell Loss Ratio greater than 1E-5.

The first group of simulation experiments has been carried out to compare the performance attained by the proposed slot based CCR against those attained when the contention attempts are carried out using minislots. Each slot conveys 4 minislots, each of them carrying the data necessary to identify the contending queue, and the 2 cell stack algorithm is generalized to accommodate multiple collisions or successful contentions per slot. The GA strategy is used.

The Fig. 4 shows that the minislot based contention is able to improve the Cell Loss Ratio performance, but the advantage is often limited, if not negligible, for the queue profiles P1, P2 and P3. In fact, the Cell Loss Ratio attained by the two arrangements is similar for threshold set to $T = 0.5 * E$, (this is the considered default threshold value) because the number of concurrently contending queues is limited to at most a few units and the slightly faster collision resolution of the minislot based contention is compensated by its higher overhead, due to the fact that an ATM cell is not transmitted in the case of successful contention. The only appreciable difference is when 22 queues with profile P3 are set-up, in which case the long persistency of the time interval in which the number A of established and active queues is greater than T has the effect of blocking the allocation of the CTO's for a significant period: the faster collision resolution of the minislot based contention is beneficial to attenuate this disadvantage. The plots of the Cell Loss Ratio are similar for P3 and P2 and tend to the same value for increasing offered loads; in fact both these sources waste many GTO's for low and medium offered loads, due to the inability of the round robin scheduling in tracking the actual requirements of the active queues, and this effect looses its relevance only at high offered load, in which case the inter GTO delay (i.e. the time elapsing between two successive allocation of GTO to the same queue) becomes critical. As for P1 the Scheduling Inefficiency is often very low and the sharp increase of the Cell Loss Ratio is due to the round robin scheduling that operates irrespectively of the number of cells backlogged at the MAC buffer of the active queues, with the result that the inter GTO delay becomes larger than the inter-cell time at the source peak rate and the MAC level buffers overflow. The oscillations of the plots are due to the truncation carried out when E is odd to meet the natural requirement of integer threshold value. In the same figure the Cell Loss Ratio for normalized offered load

equal to 0.25 is larger than that for normalized offered load equal to 0.37 because of the effect of the threshold value that is too high when 6 queues are established.

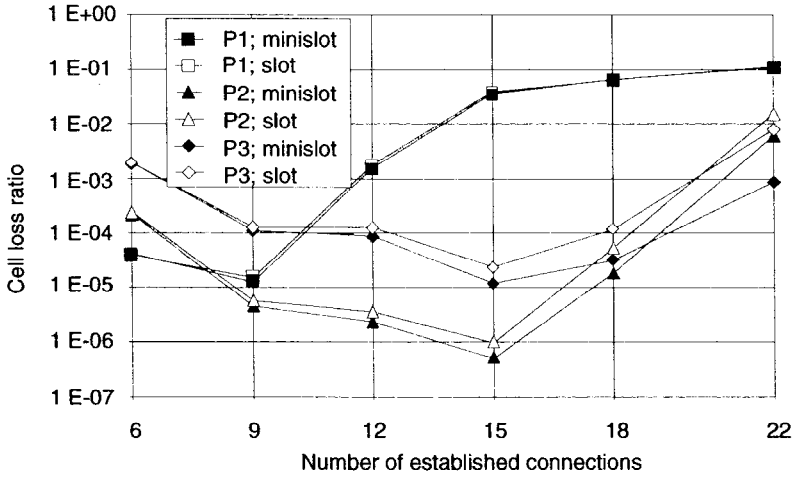

Figure 4. Minislot vs. slot contention: Cell Loss Ratio for the default threshold

The effect of the threshold value on the Cell Loss Ratio is shown in the Fig. 5 for normalized offered load equal to 0.92. P1 needs a small threshold to receive more GTO's than CTO's and its Cell Loss Ratio increases always with the threshold; the Cell Loss Ratio of P2 and P3 is very sensitive to the threshold value and presents a minimum value in correspondence to a threshold value lower than $\lceil 1 / F_{p,n} \rceil$ for both of them. This value is meaningful only if enough queues are established, otherwise the Cell Loss Ratio decreases with the threshold increasing up to E.

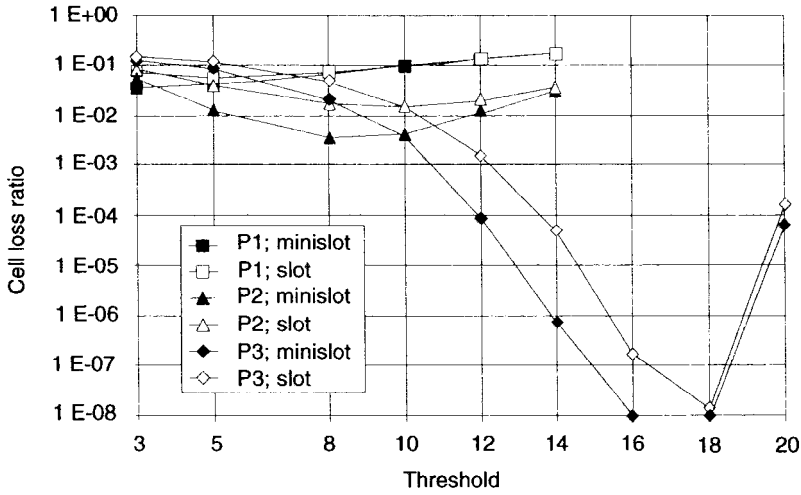

Figure 5. Effect of the threshold value on the Cell Loss Ratio

The Scheduling Inefficiency is plotted in the Fig. 6 for the GA strategy with cell based contention, for variable threshold and normalized offered loads 0.62 and 0.75. The Scheduling Inefficiency is higher for lower offered load and for lower values of the threshold, because many GTO's are allocated to the active queues that may not have an entire cell available for upstream transmission. The Scheduling Inefficiency assumes high values for P2 and even more for P3, because the number of sources concurrently active at one instant is likely to interrupt for a long period the allocation of CTO's and the UCS allocates more GTO's than those actually needed.

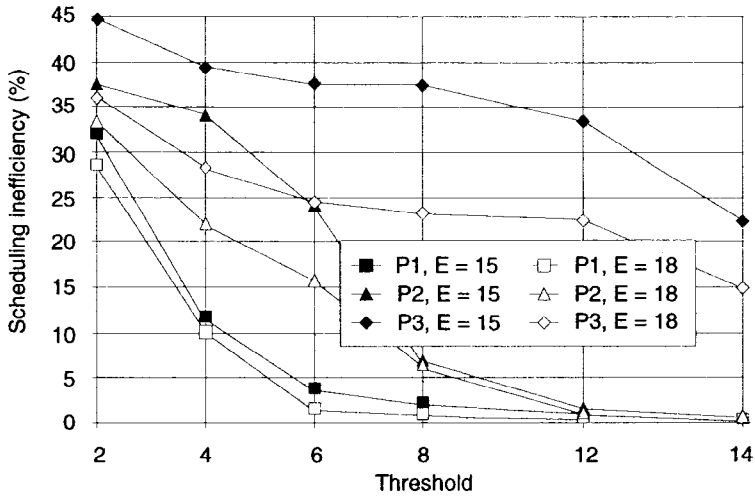

Figure 6. Scheduling inefficiency as a function of the threshold value

The average Contention Delay is plotted in the Fig. 7 for the GA strategy. For P1 the Contention Delay is equal to a few slots and lower than the inter-cell time at the source peak rate, whatever the offered load is. This is also true for P2, with the exception of the case 22 queues are established, and is not true for P3 that stresses the different behaviors of the slot based and the minislot based contention. Again, the reason is the long persistency of the period in which A is larger than the threshold and the resulting increase in the number of queues waiting for a CTO to notify their activation. It is anyway worth highlighting that the Contention Delay for $E = 22$ is about 2 cell times for P2 and less than 3 cell times for P3.

The difference between the minislot based contention and the slot based contention is magnified by the histogram plotted in the Fig. 8 for P1 and P3, for the default value of the threshold and normalized offered load equal to 0.62. This figure shows also the impact of the system latency, due only to the transmission time because the propagation delay is set to zero: the Contention Delay may only assume odd values because every CTO is always followed by either a GTO or an empty slot, to guarantee that the BS has received the contending data units before issuing a new CTO.

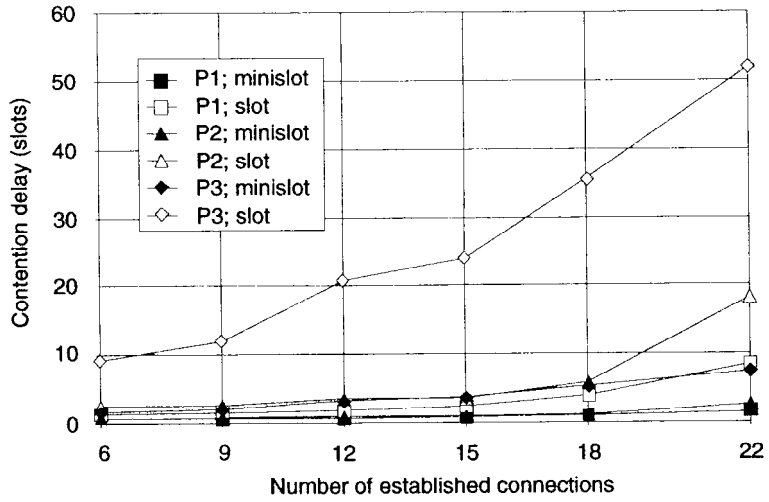

Figure 7. Minislot based vs. Slot based contention: Contention Delay

The Fig. 9 plots the Cell Loss Ratio for B=30 and B=70, for the default value of the threshold and for the slot based contention. The role played by B is evident and makes it one of the key design parameters for non real time traffic. The Cell Loss Ratio of P2 and P3 tends to very low values for low to high offered load and is significant only for very high offered load. The Cell Loss Ratio of P1 is acceptable only for low to medium values of the offered load because this profile presents tight service requirements that cannot be easily accommodated by the round robin scheduling, as explained above.

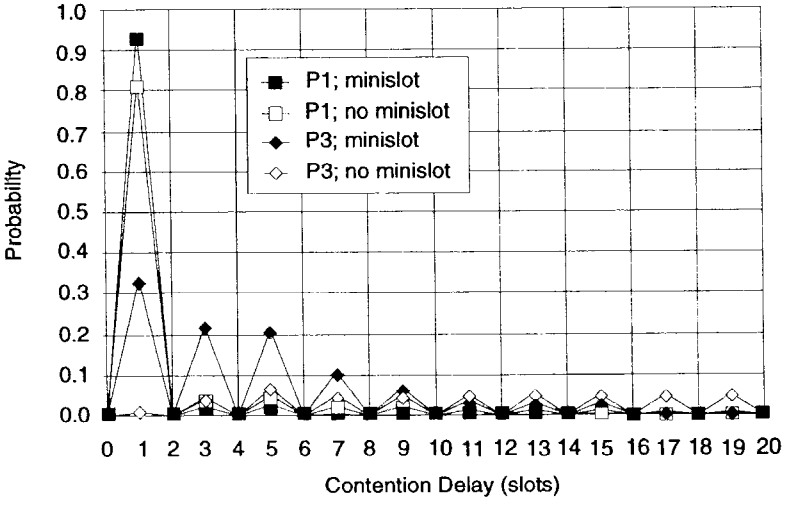

Figure 8. Minislot vs. Slot contention: histogram of the Contention Delay

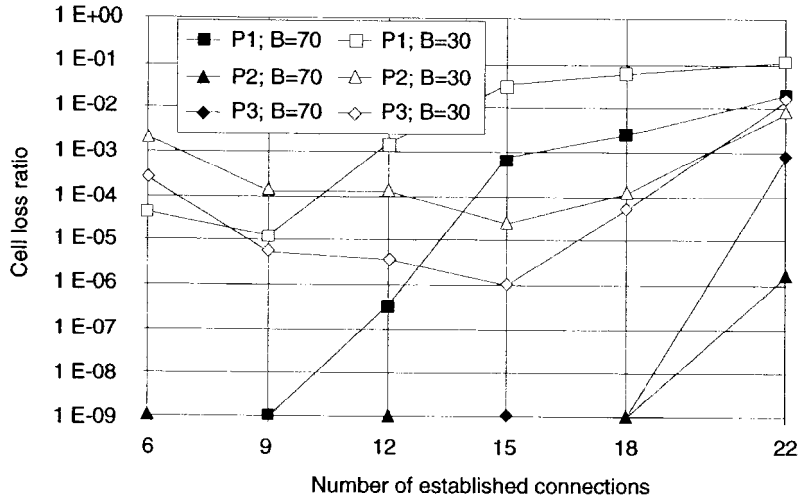

Figure 9. Short vs. long MAC level buffer: Cell Loss Ratio

The second group of simulations compare the FA and the GA strategies. No significant differences emerge between the Cell Loss Ratio (see the Fig. 10) for the low number of contending queues. This is confirmed by the fact that a CI has average duration of about 1 slot and, so, the operation of the GA strategy becomes equivalent to that of the FA strategy.

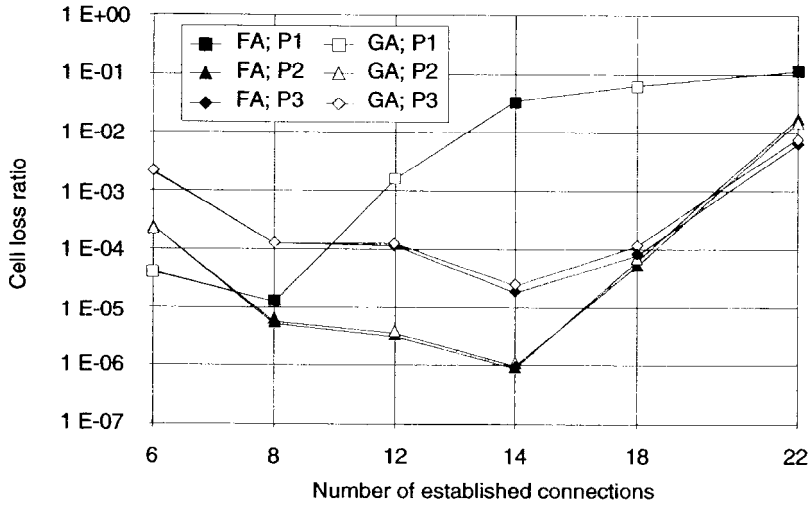

Figure 10. Free Access vs. Gated Access: Cell Loss Ratio

The Admission Delay decreases with the increase of the threshold (see the Fig. 11) because this entails the allocation of shorter sequences of consecutive GTO's, for any offered load. Conversely, the FA strategy entails slightly higher

Contention Delay (see the Fig. 12) because the on-the-fly contention admission may interfere with an in progress contention attempt.

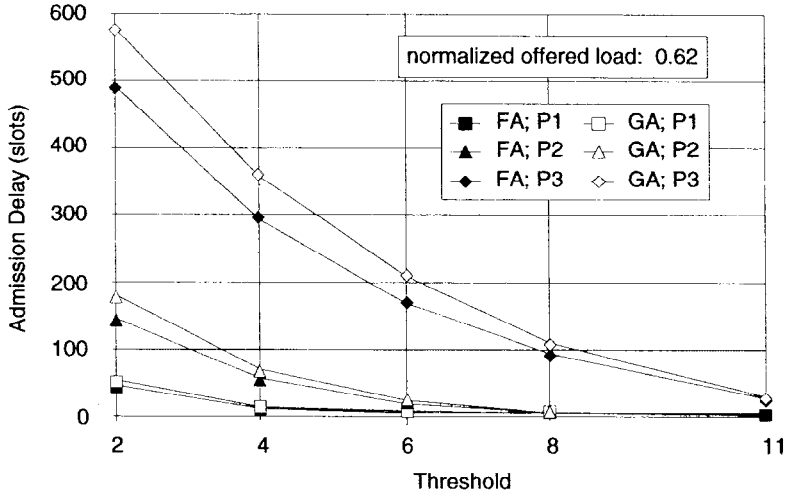

Figure 11. Free Access vs. Gated Access: Admission Delay (variable threshold)

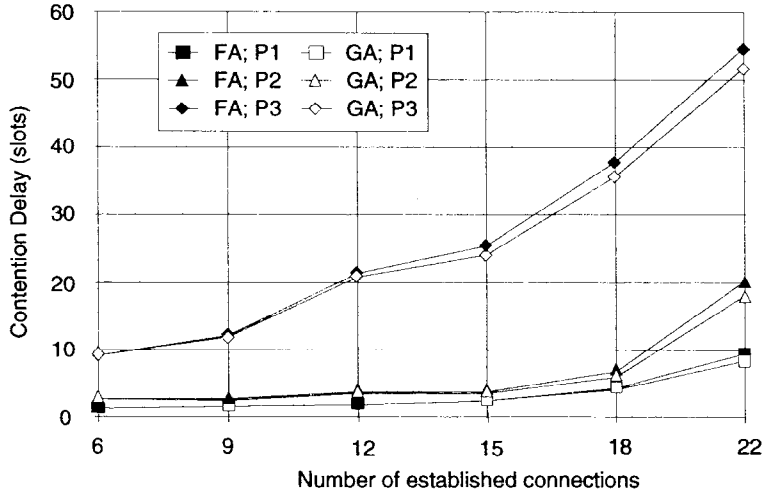

Figure 12. Free Access vs. Gated Access: Contention Delay (default threshold)

To conclude, the performance results collected so far and discussed in this section show that the rationale behind DACS seems worth of consideration but several issues are yet to be solved before it can be considered an appropriate solution for the MAC protocol in wireless ATM networks. The key open issue is to find a balance between the simplicity of the proposed mechanisms, able to operate without any coordination to the ATM protocols and thus providing a true

low overhead solution, and their actual ability in meeting the quality of service requirements negotiated at the ATM layer. A connection-oriented approach would allow a higher level of compliance to these requirements, at the expense of a higher design complexity. Alternatively, it may be appropriate to rely on a more complex piggybacking scheme, like the notification of the current backlog size at the Best Effort MAC queues; this entails anyway a larger overhead than the binary scheme here considered. Then, a heuristic to place the threshold is to be found to deal effectively with the large spectrum of source profile. Last, the simple two cell stack algorithm does not behave properly if the number of contending queues is high: in this case alternative approaches are required and some of them are being evaluated at the moment of writing this paper.

5. REFERENCES

[1] IEEE Personal Communications; August 1996, Vol. 3 No. 4; special issue on "Wireless ATM"

[2] ETSI Project on "Broadband Radio Access Networks": Terms of Reference; available at http://www.etsi.fr/bran/

[3] P.Papantoni-Kazakos: "A simple window random access algorithm with advantageous properties"; IEEE Infocom 88, New Orleans, Mar. 1988

[4] K.Y.Eng, Z.Liu, M.J.Karol: "An efficient demand assignment multiple access protocol for wireless packet (ATM) networks"; Wireless Networks, Oct. 1995

[5] B.Walke, D.Petras, D.Plassmann: "Wireless ATM: Air interface and Network Protocols of the Mobile Broadband System"; IEEE Personal Communications; Aug. 1996

[6] D.Raychaudhuri, N.D. Wilson: "ATM-based transport architecture for multiservices wireless personal communication networks"; IEEE JSAC, Oct. 1994

LOCATION MANAGEMENT IN WIRELESS ATM NETWORKS

Gopal Dommety
Computer and Information Science, Ohio State University
Columbus, OH 43210, USA

Malathi Veeraraghavan
Bell Laboratories, Lucent Technologies
Holmdel, NJ 07733, USA

ABSTRACT[1]

This paper addresses the location management problem in wireless ATM (Asynchronous Transfer Mode) networks based on the PNNI (Private Network-to-Network Interface) standard. There are two approaches to performing location management: the "mobile computing" approach and the "cellular telephony" approach. In this paper we present two mobile location management algorithms, the *mobile PNNI* scheme and the *LR (Location Register)* scheme, for wireless ATM networks. The mobile PNNI scheme represents the "mobile computing" approach and builds on the PNNI routing protocol. It uses limited-scope (characterized by a parameter S) reachability updates, forwarding pointers and a route optimization procedure. The LR scheme represents the "cellular telephony" approach and introduces location registers (such as the cellular home and visitor location registers) into the PNNI standards-based hierarchical networks. This scheme uses a hierarchical arrangement of location registers with the hierarchy limited to a certain level S. A comparison of the two schemes indicates that at low CMRs (Call-to-Mobility Ratios), the LR scheme incurs a lower average total cost (consisting of the average move and average search costs) than the mobile PNNI scheme, while at high CMRs, the mobile PNNI scheme incurs a lower cost. We also observe that the two schemes show a contrasting behavior in terms of the value to be used for the parameter S to achieve the least average total cost. At low CMRs, the parameter S should be high for the mobile PNNI scheme, but low for the LB scheme, and vice versa for high CMRs.

1. This work was partially funded by the Advanced Technology Program of the National Institute of Standards and Technology, U.S. government.

1. INTRODUCTION

Mobility management algorithms enable networks to support mobile users allowing them to move, while simultaneously offering them incoming calls, data packets and/or other services. In connection-oriented networks, *mobility management* consists of *location management* and *handoff management.* Handoff management deals with rerouting connections, on which the mobile user was communicating while moving, with minimal degradation of quality of service. Location management deals with tracking mobiles and locating them prior to establishing an incoming call. Location management schemes provide mechanisms to track and locate mobiles. The two aspects of this problem, mobile tracking and mobile locating, are also referred to as *MOVE* and *FIND* operations, respectively in [1][2]. Mobile tracking is the procedure by which the network elements update information about the location of the mobile. Mobile location is the procedure by which the network finds the exact location of the mobile. The information acquired during the tracking phase is used in the locating phase.

Since both ATM networks and cellular telephony networks are connection-oriented, the IS-41 and GSM MAP (Mobility Application Part) standards [3][4] offer a natural starting point for the design of location management algorithms in ATM networks. On the other hand, the PNNI (Private Network-to-Network Interface) routing protocol standard proposed by the ATM Forum [5], for propagating network topology, loading, and reachability information, is also a candidate starting point for an ATM location management algorithm.

Using these two starting points, we propose two algorithms for location management in PNNI standards-based ATM networks. In the first algorithm, we add features to the PNNI routing protocol to enable it to handle mobile users. We refer to this solution as the *mobile PNNI* scheme. In the second algorithm, we introduce location registers of the type used in cellular telephony networks, into the PNNI standards-based ATM networks. These location registers are databases that track mobile locations and respond to location queries. This solution is referred to as the *LR (Location Registers*) scheme.

We review prior work on location management in Section 2. Next, we describe the proposed *mobile PNNI scheme* and the proposed *LR scheme* in Section 3. A discussion of these two schemes based on an analytical comparison [6] is included in Section 4. Our conclusions are presented in Section 5.

2. PRIOR WORK

There are two approaches to performing location management: the “cellular telephony” approach and the “mobile computing” approach. The cellular IS-41 MAP (Mobility Application Part) standard [3] and several improvements proposed in [1, 7, 8, 9, 10, 11] are representatives of the cellular telephony approach, and mobile IP [12, 13] is a representative of the mobile computing approach.

Fig. 1 shows the cellular network architecture and the mobile tracking scheme defined in the IS-41 standard. The network elements consist of BS (base stations), MSCs (Mobile Switching Centers), VLRs (Visitor Location Registers) and HLRs

(Home Location Registers). Base stations provide wireless access to mobiles. MSCs are the switching centers through which connections to and from mobiles are made. HLRs and VLRs track the location of mobiles and answer queries for mobile locations. The cellular standards allow for a variation of this network architecture in which *VLRs are collocated with the MSCs.*

Mobile tracking begins with mobiles generating registrations to the network. For the purposes of this paper, we assume that *power-on, power-off,* and *zone-change registrations* are generated by each mobile. The power-on (and power-off) registrations allow the network to maintain a list of mobiles enabled to receive calls. Zone-change registrations allow the network to track a mobile user as he/she moves from one zone to the next as illustrated in Fig. 1, where a *zone* is defined to include all the base stations under an MSC. Each base station emits a periodic broadcast beacon carrying the zone identifier of the base station. This allows a listening mobile to identify when it moves into a new zone. An effect of this form of tracking is that a *page* needs

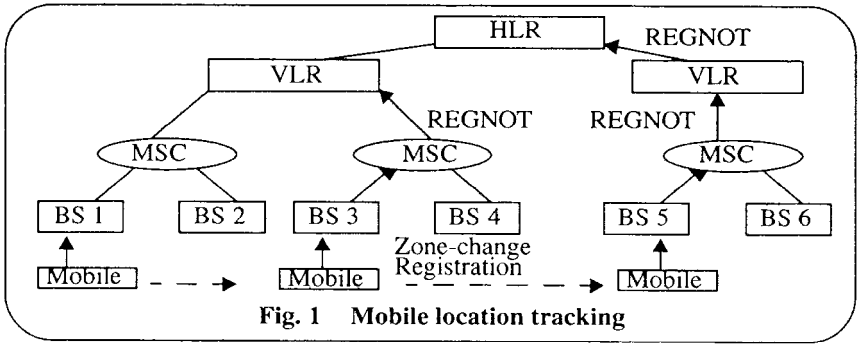

Fig. 1 Mobile location tracking

to be generated by the MSC to determine the exact base station at which a mobile is located during call setup. In Fig. 1, when a user moves from base station 1 to base station 2, it does not generate a registration, since both base stations are in the same zone. A zone-change registration is sent by the mobile when the user moves from base station 2 to base station 3. REGNOTs (Registration Notifications) are generated from MSCs to VLRs, and from VLRs to HLRs (only if the VLR changes) as shown in Fig. 1.

To deliver an incoming call to a mobile, the calling party's switch (originating switch) sends a LOCREQ (Location Request) message to the HLR of mobile. The

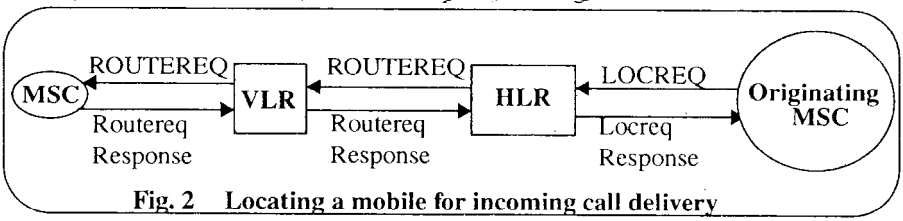

Fig. 2 Locating a mobile for incoming call delivery

HLR sends a ROUTEREQ (Route request) to the VLR from which it received the last registration message for mobile and the VLR in turn sends it to the MSC. The MSC allocates a TLDN (Temporary Location Directory Number) for this mobile and returns this value in its response. Routing of the connection through the network is done using the TLDN which indicates the far end MSC (of the called mobile). When

the call setup message reaches the MSC of the called mobile, the MSC issues a page to locate the exact base station at which the mobile is located. If VLRs and MSCs are collocated, then the message exchange between the VLR and the MSC is avoided. However, HLRs experience a heavier registration (tracking) load.

Among the improvements proposed to this scheme are the extremes of the "flat" scheme [8, 9] and the "hierarchical" scheme [10, 11]. The former proposes using a single-level hierarchy of location registers, while the latter proposes building a rooted tree of location registers. In the *flat* scheme, upon receiving a LOCREQ, the HLR assigns a temporary address (TLDN) based on the VLR/MSC at which the called mobile is located rather than require an additional message exchange from the HLR to the VLR/MSC to obtain a temporary address assignment. The mobile's permanent

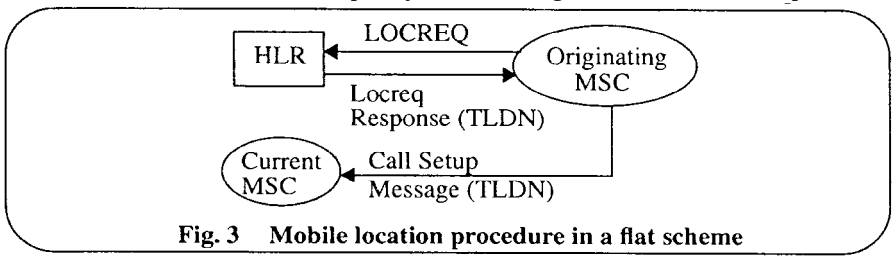

Fig. 3 Mobile location procedure in a flat scheme

address is tunnelled in the call setup message while the temporary address is used to route the connection from the call originating switch to the mobile's current switch (see Fig. 3). The *hierarchical* scheme uses a hierarchy of location registers to localize both mobile tracking and mobile locating messages. A registration is propagated up the hierarchy until it reaches a location register beyond which there is no change of information regarding the mobile's location. The call setup message (or an explicit location query) is sent up the hierarchy until it reaches a location register that knows the location of the mobile from which point the hierarchy is traced in the downward direction to reach (or determine) the exact switch where the mobile is located. Fig. 4

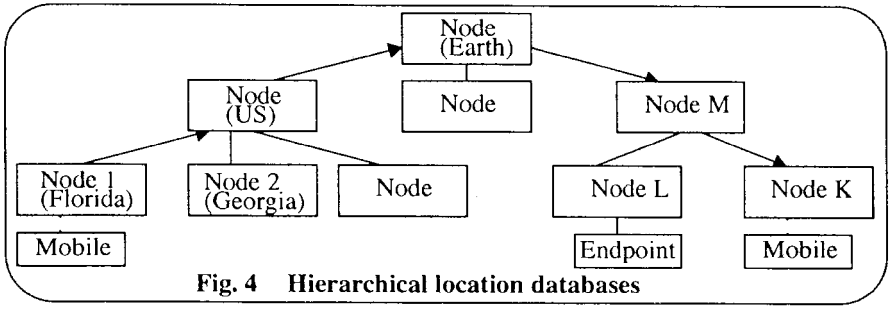

Fig. 4 Hierarchical location databases

shows the network architecture assumed in the hierarchical scheme. If the mobile, whose home is node 1, moves to node K, a chain of pointers are set up as shown in Fig. 4 from the home node (node 1) to node K. If an endpoint at node L calls the mobile, node L sends a call setup message to node M (in the direction of the home node of the mobile). Since node M has a pointer for the mobile (indicating node K), it sets up a connection to node K. Thus, long-distance signaling is avoided by using this chain of pointers to cut short searches.

The flat scheme results in lower computation costs but incurs larger communica-

tion costs than the cellular scheme, while the hierarchical scheme achieves the opposite (lower communication costs, but higher computation costs). Other improved schemes, such as the forwarding scheme of [1] and the anchor scheme of [7], are in between these two extreme schemes in terms of computation and communication costs. In contrast, the *Location Registers (LR) scheme* for wireless ATM networks is a hybrid scheme whose parameters can be set to default to one of the two improved schemes, i.e., the flat scheme or the hierarchical scheme. It essentially uses a hierarchy of location registers but limits the hierarchy by lopping off the tree at some level *S*, beyond which it resorts to the flat scheme approach of updating/consulting a home location register.

Other related work on location management includes mobile IP [12, 13] with caching and route optimization extensions [14]. Mobile IP is an extension to the Internet Protocol (IP), which enables hosts to change their point of attachment to the internet with out changing their IP address. In this protocol, a packet for a mobile host is routed to the home network of the mobile as identified by its permanent IP address. The home network tracks the current location of the mobile and tunnels the packet to the current network of the mobile. In order to prevent this "triangle" routing, route optimization extensions have been proposed [14]. These extensions provide a means for communicating-nodes to maintain a binding between the mobile and its current location, and use this binding to tunnel datagrams directly to mobile. Extensions are also provided to allow datagrams in flight when a mobile node moves, and datagrams sent based on an outdated binding information, to be forwarded directly to the mobile.

3. PROPOSED LOCATION MANAGEMENT SCHEMES

In this section, we present *two* location management schemes for wireless ATM networks. We propose both location management schemes for ATM networks based on the PNNI standards [5].

The *first scheme* enhances the PNNI routing protocol to handle mobile users, and is referred to as the *mobile PNNI scheme.* The PNNI routing protocol provides a feature to convey reachability information about the endpoints to ATM switches. This feature is exploited with minimal change to convey reachability information about mobile endpoints. In order to limit the region of nodes that receive reachability updates as mobiles move, the "scope" parameter (set to some number *S*) available in PNNI routing protocol messages is used. In this scheme, there is no explicit mobile location phase prior to connection setup. Instead, connections are set up to mobiles according to the reachability information at the switches. Switches within the region defined by *S* (relative to the position of each mobile) have the correct reachability information for mobiles. Calls originating from such switches will be routed on "shortest paths." However, calls originating at switches outside the region defined by *S* will be routed toward the home switch of the mobile, and subsequently forwarded to the current location of the mobile. This requires a *procedure for setting and updating forwarding pointers* at the home switch of the mobile, and creates the need for a *procedure to optimize the routes* of calls set up on circuitous routes.

The *second scheme* isolates the effect of mobility from the PNNI routing protocol. The cellular concept of using location registers to handle mobile users is introduced

into PNNI standards-based ATM networks. Location registers (databases) are placed within the peer group structure of these ATM networks. In this approach, location registers track the location of mobiles, and respond to location queries generated prior to connection setup. Thus, unlike the mobile PNNI scheme, the second approach, named the *LR (Location Registers)* scheme, has an explicit mobile location phase prior to connection setup.

3.1 Mobile PNNI scheme

We first give an overview of the PNNI routing protocol in Section 3.1.1 to show how *fixed endpoints* are supported in networks based on the PNNI standards. Next, we describe our proposal for supporting mobile endpoints in such networks. In Section 3.1.2, we describe the *architectural* addition needed to support mobile endpoints. In Section 3.1.3, we describe the *mobile tracking* procedure. Section 3.1.4 describes how incoming calls to mobiles are set up using this location management scheme. Finally, *route optimization* is briefly addressed in Section 3.15.

3.1.1 Overview of the PNNI routing protocol

PNNI standards-based ATM networks are arranged in hierarchical peer groups as shown in Fig. 5. At the lowest level ($l = L$), ATM switches are shown connected in

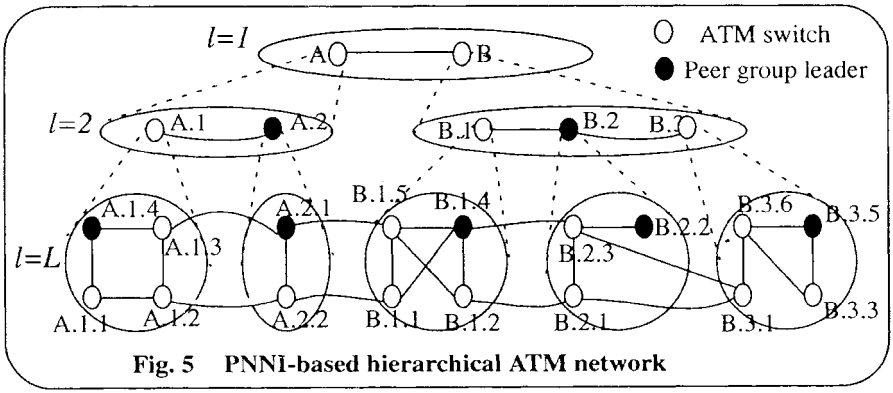

Fig. 5 PNNI-based hierarchical ATM network

arbitrary topologies. A PGL (Peer Group Leader) is elected in each peer group. This node represents the peer group at the next higher-level peer group. In this role, it is termed the LGN (Logical Group Node) representing its lower-level peer group. Nodes within a peer group exchange detailed PTSPs (PNNI Topology State Packets) and hence have complete information of the peer group topology and loading conditions. A PGL summarizes topology/loading information received in its peer group, and generates PTSPs in its role as LGN to members of the higher-level peer group. Each member of the higher-level peer group receiving this summarized information will send information to members of its child peer group (downward flow). This exchange of *topology and loading information* constitutes the *PNNI routing protocol* [5]. Using this mechanism, each node in the network has the complete topology/loading information of its lowest-level peer group, and also the topology/loading informa-

tion of its ancestor peer groups. This information is used to determine routes of connections when a call arrives.

As part of the PTSPs, *reachability information* is propagated among nodes to indicate where endpoints are located. Endpoint addressing in the ATM Forum standards is based on the NSAP (Network Service Access Point) addressing format [15]. All three forms of NSAP addressing support *hierarchical addressing, where the prefix of the address indicates the peer group in which the endpoint is located.* A switch will have exact reachability information for endpoints within its level-*L* peer group indicating the switch at which each such endpoint is located. However, for endpoints in other level-*L* peer groups, the switch will only know the higher-level peer group through which the endpoints can be reached. PTSPs carrying reachability updates also propagate up and down the hierarchy as explained earlier for the PTSPs carrying the topology and loading information. Reachability information advertised by a node has a scope associated with it. The scope denotes a level in the PNNI hierarchy and represents the highest level at which this address can be advertised or summarized [5].

As an example of a PNNI based network, consider the network shown in Fig. 5. The numbering scheme used for the nodes reflects the peer group structure. Node A.1.4 belongs to peer group A.1 at level 2, and to peer group A at level 1. Node A.1.4 is the peer group leader of peer group A.1. It advertises that all A.1 nodes are reachable through itself in peer group A. Upon receiving this advertisement, LGN A.2 sends a PTSP to all its lower-level nodes. Using this process, nodes in A.2, such as A.2.1 and A.2.2, learn that all A.1 endpoints are reachable through A.1. Similarly, nodes A.2.1 and A.2.2 learn that all nodes with the address prefix B are reachable through LGN B, since they maintain the topology, loading conditions and reachability data for all their ancestor peer groups.

3.1.2 Architecture to support mobile endpoints

Mobile endpoints can be supported in the PNNI hierarchical architecture in the following manner. Mobiles are located at base stations, which are assumed to be organized as in cellular networks, with multiple base stations connected to each switch as shown in Fig. 6. The set of all the base stations under a single switch is

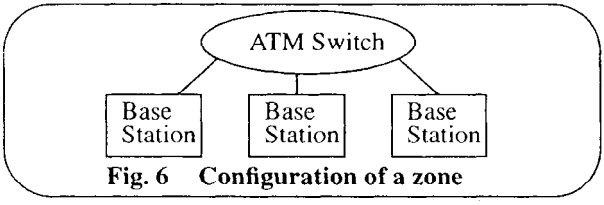

Fig. 6 Configuration of a zone

defined to be a *zone.* The concept of zones are used to limit air interface registration traffic. Instead of tracking the exact base station at which each mobile is located, the network tracks the zone locations of mobiles. This is done using zone-change registrations[1]. The configuration shown in Fig. 6 ties in with the architecture of Fig. 5 at the lowest level. In other words, some of the switches in Fig. 5 are connected to a set of base stations as shown in Fig. 6. Besides zone-change registrations, mobiles are

assumed to generate *power-up* and *power-down* registrations allowing a network to track only powered mobiles.

3.1.3 Mobile tracking

When a mobile powers on or changes locations (moves), the mobile tracking procedure uses a combination of *setting forwarding pointers* at the home and old (in case of a move) locations of the mobile, and *sending limited reachability updates* (with a scope *S*) using the PNNI routing protocol to track the location of the mobile. Forwarding pointers are set at the home in order to route the calls generated by nodes outside the neighborhood (defined by scope *S*), and at the old location (in case of a move) of the mobile to route calls generated prior to the complete propagation of the reachability information updates (resulting from the move).

Setting of forwarding pointers: If the registration received from the mobile (through the base station) is a *power-up* registration, then the switch receiving the

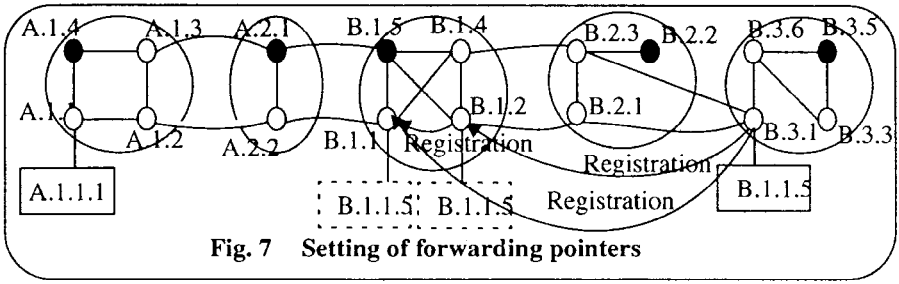

Fig. 7 Setting of forwarding pointers

registration sends a message to the home switch of the mobile. A pointer is set at the home to point to the current location. When *a mobile moves*, if the old location is outside the neighborhood (defined by scope *S*) of the new location, the forwarding pointer at the home is updated to point to the new location. A forwarding pointer is set at the old location of the mobile before distributing reachability updates to handle any calls that arrive prior to the propagation of reachability information. If the old location is within the neighborhood of the new location, the forwarding pointer at the old switch is set to point to the new location. For example, consider the mobile B.1.1.5 located under a base station connected to its home B.1.1 as shown in Fig. 7. Let $S = 3$. If the mobile moves to a base station under switch B.1.2, then a *Registration* message will be sent from B.1.2 to B.1.1 (as shown in Fig. 7). Further, if it moves to B.3.1, two *Registrations* are sent to set forwarding pointers at home and old locations (B.1.1 and B.1.2, respectively) of the mobile. The pointer at the old location can be deleted after the updated reachability information has reached all the intended nodes. In an implementation, this pointer can be deleted after a certain amount of time (with the help of a timer).

Reachability updates: After setting forwarding pointers, the feature of *sending triggered updates of topology information for significant change events* in the PNNI

1. Each base station emits a periodic broadcast beacon carrying the zone identifier of the base station. This allows a listening mobile to identify when it moves into a new zone.

routing protocol is used to generate reachability updates. The scope *S*, as explained earlier, is used to set the stopping point for reachability information propagation. If a mobile powers on at a node within the neighborhood (defined by scope *S*) of its home, the reachability updates will propagate only up to some level *K*, beyond which there is no change of reachability data. For example, in Fig. 5, if *S = 2* and a mobile A.1.1.5 powers on at A.1.2, the reachability updates only propagate to nodes in peer group A.1. No PTSP is sent from A.1 to A.2, since there is no change of reachability data stored in A.2 nodes regarding mobile A.1.1.5. If a mobile powers on at a node outside its home neighborhood, the reachability updates propagate up to level *S*. For example, in Fig. 5, if *S = 2*, reachability updates for a mobile A.1.1.5 that powers on at a base station connected to node A.2.2 will spread through peer group A.2, and then upwards (i.e., PTSPs carrying the reachability updates are sent from A.2 to all nodes of peer group A) and finally, downwards from LGNs in peer group A other than A.2 to their child peer groups. *As mobiles move*, if the old location is within the neighborhood of the new location of the mobile, reachability updates are sent up to some level *K*, beyond which there is no change of reachability data. If not, it spreads up to level *S* from the new location of the mobile. In addition, a reachability update procedure is initiated by the old switch to notify all the nodes within a region *S* relative to the old switch that the reachability of the mobile is now through its home.

Details regarding how registration messages are transported are two-fold. First, each mobile maintains the identifiers of its old and home switches, allowing it to communicate this information when registering at a switch (power-on, power-off or move). This information is used by the new switch to generate the registration message to the home or old switch. Second, we assume the availability of connectionless transport to send location management messages, such as registrations. One such transport mechanism is connectionless ATM (CL-ATM) proposed in [16]. Without this assumption, on-demand connections would need to be set up and released for the transport of every *Registration* message, which creates a considerable processing and signaling overhead.

3.1.4 Connection Setup/Mobile locating

In the mobile PNNI approach, there is no explicit mobile location procedure prior to connection setup. Instead, connection setup proceeds with every switch "believing" its reachability information. Therefore, if the *calling party is within the neighborhood (defined by scope S) of the called mobile's location*, since all the switches on the path have correct reachability information, the call is delivered directly to the mobile. If the calling party is outside the neighborhood of the mobile, since the reachability information in the switches outside the neighborhood indicate that the mobile is at its home, the call setup proceeds towards the home. In this case, the following two scenarios are possible: *i)* called mobile is within the neighborhood of its home, or *ii)* mobile is outside the neighborhood of its home.

If the *called mobile is within the neighborhood of its home*, the call is delivered directly to the mobile as illustrated in the following example. Consider that A.1.1.1 issues a call setup to the mobile B.3.1.4, currently located at B.1.2 (as shown in Fig. 8). The call setup proceeds through peer groups A.1 and A.2 and arrives at peer group

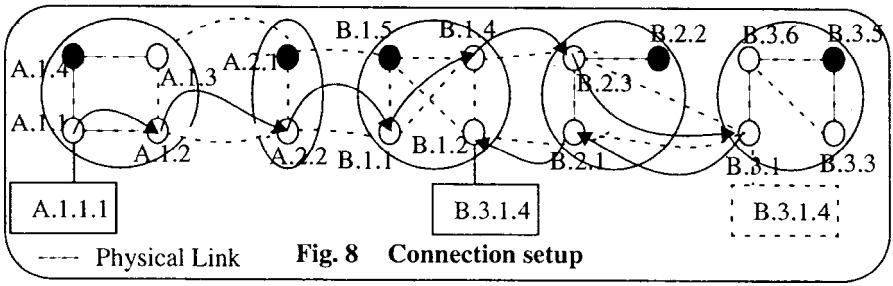

Fig. 8 Connection setup

B (at a node B.1.1, which is in the peer group B) since the reachability information in the nodes of peer group A indicate that the mobile is in its home peer group (B). Once the call setup message arrives at peer group B, it is routed efficiently to B.1.2, if *S,* the scope of limited reachability updates is 2, in which case all the nodes in the peer group B have the correct reachability information. Note that, if the call setup arrives at a switch before it is updated with the correct reachability information about the mobile, then the call is set up to the old location, which, then forwards the call to the current location. For example, consider the situation in which the mobile B.3.1.4 has just moved from B.3.1 to B.1.2. If the call setup arrives at switch B.1.1 before B.1.1 is updated with reachability information for the mobile endpoint B.3.1.4, then switch B.1.1 may choose[1] {B.1.1, B.1.4, B.2, B.3} as the shortest path by which to reach peer group B.3 based on its current reachability information for mobile B.3.1.4 (which points toward B.3). In this case, the connection route will be inefficient as shown in Fig. 8 (with the arrows indicating the connection route).

If the *called mobile is outside the neighborhood of its home*, the call setup proceeds to the home of the called mobile, which then forwards the call towards the current location of the called mobile. For example, consider that *S=3,* and a call setup request to B.3.1.4 is generated by an endpoint attached to switch B.2.1. Since *S=3*, the B.2 nodes are not updated about the move of mobile B.3.1.4 from B.3.1 to B.1.2. Hence, the connection will be inefficiently routed to the home and then to the new location (from B.2.1 to B.3.1, back to B.2.1, and then to B.1.2). The above two examples illustrate that if a call setup request for a mobile arrives between the time instant a mobile moved and the time instant that the reachability update is complete, or if it is generated in a node outside the neighborhood of the called mobile and the called mobile is not in its home neighborhood, it may be set up on a circuitous path. If the called mobile is within its home neighborhood or is in the neighborhood of the calling party, then this location management scheme results in the shortest possible connection.

3.1.5 Route optimization

As shown in the examples of Section 3.1.4, connections may be inefficiently routed due to a lack of correct reachability information. Therefore, route optimization needs to be performed. Route optimization consists of two steps. First, a "switchover

1. The first node receiving the call setup message in each peer group determines the route of the connection through that peer group.

node" at which the connection is to be rerouted from the old path to the new path is found, and a new segment is set up from the route optimization-initiating switch to the switchover node. Second, using "Tail" signals, user data is switched over from the old path to the new path without loss of cell sequence [17, 18]. Details of the route optimization scheme are presented in [19].

3.2 Location Registers (LR) scheme

In this scheme, the PNNI routing protocol reachability information is disregarded for mobile endpoints. Instead, an explicit tracking and locating procedure is overlaid on a PNNI based network using location registers. The LR scheme architecture, the mobile tracking procedure and the mobile locating procedure are defined in the following subsections.

3.2.1 LR scheme architecture

Fig. 9 shows hierarchically-organized location registers (LRs). The switches are

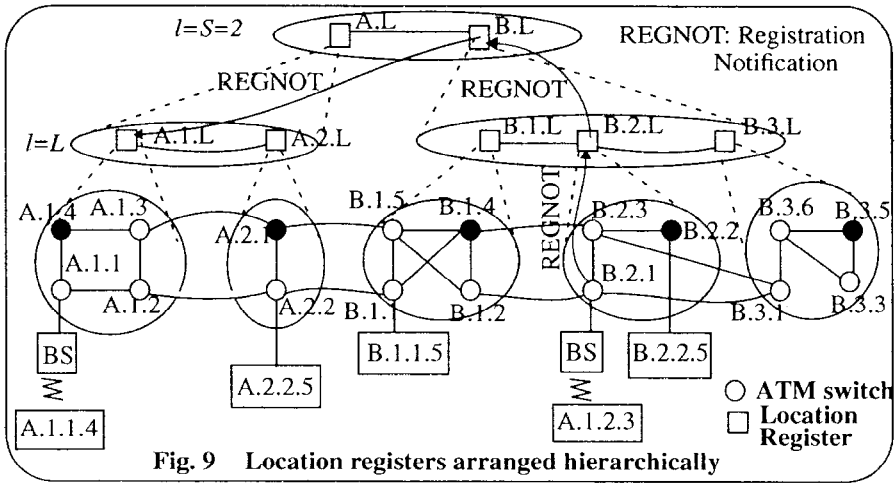

Fig. 9 Location registers arranged hierarchically

represented by circles. Location registers only exist from level *L* up to some level *S* (as explained in Section 2, we lop off the tree at level *S*). We make an assumption that each peer group is assumed to have one LR. This assumption can be relaxed and multiple LRs may be located in each peer group. This is effectively equivalent to creating a sublayer under the lowest layer of switches, and applying the same concept of allocating one LR per peer group of this new sublayer.

Location register A.1.L (we use the .L extension to avoid confusing this node with the A.1 logical node shown in Fig. 5 for the mobile PNNI scheme) is assumed to track all mobiles attached to switches within peer group A.1 (i.e., mobiles located at base stations connected to switches A.1.1, A.1.2, A.1.3 and A.1.4). Similarly, A.2.L is assumed to track all mobiles located at base stations connected to switches A.2.1 and A.2.2. A home LR is assigned to a mobile based on its permanent address, e.g. A.1.L

is the home LR of the mobile A.1.2.3.

The hierarchy of location registers helps localize mobile tracking and locating costs. However, if the hierarchy is carried to the topmost level ($l = 1$) as in the hierarchical scheme of [10], the processing requirements could be high. If computation costs are more than communication costs, it is more expensive to stop and process REGNOT (Registration Notification) or LOCREQ (Location Request) at each LR in the hierarchy, than to send one such request as a connectionless message to the home. Hence, we limit the hierarchy to level S and resort to the flat scheme approach of updating and/or querying the home LR of the mobile (see Section 2). However, if the home LR were to track the lowest level ($l = L$) LR currently tracking the mobile as in the flat scheme, the long-distance signaling costs of updating or querying the home LR would be high. Hence, the home LR only tracks the S^{th} level LR for each mobile, and only receives location queries if none of the LRs up to level S of the calling mobile's switch can respond to the query. The parameter S allows the LR scheme to be flexibly implemented as a flat structure, or as a rooted hierarchical tree, or as a mixed structure combining these extremes.

3.2.2 Mobile tracking

When a mobile *powers on*, the switch connected to its base station receives a power-on registration message. This switch sends a REGNOT (Registration Notification) to its LR at level L. This, in turn, causes REGNOTs to be generated to the ancestor LRs upstream up to an LR at level S. If the visiting switch is distinct from the home switch, the LR at level S sends a REGNOT to notify the home LR of the mobile that the mobile is currently in its domain. REGNOTs are sent as connectionless packets using the ATM NSAP address of the mobile as the destination [16]. The home LRs of all mobiles visiting at switches other than their home switch track the S^{th} level LR of the mobile in its current location. Power-off registrations are handled in a similar manner as power-on, whereby LRs up to level S are informed that a mobile powered off, and if the mobile was visiting (away from home), its home LR is also notified.

Next consider *zone-change registrations*, which are generated as mobiles move from a base station connected to one switch to a base station connected to another switch. The hierarchy of LRs is exploited to limit the propagation of registration information for such movements. On receiving the registration message, the new switch sends a REGNOT message to its level L LR. This, in turn propagates the REGNOT message upwards up to the LR which is a common ancestor of the LR corresponding to the old switch and the LR corresponding to the new switch, or up to level S, whichever is lower in the hierarchy (higher in numerical value). A message is sent by the new switch to the old switch, informing the old switch about the movement of the mobile. The old switch then generates a REGCANC (Registration Cancellation), which is sent to its level L LR. This, in turn is propagated upwards, cancelling the old information in the LRs. If the S^{th} level LR tracking the mobile changes due to the move, then the home LR of the mobile is updated.

For example, if the mobile A.1.1.4, shown in Fig. 9, moves to a base station connected to switch A.1.2, only LR A.1.L needs to be updated. One cancellation is

required at the switch. On the other hand, if it moves from switch A.1.1 to a switch A.2.2, then REGNOTs are sent to LR A.2.L from switch A.2.2, and subsequently from LR A.2.L to LR A.L, since LR A.L is the common-ancestor LR of the LRs corresponding to the old and new switches. Since the LR at level S (A.L) did not change, there is no REGNOT sent to the home LR of the mobile. However, a cancellation message is sent from A.2.2 to A.1.1, which in turn generates a REGCANC from A.1.1 to A.1.L. Finally, if the mobile moves from switch A.2.2 to switch B.1.1, REGNOTs propagate from switch B.1.1 to LR B.1.L, and then to LR B.L. Since there is a change in the S^{th} level LR tracking the mobile, LR B.L notifies home LR A.1.L. In addition, a REGCANC is generated by B.1.1 to A.2.2, which passes upwards to LR A.2.L, and then to A.L.

3.2.3 Mobile Locating

To find a mobile prior to call setup, a chain of location registers is traced. The length of the chain depends on the location of the calling party and the current location of the mobile. The called party's switch begins by checking to see if the called mobile is located at a base station in its domain. If so, it completes the call without generating any LOCREQs (Location Requests).

If the called mobile is not located at a base station within its domain, it generates a LOCREQ to its LR. Such requests are forwarded upwards in the hierarchy of LRs. If an LR at some level k has information (pointer to a child LR) regarding the location of the mobile, then it sends LOCREQs downwards toward the called mobile's current location. The location query will be resolved by the level L LR of the switch at which the called mobile is located, and the response will be sent directly to the calling party's switch.

If, however, none of the LRs, from the level L LR of the calling party's switch to the S^{th} level LR know the location of the called mobile as shown in Fig. 10, the S^{th} level LR sends a LOCREQ to the home LR of the called mobile. The called mobile's home switch will then forward this message to the home LR of the mobile. Since the home LR tracks the S^{th} level LR of its mobiles, it forwards the LOCREQ to the S^{th} level LR tracking the mobile in its current location. This LR generates downward LOCREQs according to the information it has about the called mobile. The LOCREQ will reach the level L LR of the called mobile's switch. The response is sent directly from this LR to the calling party's switch as shown in Fig. 10. The address tunnelling concept of the flat scheme described in Section 2 is also used in the LR scheme.

As *examples*, we consider call originations from *three* endpoints, B.2.2.5, B.1.1.5 and A.2.2.5, all targeted at mobile A.1.2.3 (see Fig. 9). In the *first example*, when switch B.2.2 generates a LOCREQ for mobile A.1.2.3 to its LR B.2.L, the latter can immediately respond since the called mobile A.1.2.3 is located within its region. In the *second example*, switch B.1.1 sends the LOCREQ (in response to the call setup request from its endpoint B.1.1.5 to mobile A.1.2.3) to its LR B.1.L. Since it has no pointer regarding this mobile, it simply generates a LOCREQ to the higher-level LR B.L. This register has a pointer indicating that B.2.L is tracking the mobile. Hence a LOCREQ is sent to this LR. Since B.2.L is the level L LR for the called mobile, it responds indicating that the mobile is located at switch B.2.1. This response is sent

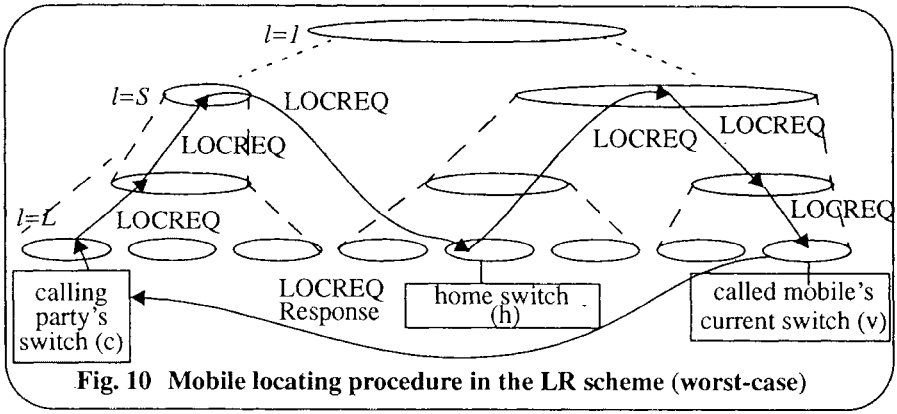

Fig. 10 Mobile locating procedure in the LR scheme (worst-case)

directly to switch B.1.1 (instead of retracing the pointers backwards) allowing it to initiate call setup to the called mobile's switch. In the *third example*, where endpoint A.2.2.5 generates the call setup to mobile A.1.2.3, the LOCREQ sent by switch A.2.2 traverses the chain of LRs, A.2.L and A.L. Since neither of these LRs have information on the location of the called mobile and $S = 2$, A.L sends a LOCREQ to the home LR of the called mobile A.1.L. This LR forwards the LOCREQ to LR B.L, since each home LR tracks the S^{th} level LR of its mobiles. LOCREQs are then sent downwards from B.L to LR B.2.L, which responds with a temporary address for the mobile indicating that the mobile is located at switch B.2.1.

4. DISCUSSION

A comparative study of the tracking and locating costs (move and search costs) of the two schemes using analytical models is presented in [6]. Based on this study, we provide a qualitative *comparison of the two schemes* in this section. We also discuss *the effect of key parameters of these algorithms*, such as S, the reachability update limiting scope, which is also the highest-level of the hierarchy of location registers in the LR scheme, h, the cost of "long-distance" signaling (needed to characterize the cost of setting forwarding pointers at the home and old switch in the mobile PNNI scheme and updating or querying the home location registers in the LR scheme, if outside the scope S of the sender), and CMR (Call-to-Mobility Ratio).

Tracking cost in the mobile PNNI scheme depends on the cost of updating reachability information, and the cost of sending messages to the home and the old locations of the mobile. Tracking cost in the LR scheme depends on the cost of updating location registers, and the cost of sending messages to the home and the old locations of the mobile. Both the cost to update and the cost to send messages depend on the relative distances between the old, new, and the home locations of the mobile, and the value of S.

Search cost in the mobile scheme is not easily quantified since there is no explicit mobile location phase in the mobile PNNI scheme. However, since some of the connections are first established on non-optimal routes, we consider the number of extra hops needed to route forwarded connections as the "search cost" associated with the mobile PNNI scheme. Thus, the search cost depends on the relative positions of the

calling party, called party, and the called party's home. If the mobile (called party) is with in scope *S* from its home or the calling party is with in scope *S* from the mobile, then the mobile PNNI search cost is zero (we expect this to hold for majority of calls). As *S* decreases (lower in value) the extent to which the reachability information is spread increases, therefore, the mobile PNNI search cost decreases. In the LR scheme, there is an explicit mobile location phase. The search cost in the LR scheme depends on the relative positions of the calling party, called party, and the called party's home. The search cost is small for all short distance calls, irrespective of the location of the mobile (home or remote location).

Typically, the search cost of the LR scheme is more than the search cost of the PNNI scheme. The intuitive explanation for this trend is that the LR scheme incurs a search cost for all calls, whereas the PNNI scheme incurs a search cost only for a small fraction of calls. In general, the tracking cost in the LR scheme is less compared to the tracking cost of the mobile PNNI scheme. The intuitive explanation for this trend is that in the LR scheme the number of databases to be updated is less compared to the number of databases in the mobile PNNI scheme. From these trends, it can be argued that the *LR scheme performs better at lower CMRs,* and the *mobile PNNI scheme perform better at higher CMRs.*

A *second level of comparison* is to understand the behavior of the two schemes relative to increasing CMR and *S*. As the value of *S* increases (the extent of spread of the reachability updates decreases), the tracking cost in the PNNI scheme tends to decrease and the locating cost tends to increase. Thus, in the mobile PNNI scheme, at higher CMRs, a low value of *S* should be selected, while at low CMRs, high values of *S* should be chosen. This is illustrated in Fig. 11. In the LR scheme, as the value of

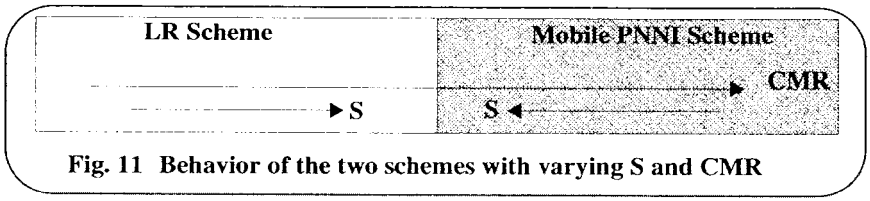

Fig. 11 Behavior of the two schemes with varying S and CMR

S is increased (there are fewer levels of hierarchy), the search cost decreases and the tracking cost increases. Thus, in the LR scheme, at higher CMRs, a high value of *S* should be selected, while at low CMRs, low values of *S* should be chosen.

The trend shown in Fig. 11 is dependent on the value of the parameter h, which is the cost of "long-distance" signaling. The trend shown in Fig. 11 is quantitatively determined in [6] for a value of *h* that was determined as being of "medium range." We discuss the effect of different values of *h* on the move and search costs of the two schemes in the following paragraphs.

Fig. 12 and Fig. 13 show the trend in the variation of the average move and search costs in the two schemes, at different ranges of h (arrows pointing upward indicate an increase in cost). These trends provide important information in helping us select values of S, at which to operate the two schemes, for a given value of h. For the mobile PNNI scheme, Fig. 12 shows that at "low and medium values" of h, the average move cost decreases with increasing S. However, at high values of h, the move cost decreases with increase in S up to a value, S_{max}, as shown in Fig. 12, beyond

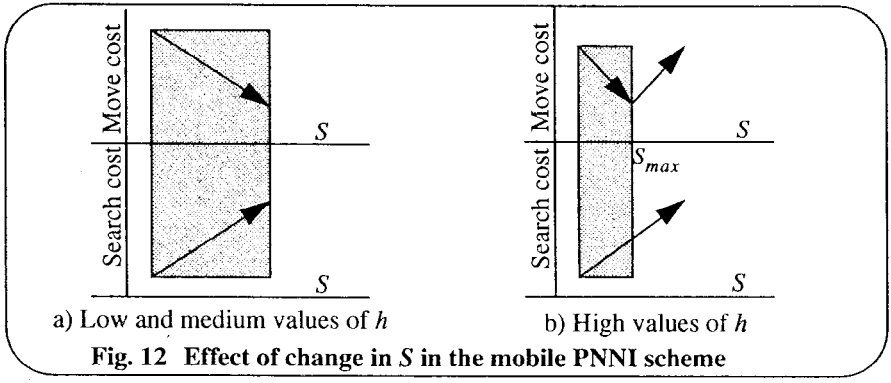

Fig. 12 Effect of change in S in the mobile PNNI scheme

which it increases. At high values of S, the mobile moves out of scope S more frequently than at low values of S. Since the home is updated, whenever the mobile moves out of scope S from its current location, at high values of S, h becomes a significant parameter. The search cost in the mobile PNNI scheme is not effected by the value of h. Since the mobile PNNI scheme search cost increases monotonically with increasing S, to minimize the average total cost, for the mobile PNNI scheme, the values of S should be chosen such that $S \leq S_{max}$.

The LR costs show a slightly different trend. At "low" values of h, the LR scheme experiences decreasing average move and search costs with increasing S, at "high" values of h, both costs increase with increasing S, and at "medium" values of h, both costs first decrease and then increase, as shown in Fig. 13. The reason for

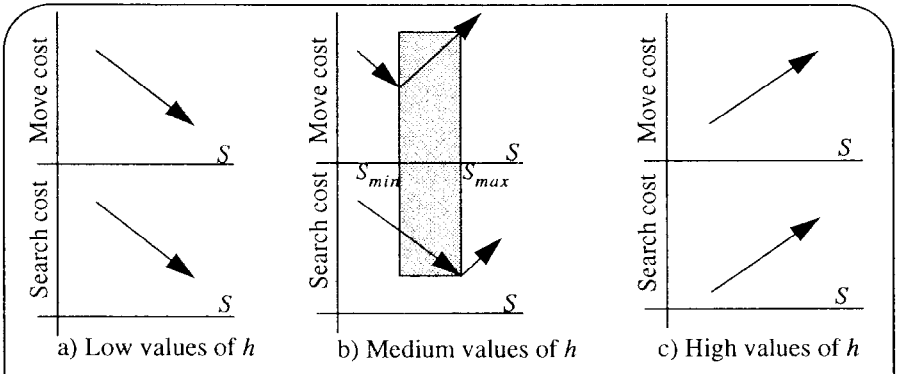

Fig. 13 Effect of change in value of S on average costs in the LR scheme

such behavior is that if h is small, the LR scheme should be operated more like the "flat" scheme of Section 2, by choosing a large S. In other words, all nodes know the home LRs of mobiles and directly send registrations and location queries to the home LRs. At large values of h, the LR scheme should be operated more like the "hierarchical" scheme of Section 2, by making S equal to 1, since the best result (lowest average total cost) is obtained at the smallest value of S. For the medium range of h, the optimal value of S depends upon the CMR. In this range, the average move cost decreases with increasing S up to S_{min}, and the search cost decreases up to a value of $S = S_{max}$. The minimum value of average total cost is obtained for some value of S that lies between S_{min} and S_{max}.

In summary, there are three important parameters, *S*, the reachability update limiting scope, which is also the highest-level of hierarchy in the LR scheme, *h*, the cost of "long-distance" signaling, and CMR (Call-to-Mobility ratio). As to which location management scheme incurs a lower average total cost, depends on these three parameters. Typically, at *low values* of *h*, operating the LR scheme at high values of *S*, leads to a lower average total cost than the PNNI scheme. On the other hand, if the cost of *long-distance signaling h is high*, either LR scheme or the mobile PNNI scheme could lead to minimal average total cost, provided the correct value of *S* is selected. For the mobile PNNI scheme, this depends on the CMR expected, but in the LR scheme, one needs to select a low *S* (preferably $S = 1$). Finally, if *h is in the medium range* (which we expect will be the range of operation), there will be a breakpoint CMR below which the LR scheme will perform better, and above which the PNNI mobile scheme will incur lower costs. A significant point to note is that for a number of cases, for example, when mobiles are located close to their home locations (which we expect will be a high-percentage), or if the calling party is close to the visiting location of the called mobile, the mobile PNNI scheme incurs a zero search cost. This leads to the mobile PNNI scheme performing better in most regions of operation expected in low-tier (i.e., slowly moving users) PCS applications.

5. CONCLUSIONS

This paper presented two mobile location management algorithms for ATM networks based on the PNNI (Private Network-to-Network Interface) standard. The first solution is called the *mobile PNNI scheme* because it builds on the PNNI routing protocol. It uses limited-scope (characterized by a parameter *S*) reachability updates, forwarding pointers (setting and clearing of these pointers occur at a cost *h*) and a route optimization procedure. The second solution is called the *LR (Location Registers) scheme* because it introduces location registers (such as the cellular home and visitor location registers) into the PNNI standards-based hierarchical networks. This scheme uses a hierarchical arrangement of location registers with the hierarchy limited to a certain level *S*. It also requires the update of home and old location registers at a cost *h*.

Discussion of the effect of key parameters on the average move, search, and total costs, of these two schemes for different values of the CMR (Call-to-Mobility Ratio) is presented. This discussion provides guidelines for selecting the critical parameters of these algorithms. It showed that at low CMRs, the LR scheme performs better than the mobile PNNI scheme. We also observed that the two schemes show a contrasting behavior in terms of the value to be used for the parameter *S* to achieve the least average total cost. For the mobile PNNI scheme, the parameter *S* should be high at low CMRs (within the range in which the mobile PNNI scheme should be used), and low at high CMRs. However, in the LR scheme, the parameter *S* should be low at low CMRs (within the range in which the LR scheme should be used), and high for high CMRs. These observations are made for a region of operation in which *h*, the cost of setting forwarding pointers and updating distant LRs, is of medium value. If this cost is low, the LR scheme always outperforms the mobile PNNI scheme. For other ranges of *h*, the scheme selected, and the S at which the scheme is operated, depends on the CMR.

References

[1] R. Jain and Y-B. Lin, "An auxiliary user location strategy employing forwarding pointers to reduce network impacts of PCS," *ACM/Baltzer Wireless Networks Journal*, Vol. 1, No. 2, pp. 197-210, July 1995.

[2] B. Auwerbuch and D. Peleg, "Online tracking of mobile users," *Journal of the Association for Computing Machinery*, Vol. 42, No. 5, pp. 1021-1058, September 1995.

[3] EIA/TIA IS-41 Rev. C, "Cellular Radio Telecommunications Intersystem Operations," November 1995, TIA/EIA PN-2991.

[4] M. Mouly and M. B. Pautet, "The GSM System for Mobile Communications," 49 rue Louise Bruneau, Palaiseau, France, 1992.

[5] The ATM Forum Technical Committee, "Private Network-Network Specification Interface v1.0 (PNNI 1.0)," March 1996, af-pnni-0055.000.

[6] M. Veeraraghavan and G. Dommety, "Mobile Location Management in ATM Networks," *IEEE Journal on Selected Areas in Communications*, under review, August 1996. (Short form appeared in Proc. of ICC '97, June 8-12, Montreal, Canada).

[7] J. S. M. Ho and I. F. Akylidiz, "Local Anchor Scheme for Reducing Location Tracking Costs in PCNs," *IEEE/ACM Transactions on Networking*, Vol. 4, No. 5, October 1996, pp. 709-725.

[8] M. Veeraraghavan, T. F. La Porta and R. Ramjee, "A Distributed Control Strategy for Wireless ATM Networks," *ACM/Baltzer Wireless Networks Journal*, pp. 323-339, 1995.

[9] G. Dommety, M. Veeraraghavan, and M. Singhal, "Flat Location Management Scheme for PCNs," to appear in *Proc. of IEEE International Conference on Universal Personal Communications (ICUPC) 1997*, October 12-16, San Diego, California.

[10] J. Z. Wang, "A Fully Distributed Location Registration Strategy for Universal Personal Communication Systems," *IEEE Journal on Selected Areas in Communications*, vol. 11, pp. 850-860, August. 1993.

[11] C. Eynard, M. Lenti, A. Lombardo, O. Marengo, and S. Palazzo, "A Methodology for the Performance Evaluation of Data Query Strategies in Universal Mobile Telecommunication Systems (UMTS)," *IEEE Journal on Selected Areas in Communications*, Vol. 13, No. 5, pp. 893-907, June 1995.

[12] Charles Perkins, "IP Mobility Support," RFC 2002, October 1996.

[13] David B. Johnson and Charles Perkins, "Mobility support in IPv6," Internet Draft, draft-ietf-mobileip-ipv6-02.txt, November 1996. Work in progress.

[14] David B. Johnson and Charles Perkins, "Route optimization in Mobile IP," Internet-Draft, draft-ietf-mobileip-optim-04.txt, February 1996. Work in progress.

[15] The ATM Forum Technical Committee, "ATM User-Network Interface (UNI) Signaling Specification Version 4.0," January 1996, ATM Forum/95-1434R9.

[16] M. Veeraraghavan, P. Pancha and G. Dommety, "Connectionless ATM using an ATM Switch Router," *Proc. of European Conference on Multimedia Applications, Services and Techniques (ECMAST) 1997*, May 21-23, Milan, Italy.

[17] M. Veeraraghavan, M Karol and K.Y. Eng, "Mobility and Connection Management in a Wireless ATM LAN," *IEEE Journal on Selected Areas in Communications*, Vol. 15, No. 1, pp. 50-68, January 1997.

[18] G. P. Pollini and K. S. Meier-Hellstern, "Efficient Routing of Information Between Interconnected Cellular Mobile Switching Centers," *IEEE/ACM Transactions on Networking*, Vol. 3, No. 6, pp. 765-774, December 1995.

[19] G. Dommety, M. Veeraraghavan, M. Singhal, "Route Optimization in Mobile ATM Networks," to appear in *Proc. of ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom)* 1997, September 26-30, Budapest, Hungary.

PHONE NUMBER TRANSLATION DELAY IN PCS SYSTEMS WITH ATM BACKBONES

Ravi Jain

Bell Communications Research
331 Newman Springs Rd, Red Bank, NJ 07701
rjain@bellcore.com

ABSTRACT

Future wired networks with an ATM backbone will need to support PCS and wireless subscriber services. One of the key functions required to support PCS and wireless subscribers with non-geographic phone numbers (NGPN) will be the ability to efficiently identify which Home Location Register (HLR) database serves the subscriber. (Note that the same functionality is also needed to serve subscribers with portable phone numbers.) In a previous paper we have presented a scheme for NGPN translation based upon distributed, dynamic hashing. The scheme uses a hash function in the Visitor Location Registers (VLR) and a set of distributed Translation Servers (TS) which store the NGPN-to-HLR mapping.

In this paper we present a preliminary investigation of the additional call setup delay introduced by the NGPN translation scheme we have proposed. We develop a queuing network model for both the one-stage and two-stage versions of our NGPN translation scheme, and we investigate the impact of two key aspects of our scheme, namely the use of hashing and of caching. A poor choice of hash function can lead to imbalanced loading at the TS. We quantify the impact of this imbalance for an example scenario. We also show that caching at the VLR can substantially reduce the mean translation delay. Finally we consider the use of the two-stage scheme and a second-level cache when the translation load increases.

1 INTRODUCTION

One of the key functions required to support Personal Communications Services (PCS) and wireless subscribers in PCS systems is mobility management, i.e.,

determining the location of a mobile user so that calls can be delivered to and from that user. In current and proposed standards for PCS and cellular systems a set of databases, called Home Location Registers (HLR) and Visitor Location Registers (VLR), is used to maintain the information about the current location of a user, and this information is consulted or updated whenever a user moves across geographical regions called Registration Areas (RA), or when a call is to be delivered to and from a user. For a tutorial on mobility management procedures, see [7, 10].

This paper considers an issue that arises when two important trends in the development of future PCS systems converge: (1) the use of high-speed ATM networks for the fixed network backbone which supports mobile users, and (2) the rise in the number of PCS users with non-geographic phone numbers (NGPN), i.e., phone numbers which do not indicate the home geographical region or the service provider of the mobile user. In this scenario, as part of the mobility management procedures, it is important that the system be able to efficiently translate the NGPN of a PCS subscriber to the identity of the HLR which maintains the current location of the subscriber. In previous work [5, 6] we have presented a scheme for NGPN translation which uses a set of distributed Translation Servers (TS). In this paper we develop a model for estimating the mean delay introduced by the NGPN translation scheme we have proposed.

We begin by describing, in this section, the NGPN translation problem and summarizing the solution we have proposed. In sec. 2 we describe our performance model for the one-stage and two-stage versions of our NGPN translation scheme, and in sec. 3 we demonstrate its application for an example scenario. Finally, we end with some conclusions.

1.1 Background on NGPN Translation

The problem of NGPN translation is to determine the identity of the signaling network database which serves a Personal Communications Services (PCS) subscriber, when the only relevant information available is a non-geographic phone number (NGPN).

Currently, fixed telephone subscribers are assigned a geographic phone number, which contains enough information to determine how the signaling messages required to set up a call to the subscriber are to be routed through the signaling network. For proposed PCS systems, subscribers will be assigned NGPNs (e.g.,

1-500-XXX-XXXX), which do not contain this information; a process called Global Title Translation (GTT) has been designed for this purpose [2, 4]. GTT is executed at signaling switches called Signaling Transfer Points (STPs), and essentially translates a subscriber's NGPN to the identity of the Home Location Register (HLR) database which serves that subscriber. For future PCS systems in which the wired backbone is an ATM network, however, signaling traffic will be likely to use the same physical transport as the user data traffic [3]. Thus STPs may not be used for signaling and GTT cannot be performed [11].

NGPN translation is required in three situations:

1. When a PCS subscriber with a NGPN is called (by a fixed or PCS subscriber), its HLR ID is required to set up the call.
2. In some implementations, NGPN translation may be required when a PCS subscriber crosses Registration Areas (RAs) served by different VLRs, since the ID of its serving HLR needs to be determined in order to update the subscriber's location information.
3. In some implementations, NGPN translation may also be required when a PCS subscriber originates a call, in order to obtain authentication and service profile information.

We will simplify the discussion by assuming that for all three cases above, the NGPN is presented to a VLR, which has the burden of performing a translation or obtaining a translation from other entities in the network, although it is understood that for call delivery from wireline phones to PCS subscribers, the operations may in fact have to be performed at a serving Service Control Point (SCP) [1].

Clearly, the process of NGPN should be *fast* (in order to reduce overall call setup time), *efficient* (in terms of signaling network and database loads), and *scalable* (as the translation load increases.) Another key requirement is that the process allow the NGPN-to-HLR mapping for any user to be changed easily and efficiently, since the mapping may need to be changed any time the user changes service providers, or if the service provider changes the HLR serving that user (for performance or administrative reasons, e.g. if the user moves permanently from one region of the country to another.) This requirement for *efficient mapping modification*, along with the requirement for scalability, are the main reasons that naive schemes for performing the NGPN translation do not suffice, necessitating the scheme that we have developed.

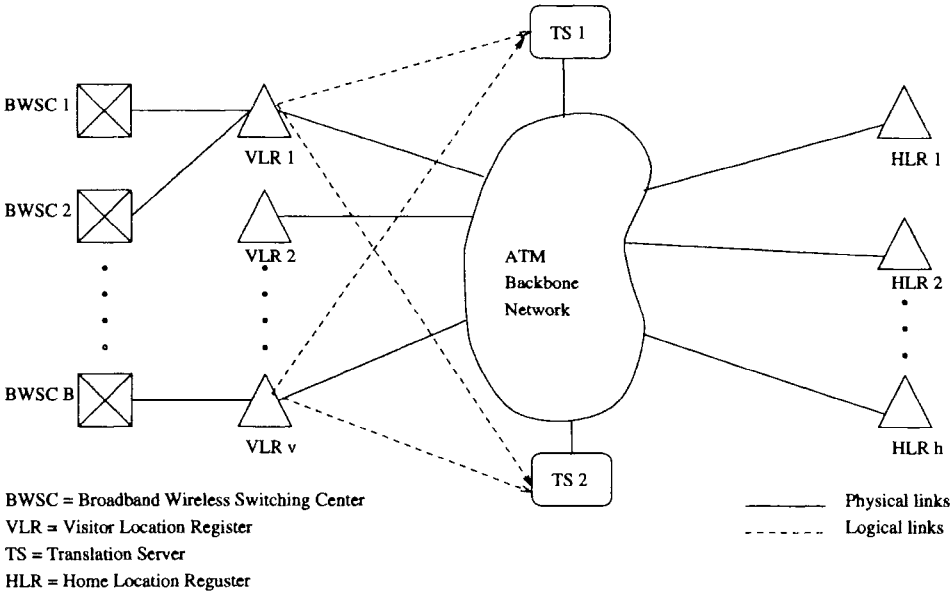

Figure 1 Architecture for the (One-Stage)NGPN Translation Scheme

1.2 The proposed NGPN translation scheme

In order to keep this document self-contained, the scheme presented in [5, 6] is briefly described here.

One-stage NGPN Translation Scheme

We first describe the one-stage version of the scheme [5]; see Fig. 1.

1. When any of the situations requiring NGPN translation occur, the non-geographical PN is presented to a switch; depending upon the architecture deployed, this may be a Mobile Switching Center (MSC) or Broadband Wireless Switching Center (BWSC), or a Service Switching Point (SSP).

2. The switch forwards the NGPN to the VLR serving that switch. The VLR performs a hash function upon the binary representation of the NGPN, to obtain a value f(NGPN), where f is the hash function. This specifies

the ID of an entity called a Translation Server (TS). Translation servers are entities not present in current and proposed PCS architectures; they were introduced for the purpose of NGPN translation. Note that TSs are logical entities; physically they may be implemented as databases and be co-located with HLRs.

3. The VLR launches a query to the TS specified by f(NGPN), passing it the value NGPN. The TS contains a table mapping the NGPN to the ID of the HLR serving that NGPN.

4. The TS responds to the VLR by returning the HLR ID.

5. The VLR uses the HLR ID to continue with the registration, call delivery, or call origination signaling operations as usual.

The VLR can maintain a cache of NGPN translations to avoid querying the TS. Thus when presented with a NGPN for the first time, the VLR performs a hash and queries the indicated TS to obtain the ID of the serving HLR. It then stores the NGPN-to-HLR mapping for that NGPN in its cache. If presented with the same NGPN a second time, the VLR can search its cache first. If the mapping is found (a *cache hit*), a hash and a query to the TS is avoided; otherwise (a *cache miss*), the VLR performs a hash and queries the TS as usual.

A simple example of the hash function, f, is the function *even*(), which returns 0 if the argument is even and 1 otherwise. Obviously, this function can only be used if there are only two TSs. In addition, if the way in which NGPNs are chosen is non-uniform (e.g., a disproportionate number of subscribers request phone numbers ending in 0), the load on the two TSs will not be balanced; similarly, if the service provider assigns NGPNs to new subscribers using some administrative process which somehow introduces some non-uniformity, the load on the TSs will not be balanced. The effect of load imbalance at the TSs will be a factor we will consider in our model.

As we have discussed in [6], we believe the one-stage NGPN translation scheme offers speed and efficiency, and the use of indirection (storing the NGPN-to-HLR mapping in the TS) offers efficient mapping modification. In order to improve scalability, we have proposed the use of a *dynamic hashing* method which can be used if necessary to increase the number of TS as the translation load grows.

Two-Stage NGPN Translation Scheme

We have also presented a two-stage version of our NGPN translation scheme in order to improve scalability (see Fig. 2). The two-stage translation scheme is identical to the one-stage scheme described above as far as step 2, i.e., the VLR applies the hash function f and determines the identity of a TS. However, step 3 is modified by introducing a second stage of TS, as follows:

3(a) The VLR launches a query to the TS specified by $f(NGPN)$, passing it the value NGPN.

3(b) This *first stage* TS applies another hash function, $g(NGPN)$, and obtains the identity of a *second stage* TS.

3(c) The first-stage TS launches a query to the second-stage TS, passing it the NGPN.

3(d) The second-stage TS contains a table mapping the NGPN to the ID of the HLR serving that NGPN.

Step 4 for the two-stage scheme is identical to that for the one-stage scheme, except that it is the the second-stage TS which responds to the VLR by returning the HLR ID. Similar to the VLR cache, the first stage TS can maintain a *TS cache* for speeding up the translation process.

2 PERFORMANCE MODEL

We have developed a queuing network model for estimating the delay entailed by our proposed NGPN translation scheme. In this section we briefly describe the model and its assumptions.

The model is developed for the two-stage NGPN translation scheme; by an appropriate choice of parameters it models the one-stage scheme also. A schematic diagram of the model is shown in Fig. 3, and a list of the parameters is shown in the first two columns of Table 1.

The model focuses on the path taken by a single translation request as it arrives at a given VLR, which we call VLR_1, and travels to a selected TS, TS_1, and its second- stage TS, $TS_{1.1}$. Referring to Fig. 3, translations arrive at VLR_1

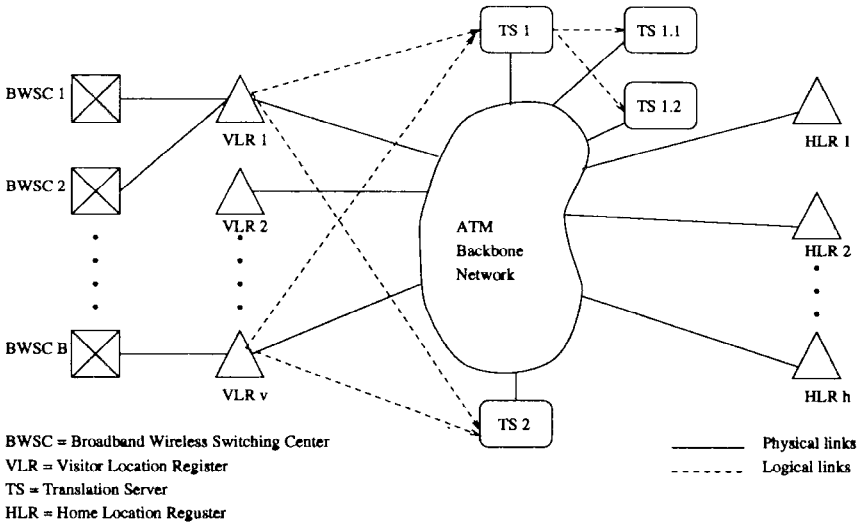

Figure 2 Architecture for the Two-stage NGPN Translation Scheme

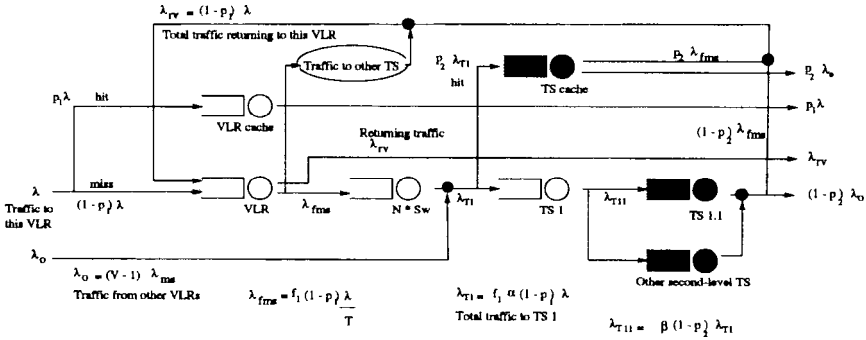

Figure 3 Model of NGPN translation delay. The shaded queues are omitted for the one-stage version of the NGPN scheme.

at a rate λ; a fraction p_1 enjoy cache hits, i.e., are served immediately by the cache and exit the system, while the remaining $(1 - p_1)\lambda$ require further processing. Recall that a hash function is applied at the VLR, so that this remaining traffic is (ideally) divided among all the T TSs in the system, i.e.,

Parameter	Symbol	Value
Direct translation load to this VLR	λ	1 - 30 trans/sec.
Service time for VLR	$1/\mu_v$	5 ms
Background load to this VLR	λ_{bv}	$.4\mu_v$
Service time for VLR cache	$1/\mu_{vc}$	1.25 ms
Cache hit ratio at VLR	p_1	0, 0.1, 0.25, 0.5
Number of TSs	T	3
Number of VLRs	V	50
Load imbalance factor	f_1	1.0 - 3.0
Avg. number of ATM switches from VLR to TS	N	3
Service time for SVC setup, per switch	$1/\mu_{sw}$	20 ms
Background load to each switch	λ_{bs}	$.4\mu_{sw}$
Cache hit ratio at first-stage TS	p_2	0, 0.1, 0.25, 0.5
Service time at first-stage TS (one-stage)	$1/\mu_{t1}$	5 ms
Service time at first-stage TS (two-stage)	$1/\mu_{t1}$	1.25 ms
Ratio of first-stage to second-stage TS	$1/\beta$	2
Service time at second-stage TS	$1/\mu_{t11}$	5 ms

Table 1 Parameters of the NGPN translation model

with no load imbalance, the traffic from VLR_1 to each TS is

$$\lambda_{ms} = \frac{(1 - p_1)\lambda}{T} \tag{1}$$

In order to model the effect of load imbalance, the traffic from VLR_1 to a selected TS, TS_1, is given in terms of a load imbalance factor f_1, where $1 \leq f_1 \leq T$. Thus, the traffic from VLR_1 to TS_1 is $\lambda_{fms} = f_1\lambda_{ms}$, and the traffic from VLR_1 to each of the other TSs is

$$\lambda_{oms} = \frac{(T - f_1)\lambda_{ms}}{T - 1} \tag{2}$$

Observe that this is a conservative model of load imbalance since only the load from a single VLR is assumed to be split non-uniformly.

To keep Fig. 3 intelligible we do not show the model for traffic to the other TSs; it is simply indicated in the oval marked “Traffic to other TS”. We assume that an ATM Switched Virtual Circuit (SVC) connection setup is required to forward the translation request, and on average, there are N switches between the VLR and the (first-stage) TS, each of which has a mean service time of

$1/\mu_{sw}$. The traffic into the switches connecting VLR_1 and $TS1$ is λ_{fms}, and the traffic into the switches connecting VLR_1 to the other TSs is λ_{oms}.

When the traffic from VLR_1 reaches the first-stage TSs, it is combined with the traffic from all the other $(V-1)$ VLRs in the system, which we denote λ_O. We assume that all the other VLRs also have an incoming load of λ each, and that the cache hit ratio is also p_1 at each. Then, assuming that the load at the other $(V-1)$ VLRs is perfectly balanced,

$$\lambda_O = (V-1)\lambda_{ms} \tag{3}$$

Thus the total traffic reaching the first-stage TS TS_1 from all the VLRs (including VLR_1) is $\lambda_{t1} = \lambda_{fms} + \lambda_O$, while that reaching the other the first-stage TSs is $\lambda_{ot1} = \lambda_{oms} + \lambda_O$.

The total traffic entering the first-stage TS can be satisfied directly by the cache with hit probability p_2. The remaining traffic is assumed to be equally divided amongst the second-level TSs, where $1/\beta \geq 1$ is the ratio of second-stage to first-stage TSs.

It is worth pointing out that the second stage is modeled as $1/\beta$ separate queues, rather than a single queue with $1/\beta$ servers. This is because the second-stage hash function g completely determines which second-stage TS an incoming translation request is sent to, i.e., incoming requests are not indistinguishable jobs which can be arbitrary sent to any of the $1/\beta$ queues.

After this second stage the traffic from the other VLRs, which is now $(1-p_2)\lambda_O$, returns to those VLRs. The traffic from VLR_1, which is $(1-p_2)\lambda_{fms}$ at TS_1 and $(1-p_2)\lambda_{oms}$ at the other TSs, returns to VLR_1; it is joined by the traffic which enjoyed a cache hit at the TS cache, so that the total returning traffic at VLR_1 is $\lambda_{rv} = (1-p_1)\lambda$. This returning traffic undergoes final processing at the VLR before exiting the translation process (and moving on to other processing, i.e. HLR querying.) Finally, we assume that the VLR and the switches both have background traffic of λ_{bv} and λ_{bs} respectively. For simplicity these background flows are not shown in Fig. 3.

For the purposes of this paper we consider that translation requests are Poisson arrivals and all the queues in the model are independent M/M/1 queues [8]. (Note once again that the second-stage TS is modeled as $1/\beta$ separate M/M/1 queues, rather than a single M/M/$(1/\beta)$ queue.) We can then calculate the total response time R, i.e, delay, of the translation process using well-known techniques [9, 8]. In the following, R_i is the residence time for the traffic from VLR_1 to TS_1 at entity i, where $i = v, vc, sw, t1, t1c, t11$ respectively represents

the VLR, VLR cache, ATM switch, TS1, TS1 cache, and TS1.1. R_{other} denotes the delay experienced by the traffic from VLR_1 which is serviced by the other TSs in the system. The binary variable *Stage* is set to 0 for a one-stage system and 1 for a two-stage system.

$$\begin{aligned} R &= R_v + R_{vc} + R_{sw} + R_{t1} + Stage * R_{t1c} \\ &\quad + Stage * R_{t11} + R_{other} \end{aligned} \tag{4}$$

where

$$R_v = \frac{2(1-p_1)}{\mu_v - (\lambda_{bv} + 2(1-p_1)\lambda)} \tag{5}$$

$$R_{vc} = \frac{p_1}{\mu_{vc} - p_1\lambda} \tag{6}$$

$$R_{sw} = \frac{(1-p_1)Nf_1}{T\mu_{sw} - (T\lambda_{bs} + (1-p_1)f_1\lambda)} \tag{7}$$

$$R_{t1} = \frac{\frac{(1-p_2)\lambda_{fms}}{\lambda}}{\mu_{t1} - (1-p_2)(\lambda_{fms} + \lambda_O)} \tag{8}$$

$$R_{t1c} = \frac{\frac{p_2\lambda_{fms}}{\lambda}}{\mu_{t1c} - p_2(\lambda_{fms} + \lambda_O)} \tag{9}$$

$$R_{t11} = \frac{\frac{(1-p_2)\lambda_{fms}}{\lambda}}{\mu_{t11} - \beta(1-p_2)(\lambda_{fms} + \lambda_O)} \tag{10}$$

The delay experienced by the traffic from VLR_1 which is serviced by the TSs other than TS_1 is given by

$$\begin{aligned} R_{other} &= \sum_{i=2}^{T} (R_{osw}(i) + R_{ot1}(i) + Stage * R_{ot1c}(i) + Stage * R_{ot11}(i)) \\ &= (T-1)(R_{osw} + R_{ot1} + Stage * R_{ot1c} + Stage * R_{ot11}) \end{aligned} \tag{11}$$

In Eq. 11 R_j is the delay at entity *j*, and $j = osw, ot1, ot1c, ot11$ represents the other switches, other first-stage TSs, other second-level cache, and other second-level TSs, and by "other" we mean the entities not associated with TS_1. We can derive expressions for the terms in R_{other} in a manner similar to those in Eq. 4.

3 EFFECT OF CACHING AND LOAD IMBALANCE

In this section we illustrate the use of the model described above for an example scenario. We focus on investigating the effect of caching and translation load imbalance. We use the values of the parameters given in Table 1 and plot the mean delay given by Eq. 4.

The translation workload is calculated assuming a low-tier PCS system, with small cells. We assume n active users per cell (base station), b base stations per base station controller, B base station controllers per Broadband Wireless Switching Center (BWSC), v BWSC per VLR, and c call originations and terminations per user during the busy hour. The translation traffic due to registrations is ignored since it is much less than that due to call originations and terminations [5]. Then the total translation workload per VLR is $\lambda = nbBvc$. *For* this paper we will first consider an example PCS system with $n = 15$, $b = 96$, $B = 8$, $v = 1$ and $c = 3$ calls/hour, so that $\lambda = 9.6$ trans/sec during the busy hour. We will plot the total translation delay for the range $1 \leq \lambda \leq 15$ trans/sec. We will later also consider a scenario where PCS penetration has increased and $n = 30$, so that $\lambda = 19.2$ trans/sec during the busy hour; for this case we will plot the delay in the range $1 \leq \lambda \leq 30$ trans/sec. As a rule of thumb, it is desirable that the mean delay due to translation alone be about 0.5 second or less.

3.1 One-stage version

To model the one-stage version using Eq. 4, we set *Stage* = 0, $p_2 = 0$, $\beta = 1$ and arbitrary positive values for $1/\mu_{t11}$ and $1/\mu_{t1c}$.

In Fig. 4 we plot the mean delay as a function of the translation load, assuming no caching at the VLR ($p_1 = 0$), the service time at TS1 is $1/\mu_{t1} = 5ms$, the load imbalance factor f_1 is varied, and the remaining parameters are as in Table 1. There is a sharp increase in the total delay R around $\lambda = 11$ trans./sec.. This occurs because the first-stage TS, TS_1, becomes saturated. The load imbalance factor is insignificant for $\lambda < 8$ trans/sec.; this is because the load imbalance at TS_1 results in higher load at TS_1 but lower load at the other TSs, so that the increased delay at TS_1 due to load imbalance is mitigated by the reduced delay at the other TSs. For $\lambda > 8$ trans./sec, the load imbalance factor begins to have some noticeable effect on the overall delay.

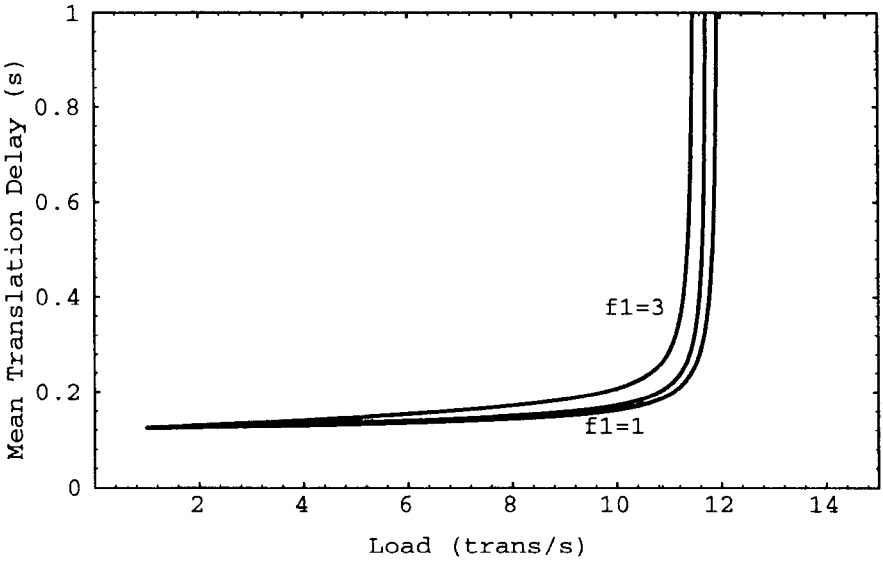

Figure 4 One-stage version: Mean translation delay for $p_1 = 0$ (no cache at VLR) and $1/\mu_{t1} = 5ms$. The load imbalance factor f_1 is varied; the three curves are for $f_1 = 3.0, 2.0, 1.0$ respectively.

In Fig. 5 we plot the mean delay assuming some load imbalance due to hashing at VLR_1, and that a cache is installed at all the VLRS. (We continue to assume that the hashing at all VLRs other than VLR_1 results in perfect load balance.) It is clear that caching can result in very substantial reductions in total translation delay.

3.2 Two-stage version

To model the two-stage version, we set *Stage* = 1. In addition, we assume that processing at the first-stage TS now simply consists of performing a second-level hash function and other minor tasks, so that $1/\mu_{t1} = 1/\mu_{t1c} = 1.25ms$, while the actual database lookup is done at the second level TS, so that $1/\mu_{t11} = 5ms$.

We assume that the two-stage version of the scheme will only be used when translation rates have become very high due to high PCS and NGPN penetra-

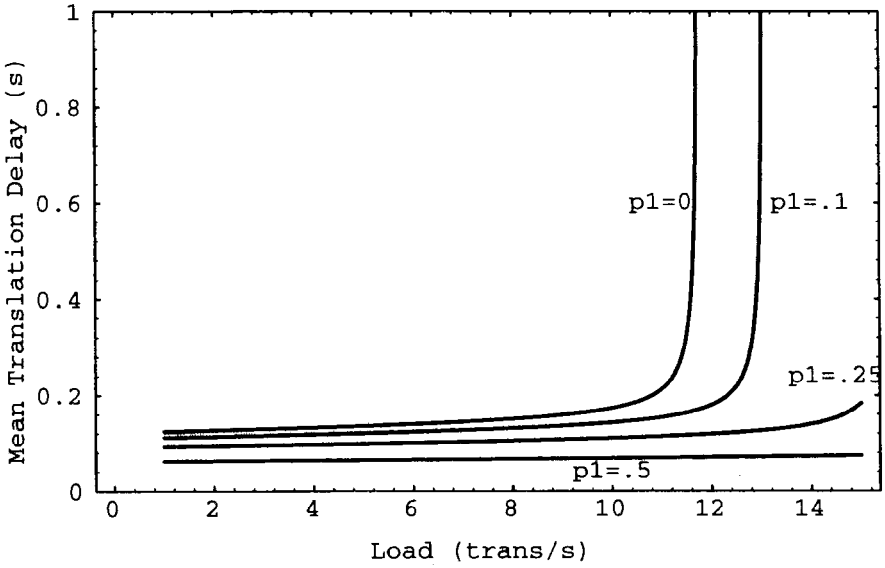

Figure 5 One-stage version: Mean translation delay for $f_1 = 2$ and $1/\mu_{t1} = 5ms$; p_1 varied.

tion. Thus we consider the scenario where the number of active users per PCS cell, $n = 30$, and $\lambda = 19.2$ in the busy hour.

Fig. 6 shows the variation of the mean delay with translation load when there is some load imbalance at VLR_1 ($f_1 = 2$), and there are two second-level TSs, i.e., $1/\beta = 2$. (We continue to assume that the hashing at all VLRs other than VLR_1 results in perfect load balance.) We consider the situation where no caching is used at the VLRs ($p_1 = 0$). We can see that even without the use of a second-level cache (i.e, $p_2 = 0$), the effect of the second-level TSs is to allow a much greater translation load to be processed; saturation occurs at an offered load of around $\lambda = 22$ instead of the $\lambda = 11$ trans/sec in the one-stage case.

The use of a second-level cache is very effective for $0.1 \leq p_2 \leq 0.25$ in this example; for $p > 0.25$ the second-level cache provides diminishing returns. Since there are far fewer TSs than VLRs, second-level caches may be more cost-effective than caching at the VLRs. Note, however, that at low loads ($\lambda \leq 20$ trans/sec.) the VLR cache reduces the mean translation delay while the second-

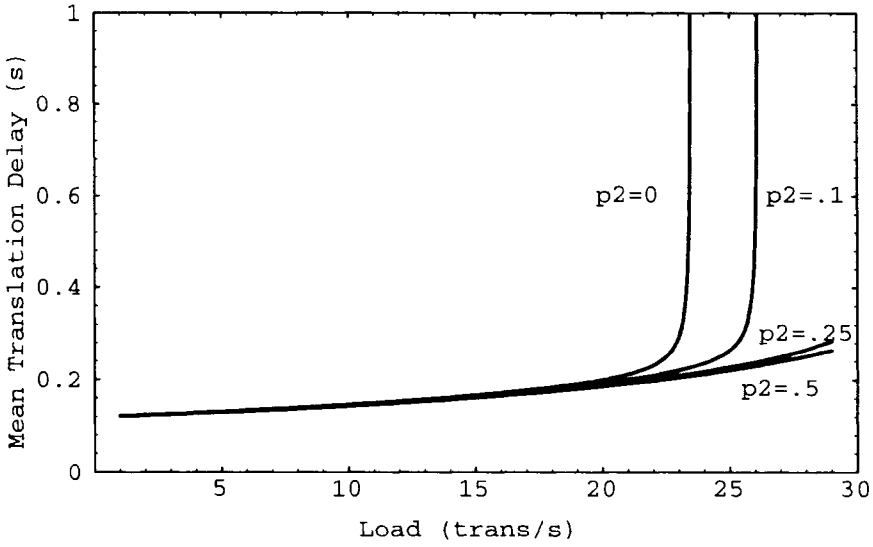

Figure 6 Two-stage version: Mean translation delay for $f_1 = 2$, $1/\mu_{t1} = 1.25\,ms$, and $1/\mu_{t11} = 5\,ms$; p_2 varied.

level cache does not; this is because at low loads the delay due to VLR processing and SVC setup is still a significant fraction of the total delay, and is not affected by second-level caching.

4 CONCLUSIONS

We have developed a simple model for the mean translation delay due to the one-stage and two-stage versions of our proposed NGPN scheme. Example calculations with the model show that the mean translation delay can be kept low (under 0.5 second), with appropriate use of caching and second-level translation servers if necessary. The model is fully parameterized and can be used to investigate other example scenarios if desired. At present the model does not provide variances or percentiles of delay; these can be incorporated if the translation delay becomes a more significant proportion of the total call setup time.

Acknowledgments. Thanks are due to Li-Fung Chang, Matthew Cheng, Greg Pollini and Subhashini Rajagopalan for useful and enlightening discussions.

REFERENCES

[1] Bellcore. Network and Operations Plan for access services to Personal Communications Services (PCS) providers. Special Report SR-TSV-002402, Bellcore, Aug. 1992.

[2] Bellcore. CCS network interface specification (CSNIS) supporting SCCP and TCAP. Generic Requirements GR-1432-CORE, Bellcore, 1993.

[3] Bellcore. Alternatives for signaling link evolution. Special Report SR-NWT-002897, Bellcore, Feb. 1994.

[4] Bellcore. PCS Network Access Services. Special Report SR-TSV-002459, Bellcore, Dec. 1994.

[5] R. Jain, S. Rajagopalan, and L.-F. Chang. A hashing scheme for phone number portability in PCS systems with ATM backbones. In *Proc. IEEE Conf. Pers. Indoor Mobile and Radio Comm. (PIMRC)*, Oct. 1996.

[6] R. Jain, S. Rajagopalan, and L.-F. Chang. Phone number portability for PCS systems with atm backbones using distributed dynamic hashing. *IEEE J. Sel. Areas Comm.*, 15(1):96–105, 1997. Special Issue on Wireless ATM.

[7] Ravi Jain, Yi-Bing Lin, and Seshadri Mohan. Location strategies for personal communications services. In J. Gibson, editor, *Mobile Communications Handbook.* CRC Press, 1996.

[8] Leonard Kleinrock. *Queuing Systems Volume II: Applications.* Wiley, 1976.

[9] E. D. Lazowska, J. Zahorjan, G. S. Graham, and K. C. Sevcik. *Quantitative System Performance.* Prentice-Hall, 1984.

[10] S. Mohan and R. Jain. Two user location strategies for PCS. *IEEE Pers. Comm. Mag.*, 1(1), First Quarter 1994. (Premiere issue).

[11] T.-H. Wu and L. F. Chang. Architectures for PCS mobility management on ATM transport architectures. In *Proc. Intl. Conf. Univ. Pers. Comm.*, pages 763–768, 1995.

This page intentionally left blank.

SUPPORTING QOS CONTROLLED HANDOFF IN MOBIWARE

Andrew T. Campbell, Raymond R.-F. Liao and Yasuro Shobatake

Department of Electrical Engineering and
Center for Telecommunications Research
Columbia University, New York City, NY 10027-6699, USA
http://comet.ctr.columbia.edu/wireless

Abstract

Currently, the ATM Forum's Wireless ATM (WATM) working group is in the process of developing a set of new specifications for ATM mobility. The work encompasses the definition of a new 25 Mbps air-interface, suitable location management and a fast handoff scheme. To date, most of the working group contributions have focussed on extensions to existing signalling specifications with little regard for the type of Quality of Service (QOS) which could be achieved during handoff. The working group considers WATM to be a direct extension to the standard ATM specification with uniformity of end-to-end QOS. In this paper we advocate an alternative approach. We argue that large-scale mobility requirements, limited radio resources, and fluctuation network conditions make the delivery of hard QOS guarantees in the wireless domain very difficult. To address this challenge, we are implementing a QOS-aware middleware platform called "mobiware" which operates between the application and radio ATM link layers. At the heart of the mobiware platform lies a QOS controlled handoff algorithm which exploits the inherent scalability of audio and video flows. Implicit in the term "QOS controlled" is the notion that audio and video flows can be represented as multi-layer scalable flows and adapted during handoff to meet fluctuating network conditions based on a user-supplied QOS adaptation policy. Novel aspects of the mobiware QOS controlled handoff signalling algorithm include the use of mobile-soft-state to represent mobile flows, aggregation techniques for bundling and transporting mobile flows to and from mobile devices, and re-routing and QOS re-negotiation anchor points which limit the impact of small-scale mobility on the wireline network.

1 Introduction

Next generation wireless communications systems such as *wireless ATM (WATM)* will be required to support the seamless delivery of voice, video and data with high quality. In this context, WATM is intended to be a direct extension of the existing fixed/wireline broadband ATM networks with uniform end-to-end QOS guarantees. Delivering hard QOS guarantees in the wireless domain is, however, rather difficult since assumptions made in providing QOS guarantees in wireline ATM networks do not always hold in their wireless extension due to large-scale mobility requirements, limited radio channel resources and fluctuating network conditions [3]. Bandwidth made available to a set of mobile devices during admission control may vary due to changes in the link quality caused by the prevalence of channel impairments and high transmission error rates. In addition, a connection with certain capacity reserved for a mobile device at a particular cell may have to be re-routed to another cell during handoff, and the new path may not have the originally required capacity. Therefore, re-negotiation of resources allocated to the connection is needed. At the same time though, the flow (e.g., audio or video) should be transported and presented 'seamlessly' to the destination device with a smooth change of perceptual quality. These conditions impact our ability to deliver hard QOS guarantees in WATM networks.

Although researchers have addressed the isolated areas of handoff criteria, connection rerouting, signalling extension [6]-[13], and QOS provision [14] in mobile ATM networks, little attention has been directed toward the development of a handoff algorithm which supports the seamless delivery of scalable multimedia flows with controlled QOS. The term "controlled QOS" is used to distinguish it from the hard QOS guarantees offered by wireline ATM networks. Implicit in this term is the notion that multimedia flows can be represented and transported as multi-layer scalable flows which can be intelligently and perceptibly scaled-up or scaled-down to match the available resources at the bottleneck base-to-mobile links during handoff. In this paper we propose a unique solution to the overall problem of providing handoff with QOS assurances. This proposed signalling algorithm is a part of a novel QOS-aware middleware platform called *mobiware.*

The structure of this paper is as follows. In Section 2, we present an overview of the mobiware platform. Following this in Section 3 we describe the QOS controlled handoff design goals, architecture and algorithms. In Section 4 we present a walkthrough of the QOS controlled handoff scheme. Finally, we present our concluding remarks in Section 5.

2 Mobiware Platform

Mobiware is a software middleware platform that runs seamlessly on mobile devices, base stations and mobile-capable ATM switches. The platform is built on distributed system and Java technology and incorporates new architecture and novel adaptive algorithms to support QOS controlled mobility. Mobiware is based on **xbind** [18] and CORBA technology [17] and is designed to operate between the application and radio-link layers of future wireless media systems. The platform provides value-added QOS support by allowing mobile multimedia applications to operate transparently during handoff and periods of persistent QOS fluctuation. The goal of the mobi-

ware adaptive algorithms is to transport scalable flows, reduce handoff dropping and improve wireless resource utilization.

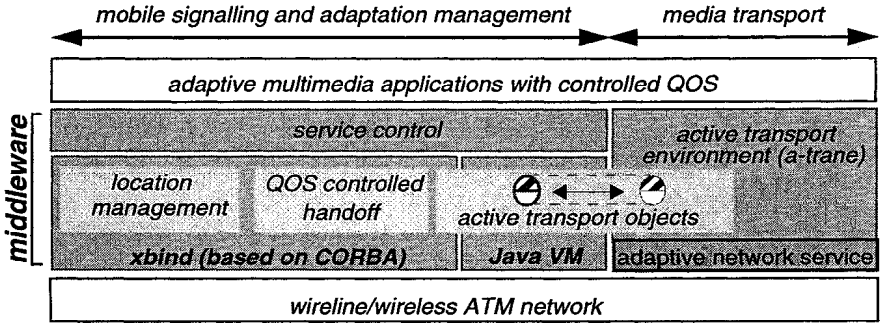

Figure 1: Mobiware Platform

Mobiware promotes the separation between mobile signalling and adaptation management on the one hand and media transport on the other. As illustrated in Figure 1 mobiware utilizes **xbind** (based on CORBA) and Java for signalling and adaptation management during handoff or periods of persistent QOS fluctuation. The Java Virtual Machine executes on mobile devices, base stations and mobile-capable ATM switches and supports the dynamic execution of *active transport objects (ATOs).* These transport objects constitute an 'active' component of the mobiware transport system which can dispatch ATOs to strategic points in the network or end-systems to provide value-added QOS support.

The realization of end-to-end QOS control and the exploitation of scalable flows is achieved in mobiware through: i) resource binding between mobile devices, base stations and ATM switches; and ii) provision of a set of QOS-aware adaptive algorithms. These algorithms operate in unison under the control of mobiware:

- *QOS controlled handoff,* provides signalling for handoff which exploits the use of: i) mobile soft-state and hard-state to represent mobile flows; ii) aggregation of mobile flows to and from mobile devices; and iii) re-routing and QOS renegotiation anchor points to limit the impact of small-scale mobility on the wider fixed network;
- *adaptive network service,* provides hard QOS guarantees to base layers (BL) and soft QOS guarantees to enhancement layers (viz. E1 and E2) of flows based on the availability of resources in the wireless environment; and
- *active transport environment*, supports the transfer of multi-layer flows through the provision of a QOS-based API and a set ATOs (e.g., media scaling [19]) and *static transport objects (STOs),* e.g., playout control. STOs are statically configured and execute at mobile and fixed devices only. In contrast, ATOs are dynamically dispatched to the mobile devices, base stations or ATM switches to support valued-added QOS at strategic nodes.

For full details on the mobiware platform see [4].

3 QOS Controlled Handoff

In this section we describe the mobiware QOS controlled handoff architecture and algorithms. First, we introduce a number of design goals which motivate the design of our scheme. Following this we present an overview of the QOS controlled handoff algorithm.

3.1 Design Goals

A number of goals motivate the design of the QOS controlled handoff scheme. These goals are based on certain assumptions concerning the nature of multimedia flows, providing QOS guarantees in the wireless domains and handoff dropping policy.

First, the handoff algorithm must support bidirectional flows where a mobile device can be both sender and receiver. For multicast flows, individual receivers (both wired and wireless) may have differing QOS capabilities to consume flows. This could be due to either fluctuating network resources with mobility or imposed by individual applications. Bridging this heterogeneity gap [3] in mobile multicast environments while simultaneously meeting the individual mobile device QOS requirements is addressed by mobiware.

Second, the approach should limit the impact of handoff on the smooth delivery of data. This is of paramount importance when considering the delivery of continuous media flows during handoff. Continuous media has a strong sense of isochronous delivery and smooth playout at receivers. In order to achieve this requirement the handoff algorithm should provide fast connection rerouting without significantly interrupting flows in progress. In addition, handoff should provide for a smooth change of the flow's perceptual quality in the case where the new location does not have sufficient resources.

Third, an important goal of QOS controlled handoff is to limit the routing and resource reservation disturbance of small-scale mobility of a mobile device on the other members of a multicast flow and on the fixed wireline network. We model small-scale mobility by limiting the amount of re-routing and QOS re-negotiation during handoff. However, handoff takes into account any potential sub-optimal routing which may evolve from this policy through the activation of a periodic route optimization algorithm.

Fourth, the handoff scheme interacts with media scaling agents [18] to filter audio and video flows to match available bandwidth. Our goal here is to reduce the handoff dropping probability and to provide better resource utilization in the wireless domain by admitting as many mobile devices as possible during handoff. Providing hard QOS guarantees in the WATM environment (as advocated by the ATM Forum) is very difficult to achieve in practice. To address this mobiware provides a QOS adaptive approach which we argue is more suitable in the wireless media environment.

Our final design goal accepts that it is better to scale-down audio and video flows and admit them during handoff rather than drop them when the maximal desired capacity is unavailable. To address this mobiware represents audio and video flows as multilayer scalable flows.

3.2 Handoff Network Model

An important contribution of our architectural approach is the consideration of two normally conflicting design goals: supporting the dynamic, transient and adaptive

nature of device mobility and mobile flows while limiting any potential disturbance incurred by QOS re-negotiation during handoff.

To achieve these design goals we introduce three novel architectural concepts:

- *mobile soft-state*, models the dynamics of mobility and service adaptation such as media scaling through the periodic re-routing and QOS re-negotiation of flows as a mobile device roams. Mobile soft-state is established between a per mobile *routing anchor point (RAP)* and the mobile device, and hard-state between the RAP and the wireline network. In this context the RAP provides an interface between the hard-state and soft-state portions of flows;
- *logical anchor points*, limit the impact of small-scale mobility of a mobile device on the wireline network. Per mobile RAP and *QOS re-negotiation anchor points (QRP)* localize the periodic re-routing and QOS re-negotiation during handoff, respectively; and
- *connection groups (CG)*, provide a common routing representation for all virtual paths and virtual circuits *to* and *from* mobile devices. Connection groups are maintained between a per mobile RAP and mobile devices and decouple handoff re-routing from resource allocation. The RAP allows collective control of all mobile flows associated with a mobile device during handoff. By aggregating flows through a single reference point (i.e., RAP) mobiware supports a fast and efficient handoff algorithm.

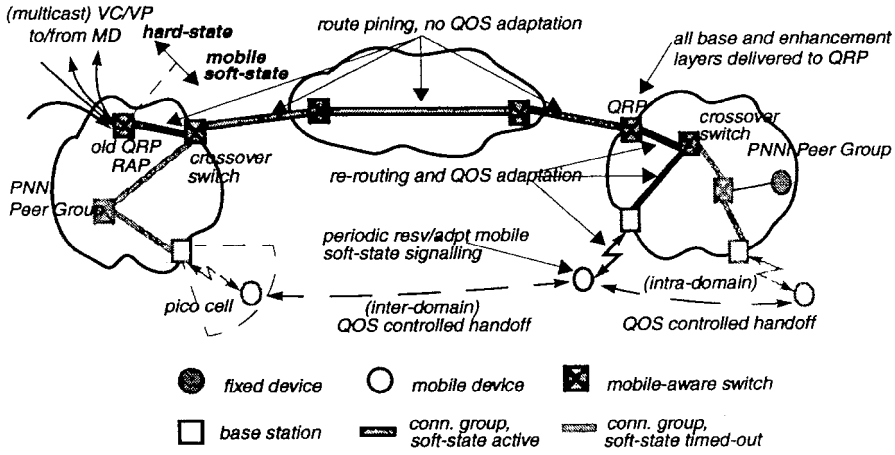

Figure 2: QOS Controlled Handoff Model

QOS controlled handoff supports multicast flows where a mobile device can be both a source and receiver. The handoff scheme uses a soft-handoff approach for flows which terminate at a mobile devices and hard-handoff for flows which originated at mobile devices. This approach helps reduce the complexity of managing flows during handoff while supporting QOS needs. The QOS controlled handoff scheme is classified as a forward handoff type with partial re-routing. As illustrated in Figure 2 the crossover switch is optimally chosen to be along the path between the mobile device and the RAP.

As illustrated in Figure 2, the concepts of mobile soft-state, connection group and logical anchor points (viz. RAP and QRP) are modeled in the wireline and wireless domain providing QOS controlled mobility. The mobiware platform models the wireless portion of the ATM network as being divided into pico-cells each served by a base station connected to a wireline ATM network. Base stations are cell relays which translate the ATM cell headers from radio ATM format to that used by standard ATM. Each base station supports signalling, QOS control and adaptation management of flows based on the semantics of an adaptive network service. The existing wireline ATM network provides connectivity between base stations. We organize the wireless network into domains. A domain consists of a set of base stations which are under the management of mobile-aware ATM switches. A domain corresponds to a logical partition of the wireline network and the physical location of the base stations in hierarchies for scalable routing (in ATM Forum PNNI routing, these domains would be peer groups).

3.3 QOS Controlled Handoff Algorithms

Mobiware models mobile flows through the semantics of an adaptive network service and QOS controlled mobility. The semantics of the adaptive network service dictate that base layers undergo full end-to-end admission testing and support hard end-to-end QOS guarantees. In contrast, enhancement layers are admitted without any such test but compete for residual capacity between the QRP and the mobile device. Mobiware assumes that network resources are scarce in the WATM domain and abundantly available in the wireline ATM domain. The semantics of QOS controlled mobility assume that as mobile devices migrate between cells then mobile flows are scaled by media scaling ATOs to match the available bandwidth in the target cell. To meet these QOS requirements, mobiware models mobile flows through a combination of mobile soft-state, logical anchor points and connection groups.

3.3.1 Mobile Soft-State A significant contribution of the QOS controlled handoff scheme is the use of mobile soft-state. Mobile soft-state is admirably suited to support the dynamics of mobile flows which operate in mobile and QOS fluctuating environments. Mobile soft-state combines self-removable routing state and periodic QOS adaptation between mobile device and the RAPs. Mobile soft-state is designed to model the dynamics of small-scale and large-scale mobility in WATM environments. As illustrated in Figure 2, mobile devices periodically send reservation (resv) messages toward the RAP in the wireline network. These resv messages carry the mobile devices desired QOS requirement and are interpreted by the base station and fixed network infrastructure according to the rules of the mobiware adaptive network service. As illustrated in Figure 2, resources are allocated to a mobile device over a particular wireless/wireline route for the duration of the mobile soft-state timeout. This is achieved when a QRP responds to a resv message by sending an adaptation (adpt) message back toward the mobile devices with an indication of the allocated resources. Soft-state provides a receiver-driven QOS adaptation mechanism for the mobile device to periodically probe the wireless network for better quality. This approach is particularly suitable when individual receivers of a multicast flow have differing QOS needs. The mobiware QOS model also operates on very fast traffic control time scales where traffic variations are countered by buffering and scheduling algorithms with resources allocated by a measurement based admission control algorithm.

The periodic *resv* message is used in combination with an adpt message to continually probe for better QOS between the mobile device and QRP. This probing occurs on the

QOS re-negotiation time scale. If during that time the infrastructure receives another resv message it refreshes the mobile soft-state held in the base station and switches between mobile devices and the RAPs. As mobile devices handoff between adjacent cell periodic *resv* messages re-establish the appropriate QOS between the mobile device and the new QRP. At the same time the mobile soft-state between the old base station and the new crossover switch expires; that is, the appropriate switch table entries and resources are automatically deallocated and made available to the remaining mobile devices.

Soft-state plays three important roles along the path between a mobile device and its logical anchor points:

- to *refresh* the soft-state timer between a QRP and the mobile device,
- to *re-negotiate* QOS between the QRP and the mobile device based on the semantics of the mobiware adaptive network service and support the active transport environment, and
- to *update* the connection group to ATM VP/VC mapping between the RAP and the mobile device.

As illustrated in Figure 2, hard-state is installed in the wireline environment above the RAP and mobile soft-state in the wireless domain between RAPs and mobile devices. QOS adaptation operates in the wireless domain between mobile devices and QRPs to share limited radio resources between mobile devices. By limiting QOS adaptation to the wireless domain, we isolate the wireline environment from periodic QOS re-negotiation that is active in the wireless domain.

Mobiware provides a receiver-oriented approach to QOS adaptation where individual members of a multicast flow specify their QOS requirements and adapt to available resources. In contracts, sources continuously send the full complement the flow. This limits the impact a source of a multicast flow can have of all receivers.

3.3.2 Logical Anchor Points A RAP and QRP are defined for each mobile device. The major goal of mobile-to-QRP signalling is for QOS adaptation using the resv/adpt messages pair. The QRP is located in the same local domain as the mobile devices to minimize the impact that small scale mobility has on the wireline portion of the network. The RAP serves as a gateway for a mobile device to the outside network as all the flows to and from this mobile device are routed through the associated RAP. The RAP's major function is soft-state management, which includes anchoring the connection group, interfacing soft/hard states, consolidating resv and generating adpt messages.

The QRPs are fixed for intra-domain handoffs and relocated for inter-domain handoffs. In general, the location of RAP remains fixed. When the mobile device comes on line for the first time a RAP and QRP are instantiated. The RAP is initially co-located with the QRP, and is fixed for the duration of the handoff to preserve the mobile soft-state connection group path as a basis for future handoff. The RAP is a constraining point regarding routing to and from the mobile device after handoff. During handoff the location of the crossover switch has to be chosen between the RAP and the mobile device. By extending the soft-state connection group path the handoff re-routing algorithm can select the optimal location of the crossover switch, as presented in Figure 3 which illustrates a zigzag handoff which is common in the pico-cellular environment.

As illustrated in Figure 3, originally the RAP and QRP are co-located in the same domain. After the first handoff, the QRP is located to the new domain while the RAP stays in the same location to maintain the existing connection group. The mobile device then migrates back to the old cell. The crossover switch is optimally chosen to be near the new location of the mobile device. The mechanism for locating QRP and RAP is simplistic. During the first connection setup, with the knowledge of the pre-assigned location of the QRP (queried from PNNI peer group leader), the soft-state is set up between the mobile and the QRP/RAP. During handoff, a *handoffSetup* message will also set up the soft-state between the new QRP and RAP.

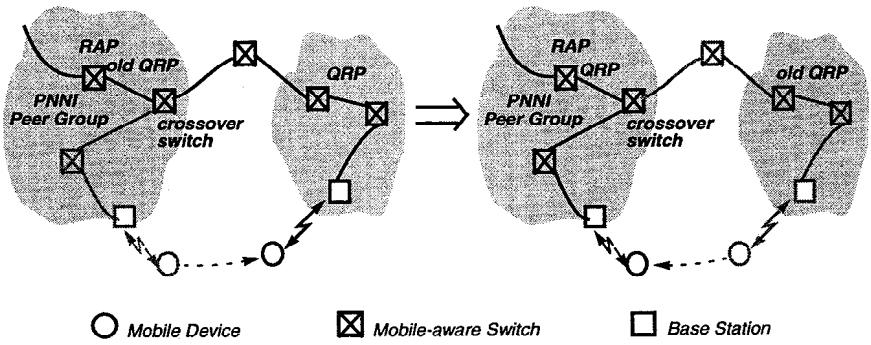

Figure 3: An Inter-domain Handoff Scenario

The RAP and QRP handle small-scale mobility and inter-domain handoff re-routing. However, suboptimal routing can still occur. This develops when a mobile device migrates further away from its initial RAP location during multiple inter-domain handoffs. To resolve this mobiware executes a path optimization algorithm. The path optimization algorithm is conservative and executes on longer time scales than handoff. The reasoning behind this conservative approach is that the path optimization requires a complete rebuild of all the flows to and from the mobile devices. This will, by its nature, be quite disruptive to all flows in operation. Hence the need to limit total rebuilds to those times when it is absolutely necessary.

3.3.3 Connection Group QOS controlled handoff uses the concept of a connection group (CG), which is analogous to virtual paths in ATM, to aggregate all flows between the mobile devices and the RAPs. Mobiware extends the connection bundling concept first introduced in [7] [13] from a VP to a higher level of aggregation allowing VC/VPs with different QOS requirements to be grouped together and transported. The connection group concept promotes the idea of supporting individual VC and VP with different QOS requirements at ATM switches and base stations.

The connection group is maintained by the *resv/adpt* mechanism. The *resv* message probes the path between the mobile device to the RAP to detect any individual VC/VP setup or timed-out/torn-down. On the return path, the *adpt* message updates connection group mapping tables in all the intermediate switches and the base station. During handoff, a new portion of connection group is set up using the multicast operation *addBranch.* After handoff, the removal of the old branch from a connection group tree is managed automatically through the soft-state operations used in the mobile environment.

4 Handoff Walk-through

In this section we provide a step-wise walk-through of the QOS controlled handoff algorithm. Each phase of the handoff is considered in turn. These phases comprise the *handoff initiation, path setup* and *path teardown* phases. Figure 4 presents each phase and it respective constituent parts.

4.1 Handoff Initiation Phase

The first phase of QOS controlled handoff determines whether a new base station can provide a stronger signal at the desired level of QOS. A QOS monitoring algorithm resident at mobile device monitors the beacon messages which are periodically broadcast by all neighboring base station. In addition to indicating the signal strength of neighboring cells, beacon messages indicate the residual capacity currently available at the neighboring base stations. By periodically monitoring beacons from all neighboring base stations the mobile device is able to determine link qualities and occupancy of adjacent base stations and use this QOS state information as a basis to initiate handoff after a suitable dwell time.

The next step in handoff is the establishment of a new signalling channel between the mobile device and the new base station. A mobile device issues a *(1) signalRequest* to the new base station over a dedicated *meta signalling* channel using the base station address found in the beacon message. This results in the creation of a signalling channel when the base station responds with a *(2) signalResponse.* The signalling channel carries all signalling and *resv, adpt* messages between the mobile and wireline network.

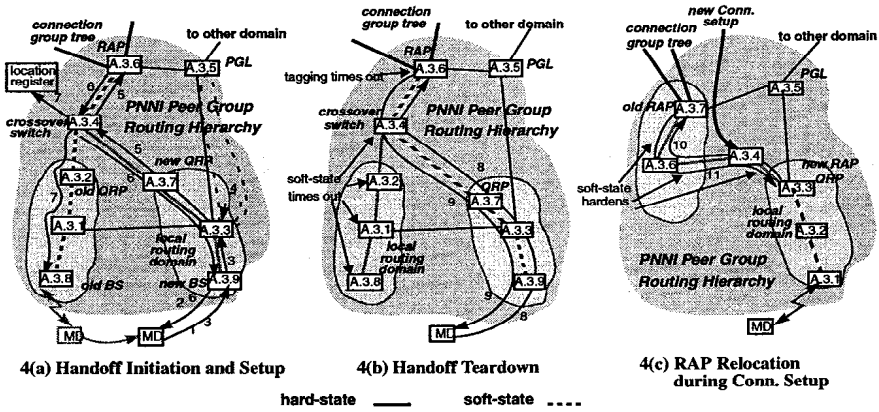

Figure 4: Mobiware Handoff Walk-through

4.2 Path Setup Phase

Once the signalling channel has been successfully created the mobile device initiates a forward handoff to the new base station. It does so by issuing a (3) *handoffSetup* message which includes connection group route state information and the desired

QOS required. Mobile devices express their desired QOS in terms of the semantics of the adaptive network service which are mapped into connection group QOS requirements in terms of connection group base layer requirements and enhancement layer requirements, respectively. Admission control located at the new base station first determines whether sufficient resources are available to support the minimum QOS of the requested handoff.

The base station's parent mobile switch uses connection group routing information in combination with PNNI routing to (4)*locateCrossover* within the range of the old mobile soft-state path and establish the shortest branch between the new base station and the crossover switch on the existing connection group tree with the desired QOS. The PNNI peer group leader is interrogated in the case where the mobile device roams outside the current peer group domain of the old base station. This suggests that the existing crossover point would provide a suboptimal routing point. An optimal re-routing point could to be established by end-to-end hierarchical routing algorithm as described in section 3.3.2.

The mobile switch then issues a (5)*handoffSetup* message to the newly identified crossover switch. The QOS values in *handoffSetup* message can be modified by the distributed handoff algorithm based on the availability of resources at the traversed nodes as the *handoffSetup* messages is forwarded toward the RAP. The returned *(6)handoffAck* message commits switch and base station resources for the new connection group branch. The semantics of the adaptive network service provides hard guarantees to the base layers and admits enhancements layers based on the availability of residual resources. Existing mobiles' reservations *above* the base layer may be altered to allow a new mobile device to enter a cell. The base station drops the handoff call (6)*handoffDrop* if insufficient residual capacity is available to meet the group connection base layer resource requirements of the new mobile device.

The handoff algorithm interacts with the adaptive transport *media scaling agents* at either the base station or the mobile-aware switches during connection group setup. In the case of media scaling, the level of media scaling is dependent on the current utilization of the wireless link, application specific desired QOS and the semantics of the adaptive network service. Media scaling may result in the "scaling-down" of mobile delivered quality when a mobile device enters a pico-cell and "scaling-up" when they leave. QOS filters support the delivery of different combinations of layers to particular mobile devices based on the available resources, which is essential for multicast to heterogeneous receivers. For full details on mobiware media scaling see [19].

Handoff registers new mobile devices with the domain location management which in turn allocates a new proxy ATM address as mobile devices roam into cells within a new PNNI Peer Group. A (7) *locationUpdate* message is used to register the mobile devices at the home location in this case and update the cache register at the old base station.

4.3 Path Teardown

QOS controlled handoff is based the notion of soft-handoff for roaming receivers and hard-handoff for roaming senders. In the case of soft-handoff, the mobile device simultaneously receives flows from the old and new base stations. Once the crossover switch sends a *handoffAck* message to the mobile device it begins to forward the new flow to the mobile devices. This results in duplicate cells arriving at the mobile device from both the old and new base stations. The crossover switch uses cell tagging to

preserve ATM cell level sequence integrity at the mobile device during handoff. This is achieved by inserting an end-to-end OAM cell for both flows for every K data cells.

After tagging has commenced mobile devices determine suitable synchronization points in the old and new flows and initiate flow switching from the old to the new flow. After flow switching the old flow is rendered redundant. Old flows continue to arrive at the mobile devices as long as the route between the old base station and mobile switching point is active; that is, old flows are switched through to the mobile while the mobile soft-state is still installed and has not timed-out. Mobile devices do not, however, have to process old flows. One advantage of this mobile soft-state tear-down approach is to allow path re-use if a mobile device comes back before the path states time-out. This is equivalent to pre-reserving network resource following the tracking of a mobile device, which can speed up handoff in a pico-cell environment.

The mobile device at the new cell starts its QOS adaptation and mobile soft-state management procedure by sending (8)*resv* message toward its QRP and RAP. The resv messages probes for higher quality enhancement layers and consolidates connection group to VP/VC mapping between the mobile device and QRP point. The *resv* message is also used to consolidate the connection group mapping between the QRP and RAP point. Note that there is no QOS adaptation between these points. The full complement of connection group base and enhancement layers are delivered to the QRP. QOS adaptation is active from the QRPs to the mobile devices. The RAP responds to a *resv* by sending an adpt messages back toward the mobile. This message confirms the connection group mapping, refreshes the mobile soft-state between the RAP and QRP. In addition, the adpt message also informs the mobile device of the resources available to it over the next interval and scales the available resources for the mobile device. Media scaling filters the flow at the base station or at the QRP.

After flow switching the mobile device refrains from sending any further periodic resv messages to the old base station and closes the signalling channel to the old base station. Once the mobile soft-state timer expires the old branch of the connection group tree between the old mobile switching point and the old base station is timed-out and removed. Media scaling is once again invoked at the old base station to determine if deallocated resources can be utilized by the any existing mobile devices at old base station. If the *resv* and *adpt* message pairs of these mobile devices confirm the residual bandwidth availability, the media scaling agent scales up the corresponding media flows.

4.4 RAP Relocation Optimization

As illustrated in 4c mobiware avoids potential sub-optimal routing to the RAP for new calls. A RAP relocation feature allows the routing algorithm to route a new connection to a merge point between the current RAP and QRP point as oppose directly to the RAP. RAP relocation optimization is illustrated in 4c. This procedure does not require any re-routing of existing flows from the old RAP location to the new one. All that is required is that the RAP is relocated and the soft-state between the new and old RAP points in hardened (10) and (11).

5 Conclusion

In this paper we have argued that large-scale mobility requirements, limited radio resources, and fluctuation network conditions make the provision of hard QOS guar-

antees in the wireless domain rather difficult. Currently we are implementing a QOS-aware middleware platform called mobiware and its QOS controlled handoff algorithm which exploits the inherent scalability of audio and video flows. Novel aspects of our QOS controlled handoff include the use of soft-state and hard-state to represent mobile flows, aggregation of flows to and from mobile devices and RAP and QRP which limit the impact of small-scale mobility on the wireline network. Mobiware contrasts approaches taken in WATM working group which advocates extensions to the traditional ATM Forum UNI and PNNI signalling stacks. Implicit in mobiware is the notion that audio and video flows can be represented as multi-layer scalable flows which can be adapted during handoff to meet fluctuating network conditions.

The mobiware testbed consists of 4 ATM switches (viz. ATML Virata, Fore ASX100/ASX200s, NEC Model 5, Scorpio Stinger) and 4 base stations. The base stations are multi-homed 200 MHz Pentium with 25 Mbps wireline access to the wireline ATM network and 2 Mbps WaveLAN air-interfaces to a number of mobile devices based on Pentium PCs and notebooks. The PCs run Linux, Windows/NT and **xbind** (based on CORBA). An early version of mobiware runs on PCs, base stations and the ASX100 ATM switch. Finally, we have implemented a beta version of the handoff protocol which supports mobile soft-state, connection groups and RAPs and QRPs. In addition we have completed the implementation of active transport object for mobile filters to scale media during handoff [19].

6 References

[1] Raychaudhuri, D., (NEC USA), Dellaverson, L., (Motorola), Umehira, M., (NTT Wireless Systems), Mikkonen, J., (Nokia Mobile Phones), Phipps, T., (Symbionics), Porter, J., (Olivetti Research), Lind, C., (Telia Research) and Suzuki, H., (NEC C&C Research), *Scope and Work Plan for Proposed Wireless ATM Working Group*, ATM Forum Technical Committee, ATM Forum/96-0530/PLEN, April, 1996.

[2] Schwartz, M., *Network Management and Control Issues in Multimedia Wireless Networks*, IEEE Personal Commun., pp8-16, June 1995.

[3] Campbell, A., *Towards End-to-End Programmability for QOS Controlled Mobility in ATM Networks and their Wireless Extension*, Proceeding of MoMuC-3, pp P.2b-1-5, Sept. 1996.

[4] Campbell, A., *Mobiware: QOS-aware Middleware for Mobile Multimedia Communications*, Proc. IFIP 7th Intl. Conf. on High Performance Networking, White Plains, New York, April 1997, (to appear).

[5] Pollini, G. P. *Trends in Handover Design*, IEEE Commun. Mag., pp82-90, March 1996.

[6] Raychaudhuri, D., (NEC USA), Dellaverson, L., (Motorola), Umehira, M., (NTT Wireless Systems), Mikkonen, J., (Nokia Mobile Phones), Phipps, T., (Symbionics), Porter, J., (Olivetti Research), Lind, C., (Telia Research) and Suzuki, H., (NEC C&C Research), *Scope and Work Plan for Proposed Wireless ATM Working Group*, ATM Forum Technical Committee, ATM Forum/96-0530/PLEN, April, 1996.

[7] Porter, J., et al., *The ORL Radio ATM System, Architecture and Implementation*, Olivetti Research Tech. Report 96.5, Jan. 1996.

[8] Porter, J. and Gilmurray D., *Tunnelled Signalling for the Support of Mobile ATM*, ATM Forum Technical Committee, ATM Forum/96-1699, Dec., 1996.

[9] Agrawal, P., Hyden, E., Krzyzanowski, P., Mishra, P., Srivastava, M. and Trotter, J., *SWAN: A Mobile Multimedia Wireless Network*, IEEE Personal Commun., Apr. 1996.

[10] Acharya, A., Li, J. and Raychaudhuri, D., *Primitives for Location Management and Handoff in Mobile ATM Networks*, ATM Forum Contribution 96-1121, August 1996.

[11] Deane, J., *WATM Interim Mobility Protocol (WIMP) Protocol*, ATM Forum Contribution 96-1118, August 1996.

[12] Akyol, B.and Cox, D., *Handling Mobility in a Wireless ATM Network*, IEEE INFOCOM'96, pp1405-1413, April 1996.

[13] Toh, C.-K., *The Design & Implementation of a Hybrid Handover Protocol for Multimedia Wireless LANs* ACM MOBICOM' 95, pp49-61, 1995.

[14] Naghshineh, M. and Acampora, A.S., *QOS Provisioning in Micro-Cellular Networks Supporting Multimedia Traffic*, IEEE INFOCOM'95, pp1075-1084, April 1995.

[15] Lazar, A., Bhonsle, S. and Lim, K.S., *A Bind Architecture for Multimedia Networks*, Journal of Parallel and Distributed Computing, Vol. 30, No. 2, Nov. 1995.

[16] Zhang, L., et al., *Resource Reservation Protocol (RSVP) - Version 1 Functional Specification*, IETF Working Draft, draft-ietf-rsvp-spec-07.ps, 1995.

[17] Object Management Group (OMG) and X/Open, *The Common Object Request Broker: Architecture and Specification*, Revision 1.2, December 1993.

[18] **xbind** project, http://www.ctr.columbia.edu/comet/xbind/xbind.html

[19] Balachandran A. and Campbell A.T., *Mobile Filters: Delivering Scaled Media to Mobile Devices*, Technical Report, Columbia University, http://comet.ctr.columbia.edu/wireless/publications/, January, 1997.

This page intentionally left blank.

BANDWIDTH ALLOCATION IN FIXED BROADBAND WIRELESS NETWORKS

Thomas K. Fong, Paul S. Henry, Kin K. Leung
Xiaoxin Qiu, and N.K. Shankaranarayanan

AT&T Labs - Research
100 Schulz Drive
Red Bank, NJ 07701

fong, psh, kkleung, xqiu, shankar@research.att.com

Abstract:

We consider use of fixed broadband wireless networks to provide packet services for telecommuting and Internet access. Each cell in such networks is divided into multiple sectors, each of them served by a sector antenna co-located with the base station (BS), and user terminals also use directional antennas mounted on the roof top of homes or small offices and pointed to their respective BS antennas. To support a target data rate of 10 Mb/s, a bandwidth of several MHz is required. Since radio spectrum is expensive, the bandwidth need to be reused very aggressively. Thus, efficient strategies for frequency reuse and managing co-channel interference are critically important. We propose here a dynamic radio-resource allocation method, to be referred to as staggered resource allocation (SRA) method, that uses a distributed scheduling algorithm to avoid major sources of interference, while allowing concurrent packet transmission and meeting a specified signal-to-interference objective. The performance of the method is studied by analytic approximations and detailed simulation.

Our results show that the combination of directional antennas plus the SRA method is highly effective in controlling co-channel interference. For reasonable system parameters, the SRA method delivers a throughput in excess of 30% per sector while permitting a given frequency band to be re-used in every sector of every cell. It also provides

higher than 90% probability of successful packet transmission at appropriate traffic load conditions. In addition, a simple control mechanism can be applied in the method to improve performance for harsh radio environments.

1 INTRODUCTION

As work-at-home, telecommuting and Internet access become very popular, the demand for broadband packet services will grow tremendously. Customers are expecting high quality, reliability and easy access to high-speed communications from homes and small businesses. High-speed services are needed in the very near future for: a) accessing World Wide Web for information and entertainment, b) providing data rates comparable to local-area networks (LAN) for telecommuters to access their computer equipment and data at the office, and c) multimedia services such as voice, image and video.

In this paper, we consider a wireless approach to this problem: fixed (i.e., non-mobile) broadband packet-switched TDMA networks with user data rates of 10 Mb/s, link lengths typically less than 10 kilometers and operating frequency in the range of 1 to 5 GHz. We specifically focus on networks using sector antennas at base stations and narrow-beam antennas at roof tops of homes and small offices. Physical and link-layer issues arising in the design of such networks include modulation, equalization, access techniques, co-channel interference, bandwidth usage, error control and multi-access schemes [1]. Our focus here is to address just one of these issues, namely, the co-channel interference.

Microwave spectrum is expensive, so efficient strategies for re-using frequencies and managing co-channel interference are critically important. To support a user data rate of 10 Mb/s in an interference-limited wireless environment, a bandwidth of several MHz is needed for TDMA. In contrast to narrowband cellular networks where radio spectrum is divided into multiple channel sets, which are re-used only in relatively distant cells [6], broadband wireless networks must re-use bandwidth very aggressively, ideally reusing the same frequency band in every cell. The need for reuse of a common radio bandwidth in all cells has also been noted by [3] and [2] for mobile broadband wireless networks. Note that although CDMA uses the same frequency band in all cells, the radio bandwidth required for supporting the target data rate of 10 Mb/s will be excessive and the associated high processing speed has not yet shown to be technologically feasible.

In the context of packet-switched networks, time slots naturally become the bandwidth resources, which are shared by all users in a sector or cell. We need to dynamically allocate time slots to various transmitters to send data packets such that a given signal-to-interference ratio (SIR) can be guaranteed at the intended receiver for successful packet reception. This results in the concept of *dynamic resource allocation.* By using a central controller, [10] and [8] propose approaches to assigning time

slots. In our fixed wireless networks, cell sectorization and directional antennas at fixed locations are key components in reducing the amount of interference from neighboring sectors and cells. In this paper, we propose a strategy for managing co-channel interference that does not require a central controller. It permits reuse of the same bandwidth in every cell, resulting in a high degree of spectral efficiency.

The rest of this paper is organized as follows. Section 2 describes the bandwidth allocation problem in fixed broadband wireless networks. Section 3 introduces the *staggered resource allocation (SRA) method* for bandwidth allocation, and discusses its properties. We outline performance models for the SRA throughput in Section 4. Using typical parameters, Section 5 presents several numerical examples, and discusses the performance characteristics of the proposed methods. Finally, Section 6 is our conclusion.

2 BANDWIDTH ALLOCATION IN FIXED BROADBAND WIRELESS NETWORKS

Consider a broadband wireless network where each cell is divided into multiple sectors, each of which is covered by a sector antenna co-located with a base station (BS) at the center of the cell. Because of the co-location, sector antennas are also referred to as BS antennas. Terminals (users) use directional antennas mounted on the roof top and pointed to their respective BS antennas. The beamwidth (angle) of each BS antenna should be just wide enough to cover the whole sector, while a terminal antenna pointing to a designated BS antenna can have a smaller beamwidth to avoid interference. The ratios of front-to-back-lobe gain (abbreviated by FTB ratio below) for BS and terminal antennas may be different, and are assumed to be finite. Time is slotted such that a packet can be transmitted in each slot. In addition, the downlink and uplink between terminals and BS are provided by time-division duplex (TDD), using the same radio spectrum.

To make our ideas concrete, let us consider a hexagonal cell layout. Each cell is divided into six sectors, each of which is served by a BS antenna with 60° beamwidth. Terminal antennas can have an angle smaller than 60°. For the hexagonal layout, using a simple path-loss model [9], we can find the major sources of interference for the downlink and uplink in a given sector. Our challenge is to develop resource allocation methods that avoid most of the major interferers.

3 THE STAGGERED RESOURCE ALLOCATION (SRA) METHOD

In the *staggered resource allocation (SRA)* method, the frame structure and the sector labeling are presented in Figure 1. Basically, a fixed number of time slots for a downlink (or uplink) are grouped into *subframes* and they are labeled by 1 to 6. Sectors are also labeled by 1 to 6 anti-clock-wise. The labeling patterns for adjacent cells differ by a

120 degrees rotation, thus creating a cluster of 3 cells whose patterns can be repeated across the whole system.

A special slot assignment order is established and followed by each sector. As shown in Figure 2, for example, a sector with label 1 first schedules packets for transmission in time slots of subframe 1 (denoted by **a**). If it has more traffic to send, it then uses subframe 4 (**b**), subframe 5 (**c**), etc. until subframe 6 (**f**). The idea here is that, if interference due to concurrent packet transmission in the same cell can be tolerated, then after using all slots in the first subframe **a**, a sector should use the first subframe of the opposite sector in the same cell, in order to make the best use of the BS directional antennas. Following that, time slots in the first subframes for the sectors next to the opposite sector are used. To avoid interference due to imperfect antenna patterns of neighboring sectors, their first subframes are used as the last resort. As the figure shows, the assignment order for the next sector is "staggered" by a right rotation by one subframe based on the order for the previous sector. For this reason, this method is called the staggered resource allocation (SRA) method. The benefits of the SRA method in terms of interference avoidance and control of concurrent packet transmissions are explained in the following.

Avoidance of Major Interference

First, let us consider the intra-cell interference. It is easy to see from Figure 2 that if all sectors have traffic load of less than one-sixth of total channel capacity, all packets are transmitted in different time subframes, thus causing no interference within the same cell. Of course, as the traffic load increases, packets are transmitted simultaneously, thus increasing the level of interference. Nevertheless, the special assignment order exploits the characteristics of directional antennas to allow multiple concurrent packet transmissions while maximizing the SIR.

Besides managing intra-cell interference, the SRA protocol also helps avoid interference from major sources in the neighboring cells. This is particularly so when traffic load is low to moderate. To see this, let us consider the downlink for Sector 1 in the middle cell of Figure 1. Sector 2 in the bottom cell and Sector 3 in the upper cell are the major sources of interference. By examining the assignment order for Sectors 1, 2 and 3, one finds that they will not transmit simultaneously, so they will not interfere with each other provided that all of them have a traffic load of less than one-third of total channel capacity (i.e., using only subframes **a** and **b** for transmission). The same comment also applies to the uplink where Sectors 2 and 5 of the bottom cell in the figure now become the major sources of interference. Due to the symmetry of the assignment order and cell layout, this same comment applies to each sector in every cell.

A Control of Concurrent Transmissions for Enhanced Quality of Service

According to a given radio environment and antenna characteristics, the SRA method can be used in conjunction with a control mechanism to improve the reception quality in terms of SIR at the receiving ends. Specifically, the control limits packet transmissions only in the first few subframes for each sector. For example, for the given radio environment, if at most three packets can be sent simultaneously by various BS or terminal antennas in the same cell to ensure the required reception quality, only time slots in subframes **a**, **b** and **c** as indicated in Figure 2 can be used for transmission by each sector. In general, we have shown in [5] that if each sector schedules packet transmission in the first k subframes in the SRA method, there are at most k packets transmitted simultaneously by various antennas in each cell at any time. The control limits the degree of concurrent transmissions, thus the amount of interference, to achieve the desirable SIR. When different grades of quality of services (QoS) such as for voice, secure data, non-realtime data calls are defined in terms of the SIR threshold, the control can be used as a mechanism for providing the required QoS. The ability of the mechanism to control the exact number of major interferers is also useful for systems using adaptive antennas [11] for suppression of interference where a specific number of interferers can be tolerated.

Optimality of the SRA Method in a Single-Cell Setting

Consider the downlink performance of the SRA method in a simple, single cell setting. Assume that the cell is divided into N equal sectors and each sector BS antenna has a perfect antenna pattern (i.e., it has a front lobe and a back lobe with a sharp distinction between them) with beamwidth of $360°/N$, but with a finite FTB ratio. All antennas transmit at a fixed power level, and always have packets ready for transmission. A power-law path-loss model and lognormal shadowing are also assumed. By considering the success probability of a packet transmission, the SRA method is shown in [5] to yield the maximum achievable throughput.

4 THROUGHPUT MODELS FOR THE SRA METHOD

An analytic approximate model and a detailed simulation model have been developed to study the downlink throughput of the SRA method. Both models consider a fixed number of terminals (users) in each sector, finite FTB ratios for BS and terminal antennas, radio path loss and lognormal shadowing effects. They treat a packet transmission as successful if the SIR at the intended receivers exceeds a given threshold and the SRA throughput defined as the number of packets successfully transmitted per time slot in each sector.

Analytic Approximate Model

Success of a packet transmission depends on the traffic load of various sectors, and the radio environment and antenna characteristics. So, the exact throughput analysis requires consideration of the interdependency among all BS antennas, thus becoming intractable. To develop an approximate model, we assume that traffic process for each BS antenna is statistically independent, but yet packet transmissions by all BS antennas are considered in estimating the probability of successful reception at a receiver. Our simulation results reveal that such an approach yields excellent throughput approximations.

Our approximate model consists of two submodels: a) a Markovian traffic submodel to capture the traffic characteristics for each sector in the SRA method, and b) an interference submodel to consider the radio environment and antenna characteristics. For simplicity, we assume here all sectors have identical offered traffic load. Thus, the same traffic submodel is applicable to all sectors. This assumption can be relaxed by applying different traffic parameters to the submodel for various sectors. Due to space limitation, details of the submodels are omitted here and they can be found in [5].

Simulation Model

The simulation model also considers the 7-hexagonal-cell layout in Figure 1. Each cell is divided into 6 sectors, each of which is served by a BS antenna co-located at the center of the cell. The beamwidth of each BS and terminal antenna is 60° and 30°, respectively, while each terminal antenna points directly to its BS antenna. The FTB ratios for the BS and terminal antenna are finite and adjustable as input parameters for the simulation. The model uses ideal antenna patterns in a way that if a receiver antenna sees the front lobe of a transmitting antenna, the radio signal is attenuated only by the path loss. Otherwise, the signal is also attenuated according to the FTB ratio. Each radio path between a pair of BS and terminal antennas is characterized by a path-loss model [9] and lognormal shadow fading. For the downlink, since there is only one radio path between all BS antennas in the same cell (which are co-located) and an arbitrary terminal, the intended signal and interference experience the same lognormal fading and path loss. However, the fading from BS antennas in different cells are different and independent. For each packet transmission, if the SIR at the intended receiver exceeds a threshold, the packet is considered to be successfully received. Otherwise, the packet is retransmitted later. Since our focus is on the throughput, the simulation model assumes no retransmission delay and that new and transmitted packets in each sector are scheduled for transmission effectively in a random order.

There are 20 terminals (or users) randomly placed in each sector. For downlink traffic, as in the analytic approximate model, each BS antenna is assumed to have a buffer to hold one packet for each user. When the buffer is occupied, subsequent packet arrivals for the user are blocked and cleared. Packets are scheduled to transmit by each

BS antenna according to the SRA method as discussed above. To avoid non-uniform performance in the outer layer of cells, only the statistics in the middle cell in Figure 1 are collected and reported below.

5 THROUGHPUT PERFORMANCE AND DISCUSSION

In this numerical study, each subframe has 12 time slots, each sector has 20 terminals, the standard deviation of lognormal shadowing is 4 dB and the path loss exponent is 4.

Figure 3 presents the downlink throughput of the SRA method as a function of normalized aggregated load from the analytic approximate model and the simulation model (represented by curves and symbols, respectively). A set of typical FTB ratios for BS and terminal antennas, denoted by B and T respectively, are considered in the examples. Clearly, the throughput depends on the SIR detection threshold. With straightforward modulation and equalization schemes (e.g., QPSK and DFE), the threshold probably lies between 10 to 15 dB. As shown in the figure, the maximum throughput in each sector for the SRA protocol with these parameters ranges from 30% to 80%. That is, while re-using the same frequency to support high user data rates in every sector of every cell, we can still achieve a throughput in excess of 30%, which translates into a very large network capacity! For the 30% to 80% throughput per sector, each cell effectively has a data capacity of 1.8 to 4.8 times of the channel data rate. Evidently, this happens because, as intuitively expected, the SRA protocol is capable of selectively allowing concurrent packet transmission to increase throughput while avoiding major interference to yield satisfactory reception. It is possible that the SRA throughput can be improved further if the FEC code and power control are adopted in a way similar to that in [2] and [3].

In terms of the quality of the analytic model, as shown in Figure 3, the approximations closely match the simulation results, except for the case with 15 dB detection threshold, and the FTB(B,T) ratios of 20 and 10 dB at high traffic conditions. Given the stringent detection threshold and antenna characteristics, packet transmission is likely to fail and subject to retransmissions in this case, thus yielding high aggregated traffic load and strengthening the interdependence among all BS's. When this happens, the approximate model, which only captures part of the correlation among BS's by use of success probability, becomes inadequate for predicting accurate throughput. For other less stringent parameter settings, the approximations are excellent.

Another way to quantify the merits of the SRA method is the success probability of packet transmission. That is, the probability that the SIR for a packet transmission exceeds a given threshold for successful reception. From our simulation results, Figure 4 shows that the probability decreases as the traffic load grows. To avoid excessive retransmission delay, it is expected that the success probability should be higher than 90%. In light of this, we observe that except at high load for the most stringent case, the probability is above 90% for all other three settings. Combining this with the through-

put results discussed above, the SRA method is capable of sustaining a throughput of 70% to 80% per sector, while yielding the satisfactory success probability. Even for the most stringent case, the target success probability can be maintained by keeping the throughput below 30%, which still represents a large network capacity.

As discussed earlier, the SRA method can be used in conjunction with a control mechanism to limit the degree of concurrent packet transmission for enhancing the reception quality. Figure 5 portrays how the SRA throughput can be improved for the most stringent case, namely, FTB(B,T)=(20dB,10dB) and the SIR threshold of 15dB. Let the degree of concurrent transmission be limited by the control to C packets per cell. Note that when the control is not applied, $C = 6$ as time slots in all six subframes can be used for transmission by each sector. As C decreases from 6, the maximum throughput first increases and reaches the peak when $C = 3$. This is so because with a smaller C value, the control effectively limits the degree of concurrent packet transmission, thus enabling more packets to exceed the SIR threshold and improving the throughput. However, as C decreases further, the throughput starts to decrease because only a small fraction of time slots are available for transmission by each sector when C is too small. Note that when $C = 1$, the SRA method has a maximum throughput of about 1/6.

We also observe from Figure 5 that for $C = 4$ to 6, the throughput first increases, reaches a peak at a certain aggregated load, and then starts to decrease if the load increases further. In contrast, the throughput for cases with $C = 1$ to 3 always increases, perhaps marginally with the traffic load. This is because for the given parameter setting, if the degree of concurrent transmissions is higher than 3, packet transmissions become more likely to fail as more packets are transmitted at high loads. On the other hand, for $C = 1$ to 3, the throughput is limited mainly by the control. Thus, the throughput still increases slightly as the traffic load increases.

In a separate study, we find that for the most stringent parameter setting, even when $C = 1$, only a few percent of randomly chosen locations such as those at the boundary of adjacent cells may have difficulty in achieving the SIR threshold of 15dB. The percentage is reduced for a lower threshold and/or higher antenna FTB ratios, and may increase for antenna patterns that are more practical than the ideal one assumed here. To provide close to 100% coverage in harsh environment, it may be desirable to enhance the SRA method to serve those "bad" locations. Other strategies to allocate radio resources to "bad" and "good" locations can also been found in [4].

In addition to improving the throughput, the control mechanism also enhances the packet success probability in the SRA method. This can be seen in Figure 6 for the most stringent case. As one would expect, at a fixed traffic load, the probability increases as C decreases simply because less concurrent transmission implies less interference. It is interesting to observe that the probability is only marginally improved when C changes from 2 to 1, while the corresponding throughput improvement is much larger in Figure 5. This characteristic of the SRA method can be explored by network designers to

obtain a balanced tradeoff between the throughput and success probability for a given radio environment and the desired QoS.

6 CONCLUSION

We have shown that the combination of directional antennas plus dynamic resource allocation in the time domain is highly effective in controlling co-channel interference in fixed wireless systems. For reasonable choices of system parameters, our *staggered resource allocation (SRA)* technique delivers high throughput while permitting a given band of frequencies to be re-used in every sector of every cell. The SRA method also provides satisfactory probability of successful packet transmission. In addition, a simple control mechanism can be applied in the SRA method to improve the throughput and success probability for harsh radio environments and poor antenna characteristics. We believe the proposed approach is a significant step forward in establishing the viability of fixed broadband wireless access networks.

Acknowledgments

Thanks are due to Lek Ariyavisitakul, Kapil Chawla, Len Cimini, Vinko Erceg, Tony Rustako and Arty Srivastava for their helpful discussions.

References

[1] E. Ayanoglu, K.Y. Eng and M.J. Karol, "Wireless ATM: Limits, Challenges, and Proposals," *IEEE Personal Communications*, pp.18-34, August 1996.

[2] F. Borgonovo, L. Fratta, M. Zorzi and A. Acampora, "Capture Division Packet Access: A New Cellular Access Architecture for Future PCNs," *IEEE Communications Magazine*, pp. 154-162, Sept. 1996.

[3] F. Borgonovo, M. Zorzi, L. Fratta, V. Trecordi and G. Bianchi, "Capture-Division Packet Access for Wireless Personal Communications," *IEEE J. Select. Areas in Commun.*, pp.609-622, Vol. 14, No.4, May 1996.

[4] K. Chawla and X. Qiu, "Resource Assignment in a Fixed Wireless System," to be presented in IEEE ICUPC'97 and to appear in IEEE Commun. Letters.

[5] T.K. Fong, P.S. Henry, K.K. Leung, X. Qiu and N.K. Shankaranarayanan, "Bandwidth Allocation in Fixed Broadband Wireless Networks," *Workshop Record of the 6th WINLAB Workshop on Third Generation Wireless Networks*, New Brunswick, NJ, March 20-21, 1997, pp.89-121.

[6] W.C.Y. Lee, *Mobile Cellular Telecommunications Systems*, McGraw-Hill, New York, 1989.

[7] A.M. Law and W.D. Kelton, *Simulation Modeling and Analysis*, McGraw-Hill, New York (1982).

[8] N. Kahale and P.E. Wright, "Dynamic Global Packet Routing in Wireless Networks," *Proc. of IEEE INFOCOM'97*, Kobe, Japan, April 1997, pp.1416-1423.

[9] T.S. Rappaport, *Wireless Communications: Principles and Practice*, IEEE Press and Prentice Hall PTR, New York, 1996.

[10] J.F. Whitehead, private communications, 1996.

[11] J.H. Winters, J. Salz and R.D. Gitlin, "The Impact of Antenna Diversity on the Capacity of Wireless Communication Systems," *IEEE Trans. on Commun.*, pp. 1740-1751, Vol.42, No.2-4, Feb.-April 1994.

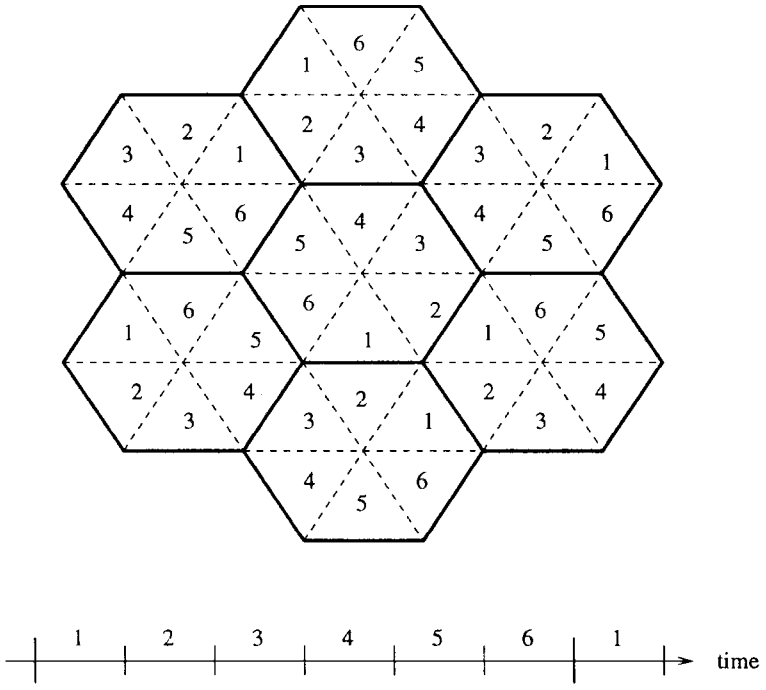

Figure 1.1 Sector Labeling and Frame Structure for the SRA Method.

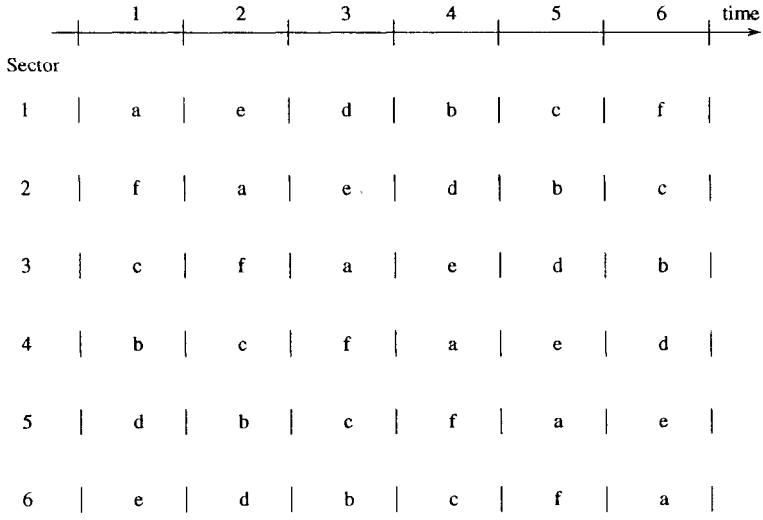

Figure 1.2 Order of Slot Assignment for the SRA Method.

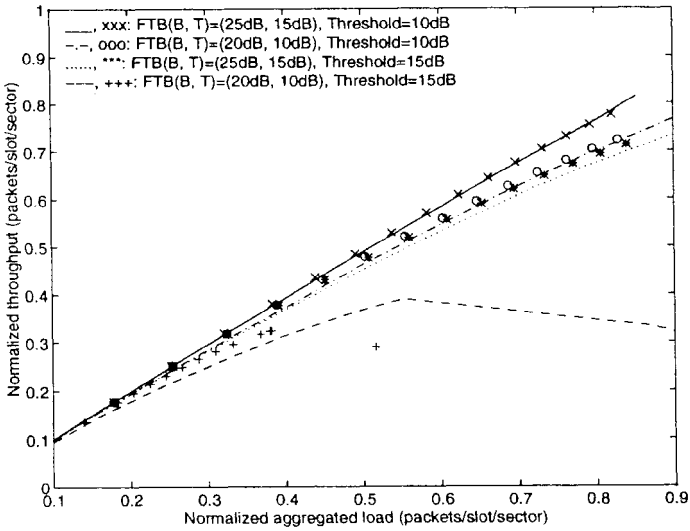

Figure 1.3 Throughput of the SRA Method.

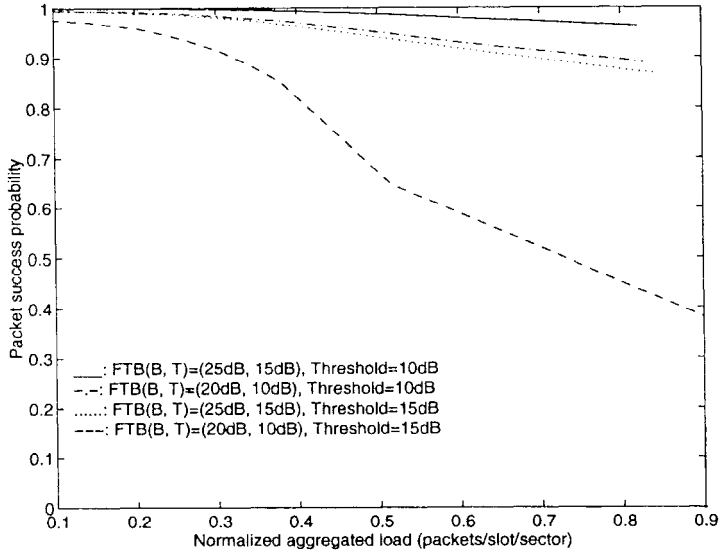

Figure 1.4 Packet Success Probability for the SRA Method.

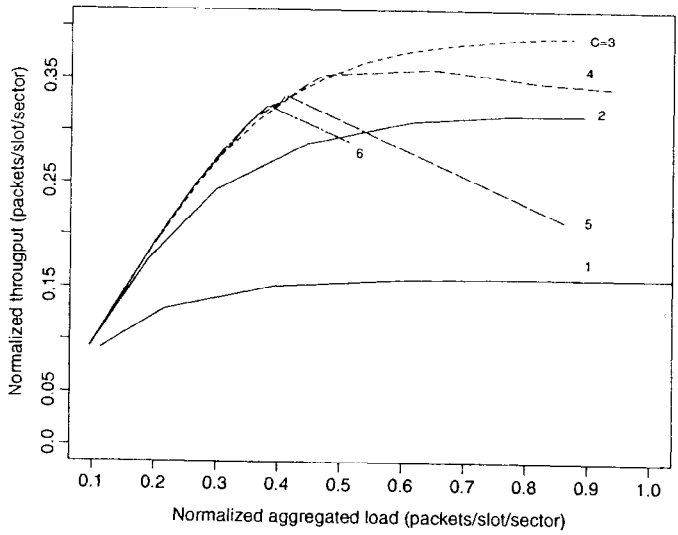

Figure 1.5 SRA Throughput With Transmission Control.

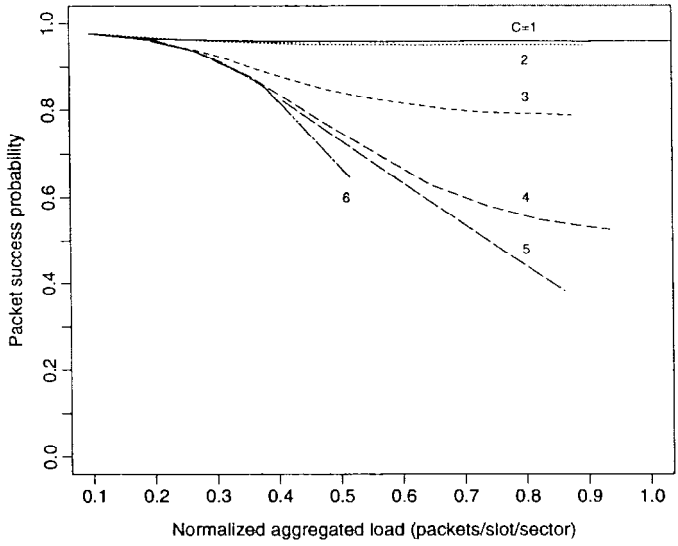

Figure 1.6 Packet Success Probability for the SRA Method With Transmission Control.

This page intentionally left blank.

A NOVEL DISTRIBUTED POWER CONTROL ALGORITHM FOR CLASSES OF SERVICE IN CELLULAR CDMA NETWORKS

Debasis Mitra

and John A. Morrison

Bell Laboratories
Lucent Technologies
Murray Hill, NJ 07974

{mitra,jam}@lucent.com

Abstract: The paper proposes a distributed power control algorithm for integrating heterogeneous transmitting sources, which have a broad range of statistical/burstiness characteristics and quality of service requirements, in wideband cellular CDMA networks. Speech sources with silence detection and data are canonical examples of service classes. Each service class is characterized by on-off transmissions with characteristic probabilities, desired minimum carrier-to-interference ratio and minimum probability with which the latter is required to be satisfied. In the power control algorithm given here, the received power for each service class at each cell is adapted locally based on only local measurements of the mean and also, importantly, the variance of the interference. The algorithm is derived from an asymptotic analysis in which the bandwidth as well as the number of mobiles are large. The analysis leads to a Gaussian approximation to the interference at each cell, which depends on the power levels. The algorithm is remarkable for its simplicity in the decoupling between classes. This is due in part to an attractive product-form in the expression for the dominant term in the asymptotic expansion of the desired power for each service class and cell. A condition for geometric convergence of the adapted power to the ideal power is obtained. The condition is remarkably unburdensome and only slightly more demanding than the condition based on the mean values of source activities. The condition also defines the capacity of the cellular CDMA network.

1 INTRODUCTION

The paper addresses the challenge of integrating heterogeneous transmitting sources with a broad range of burstiness characteristics in wideband cellular CDMA networks. Provided the right power control can be devised, CDMA offers the potential of extracting significant gain from the statistical multiplexing of such sources. In this paper we propose a new distributed power control algorithm for a range of service classes in cellular CDMA wireless networks. Each service class is characterized by its mobiles' distinctive bit rates, statistical and burstiness properties, and quality of service parameters consisting of the desired minimum carrier-to-interference ratio (CIR) and the minimum probability with which this requirement on the CIR must be satisfied. A canonical service class is voice sources with silence detection, which are modelled as on-off sources. Data sources, if also present, typically have quite different burstiness characteristics and, importantly, more stringent quality of service requirements on the CIR. In the power control algorithm that we give here, the power is adapted locally based on only local measurements of the mean and variance of the interference.

The paper is in two parts. In the first part the probabilistic quality of service specification is described. Also given is an asymptotic framework for estimating the orders of magnitude of quantities which affect the quality of service. The asymptotic scaling reflects wideband characteristics, i.e., the bandwidth as well as the number of mobiles are large. The key result of the first part are expressions for the asymptotically ideal, i.e., minimal, powers at each cell site for each service class (see (53)). The second part of the paper gives our distributed, measurement-based algorithm for solving the aforementioned expressions for the ideal powers. The analysis of the dynamical behavior of the algorithm yields a condition for the geometric convergence of the powers to their ideal values. This condition is also important because it determines the capacity of the cellular CDMA network. The condition obtained is remarkably unburdensome. The condition is only slightly more demanding than the condition based on the mean values of source activities, i.e., where the variabilities of the sources are eliminated.

This work extends some recent results in [MMO96], where the following notions were introduced: stochastic on-off traffic sources, probabilistic quality of service specifications, an asymptotic scaling for wideband networks, a Gaussian approximation for the cell-dependent interference and, importantly, a distributed power control algorithm driven not only by the mean of the local interference, which is conventional, but also the variance. New to this paper is the concept of classes of service. The power control algorithm given here (see (56)) is remarkable for the simple, decoupled dependence on the classes of service. This, in turn, is due in large part to a result obtained in the analysis, which states that the dominant term in the asymptotic expansion for the ideal powers at each cell site for the various service classes has a simple product-form, in which a term common to all service classes is multiplied by the desired minimum CIR for the service class. Moreover, the common term is the ideal power at the cell of a simple cellular network construct having a single class of service.

An important limitation of the analysis of the paper is the assumption of time-invariant path gains. The conventional argument to support this assumption is based on a separation of time scales of the power control convergence and variations in the path gains. We believe that the algorithm's sensitivity to randomness renders it more effective in

the presence of path gain variations and fluctuations. That is, the measurement-based terms in the power adaptation algorithm which reflect the mean and variance of the interference will take into account not only the effects of randomness in transmissions, as postulated in the model, but also time variations in the path gains. It should, however, be noted that this is not established here.

For the benefit of readers this paper has been written to be self-contained, which has necessitated some overlap with [MMO96]. Otherwise, given the complicated development, it would be excessively distracting and tedious.

Prior work on power control has been done in distributed [ZAN92(a), ZAN92(b), BPO92, FMI93, HAN93], distributed-asynchronous [MIT93, GZY95] and centralized [GVG93] frameworks. Hanly [HAN93] considers a cellular model with interference fluctuations. Early work on power control in spread spectrum networks is reported in [NAL83]. Recent work which unifies power control and cell site selection is reported by Hanly [HAN95] and Yates [YAT95]. Works which integrate power control and admission control are [BCP95] and [BCM95], the former focussing on active link protection and the latter on channel probing. On-off sources are considered in [VVI93] and [MMO94]. Recently, Everitt and Evans [EEV95] introduced the concept of effective interference to quantify the effects of variability in interference on the network capacity in the context of multiple service classes and cellular CDMA networks. However, the investigation is essentially independent of power control. Andersin and Rosberg [ARO96] develop a framework for power control in cellular networks, with some features in common with [MMO96], where the focus is mobility and time varying path gains. While this paper deals with CIR, and only indirectly with bit error rates (BER), it is possible to connect power control directly to BER, as in [KYH95]. The source for CDMA, specifically for matters related to interference and capacity, is Viterbi [VIT95].

2 MODEL

Let X_m be the activity indicator for mobile m, i.e., $X_m \in \{0, 1\}$ and $X_m = 1$ if and only if the mobile is active at a point in time. An active mobile m transmits at power p_m, and the gain to cell site j is g_{jm}, so that the received power at the cell site is $g_{jm} p_m$. We assume that the received power at any cell site k from all the active mobiles of class i in that cell is common and given by P_{ki}, i.e.,

$$P_{ki} \triangleq g_{km} p_m, \quad \forall m \in \{(k, i) \mid X_m = 1\} . \tag{1}$$

Since our system is spread spectrum, the interference to an active mobile's transmission is closely related to the total received power at its local site. In order to calculate such interference, we first let $I_{jk}^{(i)}$ denote the received power at cell site j due to all the mobiles in cell k of class i, i.e.,

$$I_{jk}^{(i)} \triangleq \sum_{m \in (k,i)} g_{jm} p_m X_m = P_{ki} \sum_{m \in (k,i)} \frac{g_{jm}}{g_{km}} X_m . \tag{2}$$

We isolate the component independent of power, which depends on relative gains and the random activity indicators, and denote it by $M_{jk}^{(i)}$, i.e.,

$$M_{jk}^{(i)} \triangleq \sum_{m\in(k,i)} \frac{g_{jm}}{g_{km}} X_m \quad (1 \le j \le J, \quad 1 \le k \le J, \quad 1 \le i \le I) . \tag{3}$$

Hence, $I_{jk}^{(i)} = P_{ki} M_{jk}^{(i)}$, and we note that

$$M_{jj}^{(i)} = \sum_{m\in(j,i)} X_m . \tag{4}$$

The total power received at cell j is

$$I_j \triangleq \sum_{k=1}^{J} \sum_{\ell=1}^{I} P_{k\ell} M_{jk}^{(\ell)} + \eta_j W , \tag{5}$$

where η_j $W(1 \le j \le J)$ is the local receiver noise power at cell site j, and W is the spread bandwidth (and also the processing gain). Since we are considering a wideband system, W is large. For any active mobile m in cell j of class i, $X_m = 1$ and the interference to its transmission is

$$I_{ji} \triangleq I_j - P_{ji} . \tag{6}$$

The quality of service requirement in terms of the carrier-to-interference ratio (C/I) is

$$P_{ji}/I_{ji} \ge \alpha_i/W \quad (1 \le j \le J, \quad 1 \le i \le I) , \tag{7}$$

where $\{\alpha_i\}$ are prespecified numbers; the right-hand quantity is typically small, since the processing gain W is large. From (6), we may write (7) as

$$P_{ji} \ge \frac{\tilde{\alpha}_i}{W} I_j \qquad (1 \le j \le J, \quad 1 \le i \le I) , \tag{8}$$

where

$$\tilde{\alpha}_i = \alpha_i / \left(1 + \frac{\alpha_i}{W}\right) . \tag{9}$$

Typically $\tilde{\alpha}_i \approx \alpha_i$.

Now consider the stochastic traffic model of a bursty mobile m of class i to be on-off, i.e., in cell j it is on, i.e. active, with probability w_{ji}:

$$w_{ji} = \Pr\{X_m = 1\} = 1 - \Pr\{X_m = 0\}, \quad \forall m \in (j,i) . \tag{10}$$

Although w_{ji} may not depend on the cell index j, we are retaining the generality for possible future applications. On account of the random source behavior, the quality of service requirement has to be probabilistic, with probability of compliance at best less than unity. Let the requirement be

$$\Pr\left\{P_{ji} \ge \frac{\tilde{\alpha}_i}{W} I_j\right\} \ge 1 - L_i \quad (1 \le j \le J, \; 1 \le i \le I) , \tag{11}$$

where $\{L_i\}$ are given parameters. Thus the quality of service parameters for class i are α_i and L_i.

From (10) we have

$$E(X_m) = w_{ki}, \quad \text{var}(X_m) = w_{ki}(1 - w_{ki}), \quad \forall m \in (k, i) . \tag{12}$$

Hence, from (3),

$$E\left(M_{jk}^{(i)}\right) = \sum_{m \in (k,i)} \frac{g_{jm}}{g_{km}} w_{ki} = K_{ki} G_{jk}^{(i)} w_{ki} , \tag{13}$$

where K_{ki} is the number of mobiles in cell k of class i, and

$$G_{jk}^{(i)} \triangleq \frac{1}{K_{ki}} \sum_{m \in (k,i)} \frac{g_{jm}}{g_{km}} . \tag{14}$$

The quantity $G_{jk}^{(i)}$ represents the averaged normalized gain that is used in the determination of the mean interference at cell j due to mobiles of class i in cell k. The averaging is over the latter set of mobiles and the normalization is with respect to their local gains. Note, in particular, that $G_{jj}^{(i)} = 1$. From (3) and (12) we obtain

$$\text{var}\left(M_{jk}^{(i)}\right) = \sum_{m \in (k,i)} \left(\frac{g_{jm}}{g_{km}}\right)^2 w_{ki}(1 - w_{ki}) = K_{ki} \left(H_{jk}^{(i)}\right)^2 w_{ki}(1 - w_{ki}) , \tag{15}$$

where

$$\left(H_{jk}^{(i)}\right)^2 = \frac{1}{K_{ki}} \sum_{m \in (k,i)} \left(\frac{g_{jm}}{g_{km}}\right)^2 . \tag{16}$$

The quantities $\left(H_{jk}^{(i)}\right)^2$ occur in the determination of the variance of the interference at cell j due to mobiles of class i at cell k. Their role is similar to that of $G_{jk}^{(i)}$ for the mean interference, compare (13) and (15).

We assume that g_{jm}/g_{km} are uniformly bounded for $m \in (k, i)$. Since $X_m \in \{0, 1\}$, it follows that the random variables $g_{jm} X_m / g_{km}$ are uniformly bounded for $m \in (k, i)$. Hence, from the Lindeberg theorem [FEL68], the central limit theorem holds for $M_{jk}^{(i)}$ as $K_{ki} \to \infty$. Hence

$$M_{jk}^{(i)} = K_{ki} G_{jk}^{(i)} w_{ki} + H_{jk}^{(i)} \sqrt{K_{ki} w_{ki}(1 - w_{ki})}\, Z_{jk}^{(i)} , \tag{17}$$

where $Z_{jk}^{(i)}$ is asymptotically normally distributed, with zero mean and unit variance as $K_{ki} \to \infty$. Moreover, $Z_{jk}^{(i)}$ $(1 \le k \le J,\ 1 \le i \le I)$ are independent random variables.

In summary, we have shown that for our model of many mobiles and class-dependent, random, on-off transmission of mobiles, the component which is independent of power in the expression for the received power at cell site j due to all mobiles of class i in cell k is a Gaussian random variable given in (17).

3 ASYMPTOTICS, ORDERS OF MAGNITUDE

We introduce a natural scaling, which allows us to make order of magnitude estimates and to develop meaningful and efficient approximations by dropping negligibly small terms.

3.1 Asymptotic Scaling

Inherent to wideband systems in which the bandwidth (W) is shared by a large number of users is the following scaling, in which the large parameter K is the average number of mobiles in a cell, i.e.,$K \triangleq \frac{1}{J}\sum_{j=1}^{J}\sum_{i=1}^{I} K_{ji}$:

$$\frac{\alpha_i}{W} = \frac{a_i}{K}, \quad K_{ji} = \gamma_{ji}K \quad (1 \le j \le J, \quad 1 \le i \le I)\,, \tag{18}$$

where $a_i = O(1)$ and $\gamma_{ji} = O(1)$ as $K \to \infty$ and $W \to \infty$. Note that α_i and η_j are other $O(1)$ parameters. With a small loss of generality, consider these $O(1)$ parameters to be fixed in this scaling. Also, note that, importantly, W and K are of the same order. Let

$$P_{ji} = \widehat{P}_{ji} + \frac{1}{\sqrt{K}}Q_{ji} \quad (1 \le j \le J, \quad 1 \le i \le I)\,, \tag{19}$$

where $\widehat{P}_{ji}$ and Q_{ji} are $O(1)$. The orders of magnitudes of the first order (dominant) and second order terms in the above expansion of P_{ji} are dictated by consistency, as may be checked later.

3.2 Solutions to Leading Order Terms in the Asymptotic Expansion

Now let us investigate the implications of (18) and (19) on the terms P_{ji} and $\tilde{\alpha}_i I_j$ appearing in the quality of service specifications (11). By substituting (17), (18) and (19) in the expression for I_j in (5), and using (9), it may be verified that

$$\begin{aligned}\frac{\tilde{\alpha}_i}{W}I_j &= a_i \sum_k \sum_\ell G_{jk}^{(\ell)}\gamma_{k\ell}w_{k\ell}\widehat{P}_{k\ell} + \alpha_i\eta_j \\ &+ \frac{a_i}{\sqrt{K}}\sum_k\sum_\ell \left[G_{jk}^{(\ell)}\gamma_{k\ell}w_{k\ell}Q_{k\ell} + H_{jk}^{(\ell)}\sqrt{\gamma_{k\ell}w_{k\ell}(1-w_{k\ell})}\,Z_{jk}^{(\ell)}\widehat{P}_{k\ell}\right] \\ &+ O\left(\frac{1}{K}\right)\,.\end{aligned} \tag{20}$$

To achieve $P_{ji} \ge \tilde{\alpha}_i I_j/W$ $(1 \le j \le J, \quad 1 \le i \le I)$, we compare first, the order 1 terms and, second, the coefficients of $1/\sqrt{K}$, and obtain the system of inequalities given below in (21) and (22), respectively. These constitute sufficient conditions to give $P_{ji} \ge \tilde{\alpha}_i I_j/W$, to within $O(1/K)$.

$$\widehat{P}_{ji} - \frac{\alpha_i}{W}\sum_k\sum_\ell G_{jk}^{(\ell)}K_{k\ell}w_{k\ell}\widehat{P}_{k\ell} \ge \alpha_i\eta_j \quad (1 \le j \le J, \quad 1 \le i \le I)\,, \tag{21}$$

and

$$Q_{ji} - \frac{\alpha_i}{W}\sum_k\sum_\ell G_{jk}^{(\ell)}K_{k\ell}w_{k\ell}Q_{k\ell} \ge a_i\xi_j \quad (1 \le j \le J, \quad 1 \le i \le I)\,, \tag{22}$$

where

$$\xi_j \triangleq \sum_k \sum_\ell H_{jk}^{(\ell)} \sqrt{\gamma_{k\ell} w_{k\ell}(1 - w_{k\ell})}\, \widehat{P}_{k\ell} Z_{jk}^{(\ell)} \quad (1 \le j \le J) . \tag{23}$$

Equations (21) and (22) are systems of relations, which differ qualitatively in that (21) is purely deterministic while (22) contains random variables in its right hand side. We first treat (21) before returning to (22).

In matrix form (21) is

$$\widehat{\boldsymbol{P}}_i - \frac{\alpha_i}{W} \sum_\ell \boldsymbol{F}^{(\ell)} \widehat{\boldsymbol{P}}_\ell \ge \alpha_i \boldsymbol{\eta} \quad (1 \le i \le I) , \tag{24}$$

where

$$F_{jk}^{(\ell)} \triangleq G_{jk}^{(\ell)} K_{k\ell} w_{k\ell} , \tag{25}$$

and $\widehat{\boldsymbol{P}}_i = \{\widehat{P}_{ji}\}_j$ and $\boldsymbol{F}^{(l)} = \{F_{jk}^{(l)}\}_{j,k}$. It is helpful to recall, see (13), that $F_{jk}^{(l)}$ is the mean value of the power-independent, random component $M_{jk}^{(l)}$ in the interference at cell *j* due to all mobiles of class *l* in cell *k*. Hence, it is reasonable to call $\boldsymbol{F}^{(l)}$ the *mean class l, intercellular gain matrix.*

Now define

$$\alpha \triangleq \sum_{i=1}^{I} \alpha_i, \quad \alpha \boldsymbol{F} \triangleq \sum_{i=1}^{I} \alpha_i \boldsymbol{F}^{(i)} . \tag{26}$$

Hence $\boldsymbol{F}$ is the weighted *mean intercellular gain matrix,* where the weights are proportional to the class-dependent minimum CIR requirements, see (7).

From (24), we obtain

$$\left(\boldsymbol{I} - \frac{\alpha}{W}\boldsymbol{F}\right) \sum_\ell \boldsymbol{F}^{(\ell)} \widehat{\boldsymbol{P}}_\ell \ge \alpha \boldsymbol{F} \boldsymbol{\eta} . \tag{27}$$

We assume that $\boldsymbol{F}$ is an irreducible matrix and since it is also nonnegative, it has an eigenvalue of maximum modulus, called the Perron-Frobenius eigenvalue, which is real, positive and simple. We denote this eigenvalue by r_F, and we make the important assumption that

$$\frac{\alpha}{W} r_F < 1 . \tag{28}$$

Similar capacity defining constraints occur in price studies on deterministic narrowband systems and stochastic wideband systems.The condition (28) is especially noteworthy for the particular composition of the mean intercellular gain matrix $\boldsymbol{F}$, see (25) and (26). On the other hand, since it as a condition applied to a matix of mean gains, it is the least burdensome condition we may hope for in hte determination of the capacity of the stochastic system. Also, note that r_F and W are $O(K)$ and hence the left hand quantity in (28) is $O(1)$.

We recall, see for instance [MIT93], that (28) is equivalent to the existence of the element-wise nonnegative matrix $[\boldsymbol{I} - \frac{\alpha}{W}\boldsymbol{F}]^{-1}$. Hence, from (27),

$$\sum_\ell \boldsymbol{F}^{(\ell)} \widehat{\boldsymbol{P}}_\ell \ge \left(\boldsymbol{I} - \frac{\alpha}{W}\boldsymbol{F}\right)^{-1} \alpha \boldsymbol{F} \boldsymbol{\eta} . \tag{29}$$

It follows from (24) and (29) that

$$\widehat{\boldsymbol{P}}_i \geqq \alpha_i \left(\boldsymbol{I} - \frac{\alpha}{W} \boldsymbol{F} \right)^{-1} \boldsymbol{\eta} \quad (1 \leq i \leq I) . \tag{30}$$

Equivalently,

$$\widehat{\boldsymbol{P}}_i \geqq \frac{\alpha_i}{\alpha} \boldsymbol{p}^* \quad (1 \leq i \leq I) , \tag{31}$$

where

$$\boldsymbol{p}^* \triangleq \alpha \left(\boldsymbol{I} - \frac{\alpha}{W} \boldsymbol{F} \right)^{-1} \boldsymbol{\eta} . \tag{32}$$

This is an important result, which is now summarized.

Proposition 1 The dominant, leading order term in the expansion of P_{ji} is denoted by $\widehat{P}_{ji}$, which is required to satisfy the system of inequalities (24). We assume that the nonnegative matrix $\boldsymbol{F}$ defined in (26) is irreducible and that its Perron-Frobenius eigenvalue r_F satisfies (28). Then,

(i)
$$\widehat{\boldsymbol{P}}_i^* = \frac{\alpha_i}{\alpha} \boldsymbol{p}^* \quad (1 \leq i \leq I) \tag{33}$$

is the unique solution of the system of linear equations obtained by replacing inequalities by equalities in (24). Note that $\widehat{\boldsymbol{P}}_i^* > \boldsymbol{0}$.
(ii) The solution (33) is the minimal, i.e., Pareto optimal, solution to the system of inequalities in (24) in the following sense: any other solution to (24) will have every element not less and at least one element greater than the solution in (33). ∎

The product-form in (33), wherein the multiplicative factor dependent on the class index i is simply α_i, is of crucial subsequent importance. Note that $\boldsymbol{p}^*$, see (32), is of a familiar form, see for instance [FMI93, HAN93, MIT93], since it is the minimal power solution of a network construct with a single class having CIR requirement α and mean intercellular gain matrix $\boldsymbol{F}$.

We next turn to (22) and (23), and set $\widehat{\boldsymbol{P}}_l = \widehat{\boldsymbol{P}}_l^*$ $(1 \leq l \leq I)$. But $Z_{jk}^{(l)}$ $(1 \leq k \leq J,\ 1 \leq l \leq I)$ are independent and asymptotically normally distributed, with zero mean and unit variance. Hence ξ_j is asymptotically normally distributed, with zero mean and variance σ_j^2 where, with $\sigma_j \geq 0$,

$$\sigma_j^2 \triangleq \sum_k \sum_\ell \left(H_{jk}^{(\ell)} \right)^2 \gamma_{k\ell} w_{k\ell} (1 - w_{k\ell}) \left(\widehat{P}_{k\ell}^* \right)^2 \quad (1 \leq j \leq J) . \tag{34}$$

The quality of service requirement in (11) is satisfied asymptotically if

$$\Pr \left\{ Q_{ji} - \frac{\alpha_i}{W} \sum_k \sum_\ell F_{jk}^{(\ell)} Q_{k\ell} \geq a_i \xi_j \right\} \geq 1 - L_i \quad (1 \leq j \leq J, \quad 1 \leq i \leq I) . \tag{35}$$

This condition is equivalent to the deterministic condition

$$Q_{ji} - \frac{\alpha_i}{W} \sum_k \sum_\ell F_{jk}^{(\ell)} Q_{k\ell} \geq a_i \nu_i \sigma_j \quad (1 \leq j \leq J, \quad 1 \leq i \leq I) , \tag{36}$$

where v_i ($v_i > 0$) is obtained directly from L_i and the standard Gaussian distribution thus:

$$1 - L_i = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\nu_i} e^{-y^2/2} dy \quad (1 \leq i \leq I) . \tag{37}$$

That is, in (36) v_i is the multiple of the standard deviation of the asymptotic random variable ξ_j, which indicates how large the left-hand quantity has to be for the probabilistic requirement in (35) to be satisfied.

In matrix form the system of equations (36) is

$$\boldsymbol{Q}_i - \frac{\alpha_i}{W} \sum_{\ell} \boldsymbol{F}^{(\ell)} \boldsymbol{Q}_\ell \geq a_i \nu_i \boldsymbol{\sigma} \quad (1 \leq i \leq I) . \tag{38}$$

Then, using (26), we obtain

$$\left(\boldsymbol{I} - \frac{\alpha}{W} \boldsymbol{F} \right) \sum_{\ell} \boldsymbol{F}^{(\ell)} \boldsymbol{Q}_\ell \geq \sum_{\ell} a_\ell \nu_\ell \boldsymbol{F}^{(\ell)} \boldsymbol{\sigma} . \tag{39}$$

Since $(\boldsymbol{I} - \frac{\alpha}{W} \boldsymbol{F})^{-1}$ exists and is element-wise positive, it follows that

$$\sum_{\ell} \boldsymbol{F}^{(\ell)} \boldsymbol{Q}_\ell \geq \left(\boldsymbol{I} - \frac{\alpha}{W} \boldsymbol{F} \right)^{-1} \sum_{\ell} a_\ell \nu_\ell \boldsymbol{F}^{(\ell)} \boldsymbol{\sigma} . \tag{40}$$

Hence, from (38), we have

$$\boldsymbol{Q}_i \geq a_i \nu_i \boldsymbol{\sigma} + \frac{\alpha_i}{W} \left(\boldsymbol{I} - \frac{\alpha}{W} \boldsymbol{F} \right)^{-1} \sum_{\ell} a_\ell \nu_\ell \boldsymbol{F}^{(\ell)} \boldsymbol{\sigma} \quad (1 \leq i \leq I) . \tag{41}$$

We summarize in

Propostion 2 The second order term in the expansion of P_{ji} is denoted by Q_{ji}, which is required to satisfy the system of inequalities (38) that incorporates the minimal solution given in Proposition 1 for the dominant terms.

(i) $$\boldsymbol{Q}_i^* = a_i \nu_i \boldsymbol{\sigma} + \frac{\alpha_i}{W} \left(\boldsymbol{I} - \frac{\alpha}{W} \boldsymbol{F} \right)^{-1} \sum_{\ell} a_\ell \nu_\ell \boldsymbol{F}^{(\ell)} \boldsymbol{\sigma} \quad (1 \leq i \leq I) \tag{42}$$

is the unique solution of the system of linear equations obtained by replacing inequalities by equalities in (38). Note that $\boldsymbol{Q}_i^* \geqq \boldsymbol{0}$.

(ii) The solution (42) is the minimal, i.e., Pareto optimal, solution to the system of linear inequalities in (38), i.e., $\boldsymbol{Q}_i \geqq \boldsymbol{Q}_i^*$. ∎

3.3 Lifting the Veil of Asymptotics

In the final part of this section we give relations which do not make reference to the scalings in (18) and (19). Corresponding to (19), the asymptotic approximation to the minimal solution for the power from the leading first and second order terms in the asymptotic expansion is,

$$\boldsymbol{P}_i^* \sim \widehat{\boldsymbol{P}}_i^* + \frac{1}{\sqrt{K}} \boldsymbol{Q}_i^* \quad (1 \leq i \leq I) , \tag{43}$$

where $\hat{\boldsymbol{P}}_i^*$ is given by (32) and (33), and $\boldsymbol{Q}_i^*$ by (42). Also, from (42),

$$\frac{1}{\sqrt{K}}\boldsymbol{Q}_i^* = \alpha_i \left(\frac{1}{\alpha}\boldsymbol{q}^* + \frac{\sqrt{K}}{W}\nu_i\boldsymbol{\sigma} \right) \quad (1 \le i \le I) , \tag{44}$$

where $\boldsymbol{q}^*$ is the solution of the following equation,

$$\left(\boldsymbol{I} - \frac{\alpha}{W}\boldsymbol{F} \right) \boldsymbol{q}^* = \frac{\sqrt{K}\alpha}{W^2} \sum_{\ell} \alpha_\ell \nu_\ell \boldsymbol{F}^{(\ell)}\boldsymbol{\sigma} . \tag{45}$$

We let $\boldsymbol{\pi} \triangleq \boldsymbol{p}^* + \boldsymbol{q}^*$. Then, from (33) and (43)–(45),

$$\boldsymbol{P}_i^* \sim \alpha_i \left(\frac{1}{\alpha}\boldsymbol{\pi} + \frac{\sqrt{K}}{W}\nu_i\boldsymbol{\sigma} \right) \quad (1 \le i \le I) , \tag{46}$$

and $\boldsymbol{\pi}$ satisfies the following equation,

$$\left(\boldsymbol{I} - \frac{\alpha}{W}\boldsymbol{F} \right) \boldsymbol{\pi} = \alpha\boldsymbol{\eta} + \frac{\sqrt{K}\alpha}{W^2} \sum_{\ell} \alpha_\ell \nu_\ell \boldsymbol{F}^{(\ell)}\boldsymbol{\sigma} . \tag{47}$$

Now, from (26) and (46), we obtain

$$\alpha\boldsymbol{F}\boldsymbol{\pi} = \sum_{\ell} \alpha_\ell \boldsymbol{F}^{(\ell)}\boldsymbol{\pi} \sim \alpha \sum_{\ell} \boldsymbol{F}^{(\ell)} \left(\boldsymbol{P}_\ell^* - \frac{\sqrt{K}}{W}\alpha_\ell\nu_\ell\boldsymbol{\sigma} \right) . \tag{48}$$

Hence, from (47),

$$\boldsymbol{\pi} \sim \frac{\alpha}{W} \sum_{\ell} \boldsymbol{F}^{(\ell)}\boldsymbol{P}_\ell^* + \alpha\boldsymbol{\eta} . \tag{49}$$

Then, from (46), we have

$$\boldsymbol{P}_i^* \sim \alpha_i \left[\frac{1}{W} \sum_{\ell} \boldsymbol{F}^{(\ell)}\boldsymbol{P}_\ell^* + \boldsymbol{\eta} + \frac{\sqrt{K}}{W}\nu_i\boldsymbol{\sigma} \right] \quad (1 \le i \le I) . \tag{50}$$

Now, from (18) and (34), since $\boldsymbol{P}_\ell^* \sim \hat{\boldsymbol{P}}_\ell^*$ to lowest order, it may be verified that

$$\sqrt{K}\sigma_j \sim \psi_j(\boldsymbol{P}^*) \quad (1 \le j \le J) \tag{51}$$

where $\boldsymbol{P}^* = (\boldsymbol{P}_1^*, \ldots, \boldsymbol{P}_I^*)^T$ and $\boldsymbol{P} = (\boldsymbol{P}_1, \ldots, \boldsymbol{P}_I)^T$ are IJ-dimensional column vectors, and generally,

$$\psi_j(\boldsymbol{P}) \triangleq \left[\sum_k \sum_{\ell} \left(H_{jk}^{(\ell)} \right)^2 K_{k\ell} w_{k\ell} (1 - w_{k\ell}) (P_{k\ell})^2 \right]^{1/2} \quad (1 \le j \le J) . \tag{52}$$

Hence, from (50), we obtain expressions for the asymptotically ideal powers which depend only on unscaled quantities.

$$\boxed{\boldsymbol{P}_i^* \sim \frac{\alpha_i}{W}\left[\sum_{\ell} \boldsymbol{F}^{(\ell)}\boldsymbol{P}_\ell^* + \boldsymbol{\eta}W + \nu_i\boldsymbol{\psi}(\boldsymbol{P}^*)\right] \quad (1 \le i \le I)\,.} \tag{53}$$

Note that $\nu_i\psi(\boldsymbol{P}^*)$ is the term which represents the combined effect of the probabilistic quality of service requirement, and the variance of the interference at the cell sites, which, importantly, depends on all the powers at all the cell sites. Also, note that if $w_{kl} = 1$ $(1 \le k \le J,\ 1 \le l \le I)$, so that all the mobiles are always active, then $\psi(\boldsymbol{P}^*) = \boldsymbol{0}$.

4 ALGORITHM AND ITS CONVERGENCE

We now propose a distributed algorithm for power control, based on (53). In view of (5), (13) and (15), we introduce the measured quantities at cell j during time period n:

$$\overline{I}_j(n) = \sum_k \sum_\ell G_{jk}^{(\ell)} K_{k\ell} w_{k\ell} P_{k\ell}(n) + \eta_j W \quad (1 \le j \le J)\,, \tag{54}$$

and

$$\overline{V}_j(n) = \sum_k \sum_\ell \left(H_{jk}^{(\ell)}\right)^2 K_{k\ell} w_{k\ell}(1 - w_{k\ell}) P_{k\ell}^2(n) \quad (1 \le j \le J)\,, \tag{55}$$

where $P_{kl}(n)$ is the received power at cell site k from active mobiles of class l in the cell during time period n. That is, $\overline{I}_j(n)$ and $\overline{V}_j(n)$ are, respectively, the measured values of the mean and variance of the total interference at cell j in time period n. The distributed algorithm for power control that we propose is

$$\boxed{P_{ji}(n+1) = \frac{\alpha_i}{W}\left[\overline{I}_j(n) + \nu_i\sqrt{\overline{V}_j(n)}\right] \ (1 \le j \le J,\ 1 \le i \le I,)\ n = 0, 1, \ldots,} \tag{56}$$

where $\{\nu_i\}$, defined in (37), are fixed, known quantities. The adaptation of power is thus driven by *local* measurement of the mean and variance of the total interference.

In vector form, in view of (25) and (52), the algorithm is

$$\boldsymbol{P}_i(n+1) = \frac{\alpha_i}{W}\left[\sum_\ell \boldsymbol{F}^{(\ell)}\boldsymbol{P}_\ell(n) + \boldsymbol{\eta}W + \nu_i\boldsymbol{\psi}(\boldsymbol{P}(n))\right] (1 \le i \le I),\ n = 0, 1, \cdots. \tag{57}$$

Hence, from (53),

$$\begin{aligned}\boldsymbol{P}_i(n+1) - \boldsymbol{P}_i^* \\ = \frac{\alpha_i}{W}\left\{\sum_\ell \boldsymbol{F}^{(\ell)}\left[\boldsymbol{P}_\ell(n) - \boldsymbol{P}_\ell^*\right] + \nu_i\left[\boldsymbol{\psi}(\boldsymbol{P}(n)) - \boldsymbol{\psi}(\boldsymbol{P}^*)\right]\right\} (1 \le i \le I)\,.\end{aligned} \tag{58}$$

Let

$$\boldsymbol{\Phi} = (\alpha_1 \boldsymbol{I}, \ldots, \alpha_I \boldsymbol{I})^T \left(\boldsymbol{F}^{(1)}, \ldots, \boldsymbol{F}^{(I)} \right) , \tag{59}$$

and

$$\boldsymbol{\tau} = (\alpha_1 \nu_1 \boldsymbol{I}, \ldots, \alpha_I \nu_I \boldsymbol{I})^T , \tag{60}$$

where $\boldsymbol{I}$ denotes the $J \times J$ identity matrix. Then,

$$\boldsymbol{P}(n+1) - \boldsymbol{P}^* = \frac{1}{W} \boldsymbol{\Phi} [\boldsymbol{P}(n) - \boldsymbol{P}^*] + \frac{1}{W} \boldsymbol{\tau} [\boldsymbol{\psi}(\boldsymbol{P}(n)) - \boldsymbol{\psi}(\boldsymbol{P}^*)], \quad n = 0, 1, \cdots . \tag{61}$$

Hence,

$$\| \boldsymbol{P}(n+1) - \boldsymbol{P}^* \| \leq \frac{1}{W} \| \boldsymbol{\Phi} [\boldsymbol{P}(n) - \boldsymbol{P}^*] \| + \frac{1}{W} \| \boldsymbol{\tau} [\boldsymbol{\psi}(\boldsymbol{P}(n)) - \boldsymbol{\psi}(\boldsymbol{P}^*)] \| . \tag{62}$$

We will use the norm defined in terms of the element-wise positive Perron-Frobenius eigenvector $\boldsymbol{u}$ corresponding to Φ, i.e.,

$$\boldsymbol{\Phi u} = r_\Phi \boldsymbol{u}, \quad \boldsymbol{u} > \boldsymbol{0} . \tag{63}$$

Specifically,

$$\| \boldsymbol{x} \| = \max_{1 \leq m \leq IJ} |x_m| / u_m . \tag{64}$$

Let $\boldsymbol{u} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_I)^T$. Then, from (59),

$$\alpha_i \sum_{\ell} \boldsymbol{F}^{(\ell)} \boldsymbol{u}_\ell = r_\Phi \boldsymbol{u}_i . \tag{65}$$

Using (26), we obtain

$$(\alpha \boldsymbol{F} - r_\Phi \boldsymbol{I}) \sum_{\ell} \boldsymbol{F}^{(\ell)} \boldsymbol{u}_\ell = \boldsymbol{0} . \tag{66}$$

The element-wise Perron-Frobenius eigenvector $\boldsymbol{v}$ of $\boldsymbol{F}$ satisfies

$$\boldsymbol{F v} = r_F \boldsymbol{v}, \quad \boldsymbol{v} > \boldsymbol{0} . \tag{67}$$

Hence $r_\Phi = \alpha r_F$, since r_Φ and r_F are the eigenvalues of maximum modulus. We choose $\sum_\ell \boldsymbol{F}^{(\ell)} \boldsymbol{u}_\ell = r_\Phi \boldsymbol{v}$. Then, from (65), $\boldsymbol{u}_i = \alpha_i \boldsymbol{v}$. Also,

$$\| \boldsymbol{\Phi} [\boldsymbol{P}(n) - \boldsymbol{P}^*] \| \leq r_\Phi \| \boldsymbol{P}(n) - \boldsymbol{P}^* \| = \alpha r_F \| \boldsymbol{P}(n) - \boldsymbol{P}^* \| . \tag{68}$$

It is shown in the Appendix that

$$\| \boldsymbol{\tau} [\boldsymbol{\psi}(\boldsymbol{P}(n) - \boldsymbol{\psi}(\boldsymbol{P}^*)] \| \leq \beta \| \boldsymbol{P}(n) - \boldsymbol{P}^* \| , \tag{69}$$

where

$$\beta = \delta \sqrt{IJ} \sqrt{\sum_{i=1}^{I} \alpha_i^2 \nu_i^2} \; \frac{\max\limits_{1 \leq i \leq I} \alpha_i \max\limits_{1 \leq k \leq J} v_k}{\min\limits_{1 \leq i \leq I} \alpha_i \min\limits_{1 \leq k \leq J} v_k} , \tag{70}$$

and

$$\delta^2 = \max_{\substack{1\le k\le J\\ 1\le \ell\le I}} \sum_{j=1}^{J} \left(H_{jk}^{(\ell)}\right)^2 K_{k\ell} w_{k\ell}(1-w_{k\ell}), \quad \delta \ge 0 . \tag{71}$$

Hence, from (62), (68) and (69),

$$\|\boldsymbol{P}(n+1) - \boldsymbol{P}^*\| \le \frac{1}{W}(\alpha r_F + \beta)\,\|\boldsymbol{P}(n) - \boldsymbol{P}^*\|, \quad n = 0, 1, \cdots . \tag{72}$$

It follows, by induction, that

$$\|\boldsymbol{P}(n) - \boldsymbol{P}^*\| \le \left[\frac{1}{W}(\alpha r_F + \beta)\right]^n \|\boldsymbol{P}(0) - \boldsymbol{P}^*\| . \tag{73}$$

Proposition 3 If $(\alpha r_F + \beta)/W < 1$, then $\boldsymbol{P}(n)$ converges geometrically to $\boldsymbol{P}^*$. ■

We note that in our asymptotic analysis, with the scalings in (18), $\beta/W = O(1/\sqrt{K})$. Since, as noted in Sec. 3.2, $\alpha r_F / W = O(1)$, the above condition for geometric convergence is only slightly more stringent than $\alpha r_F / W < 1$, the condition which arises from consideration of just mean activity values. The condition in Proposition 3 is also the condition which determines the capacity of the cellular CDMA system.

5 CONCLUSIONS

We have given an algorithm in (56) for distributed power control in cellular CDMA networks which supports heterogeneous transmitting sources having diverse statistical/burstiness characteristics and quality of service requirements. The algorithm is derived from an asymptotic analysis based on a scaling appropriate for wideband systems. The analysis of the dynamical behavior of the algorithm considers the geometrically fast convergence of the adapted power levels to their optimum values.

A topic for future work is the extension of the approach developed in [MM096] and here to correctly incorporate random fluctuations in the transmission medium. Since we are already taking into account statistical fluctuations in the interference due to transmissions, it should be possible to extend the calculations to take into account explicitly fluctuations in the medium.

Appendix

We here derive the inequality (69). If x_{kl} and y_{kl} $(1 \le k \le J,\ 1 \le l \le I)$ are real, then Cauchy's inequality implies that

$$\left(\sum_{k=1}^{J}\sum_{\ell=1}^{I} x_{k\ell}^2\right)^{1/2} \left(\sum_{k=1}^{J}\sum_{\ell=1}^{I} y_{k\ell}^2\right)^{1/2} \ge \sum_{k=1}^{J}\sum_{\ell=1}^{I} x_{k\ell} y_{k\ell} . \tag{74}$$

It follows that

$$\left[\left(\sum_{k=1}^{J}\sum_{\ell=1}^{I} x_{k\ell}^2\right)^{1/2} - \left(\sum_{k=1}^{J}\sum_{\ell=1}^{I} y_{k\ell}^2\right)^{1/2}\right]^2 \le \sum_{k=1}^{J}\sum_{\ell=1}^{I} (x_{k\ell} - y_{k\ell})^2 . \tag{75}$$

Hence, from (52), we obtain

$$\sum_{j=1}^{J} [\psi_j(\boldsymbol{P}) - \psi_j(\boldsymbol{P}^*)]^2 \leq \sum_{j=1}^{J} \sum_{k=1}^{J} \sum_{\ell=1}^{I} \left(H_{jk}^{(\ell)}\right)^2 K_{k\ell} w_{k\ell}(1 - w_{k\ell}) (P_{k\ell} - P_{k\ell}^*)^2 \ . \tag{76}$$

Thus, with δ defined by (71),

$$\sum_{j=1}^{J} [\psi_j(\boldsymbol{P}) - \psi_j(\boldsymbol{P}^*)]^2 \leq \delta^2 \sum_{k=1}^{J} \sum_{\ell=1}^{I} (P_{k\ell} - P_{k\ell}^*)^2 \ . \tag{77}$$

Now, from (60),

$$(\|\boldsymbol{\tau}[\boldsymbol{\psi}(\boldsymbol{P}) - \boldsymbol{\psi}(\boldsymbol{P}^*)]\|_2)^2 = \sum_{i=1}^{I} \alpha_i^2 \nu_i^2 \sum_{j=1}^{J} \left[\psi_j(\boldsymbol{P}) - \boldsymbol{\psi}_j(\boldsymbol{P}^*)\right]^2 \ , \tag{78}$$

where

$$\|\boldsymbol{x}\|_2 = \left(\sum_{m=1}^{IJ} x_m^2 \right)^{1/2} \ . \tag{79}$$

From (77)–(79) we obtain

$$\|\boldsymbol{\tau}[\boldsymbol{\psi}(\boldsymbol{P}) - \boldsymbol{\psi}(\boldsymbol{P}^*)]\|_2 \leq \delta \left(\sum_{i=1}^{I} \alpha_i^2 \nu_i^2 \right)^{1/2} \|\boldsymbol{P} - \boldsymbol{P}^*\|_2 \ . \tag{80}$$

But,

$$\max_{1 \leq m \leq IJ} |x_m| \leq \|\boldsymbol{x}\|_2 \leq \sqrt{IJ} \max_{1 \leq m \leq IJ} |x_m| \ . \tag{81}$$

Hence, from (64), we have

$$\min_{1 \leq m \leq IJ} u_m \|\boldsymbol{x}\| \leq \|\boldsymbol{x}\|_2 \leq \sqrt{IJ} \max_{1 \leq m \leq IJ} u_m \|\boldsymbol{x}\| \ . \tag{82}$$

Note that, since $\boldsymbol{u} = (\alpha_1, \ldots, \alpha_I)^T \boldsymbol{v}$,

$$\min_{1 \leq m \leq IJ} u_m = \min_{1 \leq i \leq I} \alpha_i \min_{1 \leq k \leq J} v_k \ , \tag{83}$$

and

$$\max_{1 \leq m \leq IJ} u_m = \max_{1 \leq i \leq I} \alpha_i \max_{1 \leq k \leq J} v_k \ . \tag{84}$$

The inequality (69) follows from (80) and (82)–(84), where β is defined by (70) and (71).

References

[ARO96] M. Andersin and Z. Rosberg, "Transmission power cost of mobility in cellular networks," preprint, April 1996.

[BCM95] N. Bambos, S. C. Chen and D. Mitra, "Channel probing for distributed access control in wireless communication networks," Proc. Globecom 95.

[BCP95] N. Bambos, S. C. Chen and G. J. Pottie, "Radio link admission algorithms for wireless networks with power control and active link quality protection," Proc. INFOCOM 95.

[BPO92] N. Bambos and G. J. Pottie, "On power control in high capacity cellular radio networks," Proc. Globecom 92, vol. 2, pp. 863–867.

[EEV95] D. Everitt and J. Evans, "Traffic variability and effective interference for CDMA cellular networks," Proc. 9th ITC Specialists Seminar on Teletraffic Modelling and Measurements in Broadband and Mobile Communications, Leidschendam, The Netherlands, Nov. 1995, pp. 165–184.

[FEL68] W. Feller, *An Introduction to Probability Theory and Its Applications*, vol. 1, John Wiley, New York, 1968, p. 254.

[FMI93] G. J. Foschini and Z. Miljanic, "A simple distributed autonomous power control algorithm and its convergence," *IEEE Trans. Vehic. Tech.*, **42**(4), Nov. 1993, pp. 641–646.

[GVG93] S. A. Grandhi, R. Vijayan, D. J. Goodman and J. Zander, "Centralized power control in cellular radio systems," *IEEE Trans. Vehic. Tech.*, **42**(4), Nov. 1993, pp. 466–468.

[GZY95] S. A. Grandhi, J. Zander and R. Yates, "Constrained power control," *Wireless Personal Communications*, **1**(4), 1995.

[HAN93] S. V. Hanly, "Information Capacity of Radio Networks," Ph.D. Thesis, Cambridge University, Aug. 1993.

[HAN95] S. V. Hanly, "An algorithm for combined cell-site selection and power control to maximize cellular spread spectrum capacity," *IEEE J. Sel. Areas Commun.*, **13**(7), Sept. 1995, pp. 1332–1340.

[KYH95] P. S. Kumar, R. D. Yates and J. Holtzman, "Power control with bit error rates," Proc. MILCOM 95, San Diego, 1995.

[MIT93] D. Mitra, "An asynchronous distributed algorithm for power control in cellular radio systems," Proc. 4th WINLAB Workshop on Third Generation Wireless Information Networks, 1993, pp. 249-257.

[MMO94] D. Mitra and J. A. Morrison, "Erlang capacity and uniform approximations for shared unbuffered resources," *IEEE/ACM Trans. Networking*, **2**(6), Dec. 1994, pp. 558–570.

[MMO96] D. Mitra and J. A. Morrison, "A distributed power control algorithm for bursty transmissions in cellular, spread spectrum wireless networks," in *Wireless Information Networks*, (Proc. 5th WINLAB Workshop, 1995), Ed. J. M. Holtzman, Kluwer, 1996, pp. 201–212.

[NAL83] R. W. Nettleton and H. Alavi, "Power control for spread-spectrum cellular mobile radio system," *Proc. IEEE Vehic. Tech. Conf.*, **VTC-83,** 1983, pp. 242–246.

[VIT95] A. J. Viterbi, *CDMA, Principles of Spread Spectrum Communication*, Addison-Wesley, 1995.

[VVI93] A. M. Viterbi and A. J. Viterbi, "Erlang capacity of a power controlled CDMA system," *IEEE J. Sel. Areas Commun.*, **11**(6), Aug. 1993, pp. 892–900.

[YAT95] R. D. Yates, "A framework for uplink power control in cellular radio systems," *IEEE J. Sel. Areas Commun.*, **13**(7), Sept. 1995, pp. 6341–1347.

[ZAN92(a)] J. Zander, "Performance of optimum transmitter power control in cellular radio systems," *IEEE Trans. Vehic. Tech.*, **41**(1), Feb. 1992, pp. 57–62.

[ZAN92(b)] J. Zander, "Distributed cochannel interference control in cellular radio systems," *IEEE Trans. Vehic. Tech.*, **41**(3), Aug. 1992, pp. 305–311.

FAST POWER CONTROL IN CELLULAR NETWORKS BASED ON SHORT-TERM CORRELATION OF RAYLEIGH FADING

Z. Rosberg

Radio Commun. Systems Lab., KTH, Stockholm, Sweden
and
Haifa Research Lab., Science and Technology, IBM Israel
rosberg@haifa.vnet.ibm.com

Abstract: The fast transmission power control problem in a cellular network is studied when the signal attenuation is subject to a fast time-varying multipath Rayleigh fading. An asymptotic representation of the link gain evolution in time is first derived, and then used for constructing a Fast Time Variant distributed constrained Power Control (FTVPC). The proposed FTVPC power control successfully counteracts Rayleigh fading and demonstrates significant lower outage probabilities in comparison with the Constant-Received Power control (CRP). It is shown that a potential improvement of 4/3 in the spectrum utilization can be obtained by FTVPC compared to CRP. A comparison with a network *without Rayleigh fading* under a previously studied Time Variant Power Control (TVPC), is also made. Not surprisingly, it turned out that Rayleigh fading is very difficult to counteract, and it adds a substantial amount of outages.

INTRODUCTION

Transmitter power control has proven to be an efficient method to control cochannel interference in cellular PCS, and to increase bandwidth utilization. Power control can also improve channel quality, lower the power consumption, and facilitate network management functions such as mobile disconnection, hand-offs, base-station selection and admission control.

Distributed *Quality-based* power control, where transmitters adapt their power to meet a *signal quality target* at each receiver, have been extensively studied in the literature under various assumptions. Recent studies with continuous power levels and stationary conditions are given in [1, 6, 8, 9, 12, 15, 23, 21]. Algorithms for *random interference*, have been studied in [16, 18, 20]. Distributed power control with *discrete power levels and a SIR quality measure*, has been studied in [2, 22], and with *continuous power control and Bit Error Rate (BER)*

quality measure, in [14]. Power controls which cope with time-dependent link gains have been studied in [3, 4].

In all the studies above, it has been assumed that the **power control converges much faster compared to the changes in the link gains resulting from mobility.** This assumption has motivated a snapshot evaluation of the algorithms (where link gains are fixed in time), which over looks the random changes in the link attenuation. Such analysis under-estimates the outage and the signal quality measure target (see e.g., [4], and [3]). To compensate this under-estimation, coarse over-allocation of bandwidth is being used for designing a cellular network. In future PCS environments bandwidth would more carefully be allocated, and snapshot analysis should be refined by a time variant analysis to provide better quality measure targets.

Power controls which cope with time-dependent slow shadow fading have been studied in [4] and [3]. It has been assumed in these studies, that fast multipath Rayleigh fading is resolved by coding and interleaving, and therefore not included in the model. Preliminary results in [4] have revealed that the quality measure target must be set significantly higher than the target determined by a snapshot analysis. This study however, has not provided concrete rules to determine that quality target. In [3], an asymptotic representation of the link gain evolution in time (without Rayleigh fading) has been derived, and a framework to evaluate the channel quality in a time varying system is specified. Using that framework, a time-dependent power control algorithm which successfully copes with *slow shadow fading* has been devised and evaluated.

In this paper we study a network where the signal attenuation is subject also to a fast time-varying multipath Rayleigh fading. We show that our proposed Fast Time Variant distributed constrained Power Control (FTVPC) obtains an improvement of 4/3 in the spectrum utilization compared to CRP.

RADIO LINK MODEL

Consider a generic channel in a cellular network which is being accessed by N transmitters, where each is communicating with exactly one receiver. For the uplink case, the transmitters are the mobiles and the receivers are their corresponding base stations; and for the downlink case, their roles are reversed.

When transmitter j $(1 \le j \le N)$ is transmitting at time t, it uses power $p_j(t) \le \overline{p}_j$, where $\overline{p}_j$ is its maximum transmission power. Given that at time t, the link gain between transmitter j and receiver i is $g_{ij}(t)$ $(1 \le i, j \le N)$, the Signal to Interference Ratio at receiver i, $SIR_i(t)$, is defined by

$$SIR_i(t) = \frac{g_{ii}(t)\ p_i(t)}{\nu_i + \sum_{j:j \neq i}\ g_{ij}(t)\ p_j(t)} \qquad (1 \le i \le N)\ ,$$

where $\nu_i > 0$ is a time independent background noise power. The numerator is the received signal power at receiver i, and the denominator is the interference power experienced by receiver i. The SIR is a standard measure to evaluate the channel quality, and it is highly correlated with its error rate. Let γ_i be the SIR target for the channel between transmitter i and its corresponding receiver. We say that channel i is *supported at time t,* if

$$SIR_i(t) \ge \gamma_i\ .$$

To incorporate transmitters and receivers mobility resulting in the time variant link gains, we have to specify the process $(g_{ij}(t) \mid t \geq 0)$ for every pair $1 \leq i, j \leq N$. To investigate slow shadow and fast multipath fading affects on power control, we incorporate both into our link model. For every time instant t, the link gain (in power units) is modeled as a product of a distance dependent propagation loss, a slow shadow fading component, and a fast multipath Rayleigh fading. That is,

$$g_{ij}(t) = L_{ij}(t) \cdot S_{ij}(t) \cdot R_{ij}^2(t) \ . \tag{1.1}$$

We assume that $L_{ij}(t) = D_{ij}^{-\alpha}(t)$, where $D_{ij}(t)$ is the distance between transmitter j and receiver i at time t, and α is a propagation constant. The $S_{ij}(t)$ is assumed to be log-normally distributed with a log-mean of 0 *dB*, and a log-variance of σ_S^2 *dB*. That is,

$$Z_{ij}(t) \stackrel{\text{def}}{=} \frac{10}{\sigma_S} \log_{10} S_{ij}(t) \ ,$$

is the standard normal random variable. The third factor $R_{ij}^2(t)$, is the power attenuation due to fast multipath fading which is assumed to be governed by a Rayleigh distribution $R_{ij}(t)$, with $\sigma_R = \sqrt{0.5}$. That is, $E(R_{ij}^2(t)) = 1$.

We assume that the link gain processes are mutually independent, and the evolution of each link gain process $(g_{ij}(t) \mid t \geq 0)$ is governed by the following correlated processes.

Let v be the average mobile velocity, and t_0 be an arbitrary time reference. For the slow shadow fading process we assume that for every $t > 0$,

$$Z_{ij}(t_0 + t) = Z_{ij}(t_0) \cdot e^{-\frac{vt}{X}} + \tilde{Z}_{ij}(t) \cdot \left(1 - e^{-\frac{2vt}{X}}\right)^{\frac{1}{2}} \ , \tag{1.2}$$

where $\{\tilde{Z}_{ij}(t)\}$ are independent standard normal random variables, and X is the effective correlation distance of the shadow fading. The parameter X is environment dependent and determines the rate at which the shadow fading correlation is decreasing with the distance. We further assume that for every $t > t_0$, $\tilde{Z}_{ij}(t)$ and $Z_{ij}(t_0)$ are independent.

The time variant shadow fading process with the exponential correlation function in (1.2), has been proposed in [11] based on field experimental data. For notational convenience, we introduce the *normalized velocity*, $u = 2v/X$. The evolution in (1.2) then becomes,

$$Z_{ij}(t_0 + t) = Z_{ij}(t_0) \cdot e^{-\frac{ut}{2}} + N_{ij}(t) \cdot \left(1 - e^{-ut}\right)^{\frac{1}{2}} \ .$$

The time variant Rayleigh fading is represented by the following Jake's model, [13] (see also [7] and [17] for substantiating studies). Let $\{Z_{ij}^{(k)}(t)\}$, $k = 1, 2$, be two independent Gaussian processes with stationary means and variances, 0 and σ_R^2, respectively. Each process evolves independently of the rest of the system according to the following correlated process. For every time reference t_0, and time distance $t > 0$,

$$Z_{ij}^{(k)}(t_0 + t) = \rho(t) \cdot Z_{ij}^{(k)}(t_0) + \sqrt{1 - \rho^2(t)} \cdot \tilde{Z}_{ij}^{(k)}(t) \ , \quad k = 1, 2 \ , \tag{1.3}$$

where $\{\tilde{Z}_{ij}^{(k)}(t)\}$ are independent normal random variables with mean 0 and variance σ_R^2, and $\rho(t)$ is the zero order Bessel function of the first kind evaluated at $\frac{2\pi \cdot v \cdot f \cdot t}{C}$. That is,

$$\rho(t) = J_0\left(\frac{2\pi \cdot v \cdot f \cdot t}{C}\right) .$$

Here, f is the carrier frequency and C is the speed of light.

Observe that the Rayleigh process $\{R_{ij}(t) \mid t \geq 0\}$ is the envelope of the complex correlated Gaussian process $\{Z_{ij}^{(1)}(t) + i \cdot Z_{ij}^{(2)}(t) \mid t \geq 0\}$, whose evolution is modulated by the correlated processes defined in (1.3).

The derivation of a fast power control algorithm can be facilitated by an asymptotic representation of $g_{ij}(t_0 + t)$, for small t. Assume that mobiles move with constant velocity v. It has been shown in [3] that the distance dependent and the slow shadow fading factors have the following asymptotic representations.

$$\begin{aligned} L_{ij}(t_0 + t) &= L_{ij}(t_0) + o(ut) , \\ S_{ij}(t_0 + t) &= S_{ij}(t_0)\left(1 + a \cdot (ut)^{1/2} \cdot \tilde{Z}_{ij}(t)\right) \cdot 10^{o((ut)^{1/2})} , \end{aligned} \tag{1.4}$$

where $a = \frac{\sigma_S}{10} \ln(10)$, and $o(x)$ is a function of x with the property $\lim_{x \to 0} o(x)/x = 0$.

From the Rayleigh power definition and (1.3) we have,

$$\begin{aligned} R_{ij}^2(t_0 + t) &= \rho^2(t) \cdot R_{ij}^2(t_0) + (1 - \rho^2(t)) \cdot \tilde{R}_{ij}^2(t) \\ &\quad + 2\rho(t)\sqrt{1 - \rho^2(t)}\left(Z_{ij}^{(1)}(t_0) \cdot \tilde{Z}_{ij}^{(1)}(t) + Z_{ij}^{(2)}(t_0) \cdot \tilde{Z}_{ij}^{(2)}(t)\right) , \end{aligned} \tag{1.5}$$

where $\{\tilde{R}_{ij}^2(t)\}$ are independent exponential random variables with mean 1 (i.e., with the same marginal distribution as $R_{ij}^2(t)$).

The expansion in (1.5) is difficult to apply, and therefore we develop *a stochastically dominating approximation* whose tail distribution is very close to the approximated tail. As will be seen below this is required for our power control construction.

Applying numerical methods we obtain the following approximation for small values of t.

$$\hat{R}_{ij}^2(t_0 + t) = \rho^{a(t)}(t) \cdot R_{ij}^2(t_0) + b(t)\sqrt{1 - \rho^4(t)} \cdot \tilde{R}_{ij}^2(t) , \tag{1.6}$$

Here, $a(t)$ and $b(t)$ are constants which depend on the time horizon t, the carrier frequency and the mobile speed. E.g., for 900 MHz and 90 Km/h, $a(10ms) = 2$, $b(10ms) = 1.25$ and $a(t) = 1.1$, $b(t) = 1.25$, for $t = 0.1, 1\ ms$. For 30 Km/h, $a(10ms) = 2$, $b(10ms) = 1.5$, and $a(t) = 1.1$, $b(t) = 1.25$, for $t = 0.1, 1\ ms$. The complement distribution functions (CDF) of $R_{ij}^2(t_0 + t)$ and $\hat{R}_{ij}^2(t_0 + t)$ are depicted for different values of $(Z_{ij}^{(1)}(t_0) + i \cdot Z_{ij}^{(2)}(t_0))$ and t's, in Figures 1.1 - 1.2. It can be observed that our approximation objectives are met and that the smaller t is, the better is the approximation. For practical power

update rates in our system, we need the approximations for $t \leq 1\,ms$, which are very good.

To derive a fast power control we apply the following asymptotic representation of the link gains. From (1.4) and (1.6) it follows that for a short time horizon t, the distance dependent and the slow shadow fading time variation is negligible compared to the fast Rayleigh fading variation. Therefore, $g_{ij}(t_0 + t)$ can be approximated as follows. Let

$$\tilde{g}_{ij}(t) = L_{ij}(t) \cdot S_{ij}(t)$$

denote the link gain after averaging out the Rayleigh fading. Then we have from (1.4) and (1.6)

$$g_{ij}(t_0 + t) \approx \rho^{a(t)}(t) \cdot g_{ij}(t_0) + b(t)\sqrt{1 - \rho^4(t)} \cdot \tilde{g}_{ij}(t_0) \cdot \tilde{R}^2_{ij}(t) \ . \qquad (1.7)$$

Remark 1.0.1 *Observe that for a short time horizon, each link gain can be represented by weighted sum of its previous value, and an independent exponentially distributed random variable. The mean of the random component is a factor of the previous link gain without the Rayleigh fading.*

FAST TIME VARIANT POWER CONTROL

In this section, we propose a Fast Time Variant distributed Power Control (FTVPC) which utilizes the known correlation function of the multipath Rayleigh fading. This power control is build upon the following Distributed Constrained Power Control (DCPC) from [9]. For time variant link gains, DCPC updates the power according to

$$p_i(t + dt) = \begin{cases} \min\left\{\bar{p}_i,\ \frac{\gamma_i}{g_{ii}(t)}\left(\nu_i + \sum_{j:j\neq i}\ g_{ij}(t)\ p_j(t)\right)\right\}\ , & \text{if } i \in U(t), \\ p_i(t)\ , & \text{otherwise,} \end{cases}$$

where $U(t)$ is an arbitrary set of transmitters.

Observe that the right element within the curled parenthesis is the SIR target times the ratio between the interference power (including the background noise) at receiver i, and the link gain $g_{ii}(t)$. Since the interference power can be measured, and $g_{ii}(t)$ can be detected by the transmitter from the base station pilot signal (assuming a reciprocal system), this algorithm can be implemented in a distributed manner. Another SIR estimation procedure which is based on subspace estimation is devised in [5].

Since link gains vary in time, the interference and the gain of the allocated channel i, are evaluated by intensive sampling and averaging. Hence, the implemented DCPC would update the powers according to

$$p_i(t + dt) = \begin{cases} \min\left\{\bar{p}_i,\ \frac{\gamma_i}{\bar{g}_{ii}(t)}\left(\nu_i + \sum_{j:j\neq i}\ \bar{g}_{ij}(t)\ p_j(t)\right)\right\}\ , & \text{if } i \in U(t), \\ p_i(t)\ , & \text{otherwise,} \end{cases} \qquad (1.8)$$

where $\{\overline{g}_{ij}(t)\}$ are the link gain averages in a small time interval around t.

In this study we exploit the following fundamental property of the iterations in (1.8) (see [9] and [21]). *If the link gain averages are stationary in time (i.e., $\overline{g}_{ij}(t) = \overline{g}_{ij}$), then from every initial power vector, the iterated powers in (1.8) converge to a unique and positive fixed-point solution* $(p_1, p_2, \ldots, p_N)$,

$$p_i = \min\left\{\bar{p}_i,\ \frac{\gamma_i}{\overline{g}_{ii}}\left(\nu_i + \sum_{j:j\neq i}\overline{g}_{ij}\ p_j\right)\right\}\ ,\quad (1 \le i \le N)\ .$$

Stationary $\overline{g}_{ij}(t)$'s in (1.8) are essential for convergence, however in practice, they are not. We will show how to force "short-term stationarity" in the power iterations, and how to anticipate a near future random variation in the SIR values. Let t_0 be an arbitrary time reference and $(p_1(t_0), \ldots, p_N(t_0)))$ be its corresponding power vector. For every $t > t_0$ we will consider the iterated powers in (1.8), conditioned on the link gain values at time t_0. (In probability theory terminology, we examine the iterated powers given the sub-σ-field at time t_0.)

Let τ denote the time between two consecutive power updates, and assume intensive sampling within the time intervals $[t_0, k\cdot\tau]$, $k = 1, 2, \ldots$ For small values of τ's and k's (i.e., fast power control), (1.4) - (1.5), and highly significant sample averages imply that

$$\begin{aligned}
\overline{L}_{ij}(t_0 + k\tau) &\approx L_{ij}(t_0)\ ,\\
\overline{S}_{ij}(t_0 + k\tau) &\approx S_{ij}(t_0)\ ,\\
\overline{R}^2_{ij}(t_0 + k\tau) &\approx \overline{\rho}^2(k\tau)\cdot R^2_{ij}(t_0) + (1 - \overline{\rho}^2(k\tau))\ ,
\end{aligned} \tag{1.9}$$

where $\overline{X}(t_0 + k\tau)$is the sample average of $X(t)$ in the interval $[t_0, k\cdot\tau]$, and $x \approx y$ means that $|x - y|$ is negligible and vanishes as $\tau \to 0$. Thus, from (1.1) and (1.9) it follows that for small τ's and k's

$$\overline{g}_{ij}(t_0 + k\tau) \approx \overline{\rho}^2(k\tau)\cdot g_{ij}(t_0) + (1 - \overline{\rho}^2(k\tau))\cdot \tilde{g}_{ij}(t_0)\ . \tag{1.10}$$

That is, the average power attenuation after a short time period is a weighted average of the previous power attenuation, with and without Rayleigh fading. Hereinafter, t_0 and τ are fixed and will not be carried over the following notations. For every k, let $p_i(k) \stackrel{\text{def}}{=} p_i(t_0 + k\tau)$, $\overline{\rho}^2(k) \stackrel{\text{def}}{=} \overline{\rho}^2(k\tau)$, and denote

$$I_i(k) = \nu_i + \sum_{j:j\neq i}\overline{g}_{ij}(t_0 + k\tau)\ p_j(k)\ , \tag{1.11}$$

$$I^1_i(k) = \nu_i + \sum_{j:j\neq i} g_{ij}(t_0)\ p_j(k)\ , \tag{1.12}$$

$$I^2_i(k) = \nu_i + \sum_{j:j\neq i}\tilde{g}_{ij}(t_0)\ p_j(k)\ . \tag{1.13}$$

From (1.10) - (1.13) we have,

$$I_i(k) = \bar{\rho}^2(k) I_i^1(k) + (1 - \bar{\rho}^2(k)) I_i^2(k) \ . \tag{1.14}$$

Observe that with these notations, the power under DCPC in (1.8) can be approximated for small τ's and k's by,

$$p_i(k+1) = \min \left\{ \bar{p}_i, \ \frac{\gamma_i \cdot I_i(k)}{\bar{g}_{ii}(t_0 + k\tau)} \right\} \ , \ \ i \in U(t) \ .$$

The objective of our power update iterations is to drive the powers (within a pre-specified number of iterations k^*) to a level where the SIR targets at time $t_0 + k^*\tau$ are met with a high probability. From (1.7) and (1.14), it is apparent that the power iteration must involve $I_i^1(k)$, $I_i^2(k)$, $g_{ii}(t_0)$ and $\tilde{g}_{ii}(t_0)$. Hence, every transmitter-receiver pair needs to estimate them based on local measurements. The estimation procedure will be presented below after the power iteration definition.

Accounting for the random link gains, the channel quality requirement has to be probabilistic. We require that for every time reference t_0, the conditional probability given the link gains and powers at time t_0, will satisfy

$$P_{t_0}[SIR_i(t_0 + t^*) \geq \gamma_i] \geq 1 - \beta \ . \tag{1.15}$$

Here, $P_{t_0}[Y \in A] = P(Y \in A \mid \{g_{ij}(t_0)\}, \{p_i(t_0)\})$, and β is a given positive parameter.

To address this requirement we first use the approximation in (1.7) to project appropriate percentiles for the SIR levels at time $t_0 + \tau k^*$. We also assume ideal conditions where "convergence is reached" within k^* iterations.

Remark 1.0.2 *The practical role of k^* is twofold. One is to stabilize the power oscillation within the anticipated convergence horizon. The other is to be used as a time buffer between the power update rate and the SIR information update rate. Whereas the former is performed by the transmitter, the latter is done at the receiver and requires more resources. Thus, k^* can be set to the ratio between these two update rates.*

In practice, cochannel interference is dominated by a small number of interferers, to be denoted by N_0 (usually 2 to 4). Even when the dominant interferers are unknown, we may still regard the interference as if it is being generated by N_0 transmitters. Let $0 < \beta_1 < 1$ and $0 < \beta_2 < 1$, be two parameters such that

$$\beta_2 \ (1 - \beta_1) = 1 - \beta \ , \tag{1.16}$$

where $(1 - \beta)$ is the SIR quality parameter in (1.15).

The 1^{st} version of our *Fast Time Variant Power Control (FTVPC)* is parameterized by $(\beta_1, \beta_2, \tau, k^*)$. Let ξ_1 be the $1 - (1 - \beta_1)^{\frac{1}{N_0}}$ percentile of the exponential random variable with mean 1, $R_{ij}^2(t)$, and ξ_2 be its β_2 percentile. From the exponential distribution function and (1.16), we have

$$\begin{aligned} \xi_1 &= -\ln(1 - (1 - \beta_1)^{\frac{1}{N_0}}) \ , \\ \xi_2 &= -\ln \beta_2 = -\ln \tfrac{1-\beta}{1-\beta_1} \ . \end{aligned} \tag{1.17}$$

Define,

$$w(k^*, \xi_i) \quad = \frac{b(k^*\tau)\sqrt{1-\rho^4(k^*\tau)}}{\rho^a(k^*\tau)(k^*\tau)}\ \xi_i \ . \tag{1.18}$$

where $a(t)$ and $b(t)$ are defined in (1.6). FTVPC Version 1 is then specified as follows.

FTVPC Algorithm - Version 1

For any given set of parameters $(\beta_1, \beta_2, \tau, k^*)$, *each transmitter* i *updates its transmission power according to the following cyclic loop with* $k = 0, 1, ..., k^* - 1$:

$$p_i(k+1) = \begin{cases} \min\left\{\bar{p}_i,\ \dfrac{\gamma_i \cdot \left(I_i^1(k) + w(k^*, \xi_1) \cdot I_i^2(k)\right)}{g_{ii}(t_0) + w(k^*, \xi_2) \cdot \tilde{g}_{ii}(t_0)}\right\}\ , & \text{if } \ i \in U(t) \\ p_i(k)\ , & \text{otherwise.} \end{cases} \tag{1.19}$$

Remark 1.0.3 *Note that as in DCPC, powers are factored by the inverse proportion of their corresponding signal to interference ratio. However, under FTVPC, the interference and the signal powers are taken as weighted averages of their values with and without the Rayleigh fading. Further, percentiles factors are used to anticipate with high probability, the SIR values in the near future.*

It is straightforward to verify that the right-hand side of equation (1.19) satisfies the convergence conditions of Yates' Theorem in [21] (i.e., it is a "Standard interference function"). Therefore, under the ideal convergence conditions above, it follows from (1.19) that $(p_1(k^*), (p_2(k^*), \ldots, (p_N(k^*))$ is a fixed point solution satisfying

$$p_i(k^*) = \min\left\{\bar{p}_i,\ \frac{\gamma_i \cdot \left(I_i^1(k^*) + w(k^*, \xi_1) \cdot I_i^2(k^*)\right)}{g_{ii}(t_0) + w(k^*, \xi_2) \cdot \tilde{g}_{ii}(t_0)}\right\} \qquad (1 \le i \le N)\ , \tag{1.20}$$

for every given realization of the link gains and power vector at time t_0 .

Ignoring negligible elements (reasonable for small τ's and k^*'s), it is easy to verify by simple inequalities that for every channel i, if the equality in (1.20) is attained by

$$p_i(k^*) = \frac{\gamma_i \cdot \left(I_i^1(k^*) + w(k^*, \xi_1) \cdot I_i^2(k^*)\right)}{g_{ii}(t_0) + w(k^*, \xi_2) \cdot \tilde{g}_{ii}(t_0)} \tag{1.21}$$

(i.e., whose power converges to a value below the maximum transmission power), then (1.7), (1.12), (1.13), (1.16) - (1.18) and (1.21) imply that

$$P_{t_0}\left[SIR_i(t_0 + k^*\tau) \ge \gamma_i\right] \ge \beta_2 \cdot (1 - \beta_1) = 1 - \beta \ .$$

This is the condition we were aiming at in a time variant case. That is, within short time windows, powers are driven toward a level where the SIR targets are met with high probability.

Note that there are many pairs of (ξ_1, ξ_2) which satisfy (1.16). It can be shown that $(p_1(k^*), (p_2(k^*), \ldots, (p_N(k^*))$ is minimized, if in every iteration the

selected (ξ_1, ξ_2) minimizes the right-hand side of (1.19). This motivates the following algorithm.

FTVPC Algorithm - Version 2

For any given set of parameters (β, τ, k^*), *each transmitter i updates its transmission power according to the following cyclic loop with* $k = 0, 1, ..., k^* - 1$:

$$p_i(k+1) = \begin{cases} \min\left\{\bar{p}_i, \min\limits_{(\xi_1,\xi_2)} \frac{\gamma_i \cdot \left(I_i^1(k) + w(k^*,\xi_1) \cdot I_i^2(k)\right)}{g_{ii}(t_0) + w(k^*,\xi_2) \cdot \tilde{g}_{ii}(t_0)}\right\}, & \text{if } i \in U(t), \\ p_i(k), & \text{otherwise .} \end{cases}$$

From [21], a minimum of "standard interference functions" is also a "standard interference function", and therefore the convergence to a fixed point solution also applies.

Estimation Procedure

A method to estimate $I_i^1(k)$, $I_i^2(k)$, $g_{ii}(t_0)$ and $\tilde{g}_{ii}(t_0)$ which are required for the FTVPC algorithm, is as follows. Consider the equations

$$\begin{aligned} V_i(k) &= \bar{\rho}^2(k) I_i^1(k) + (1 - \bar{\rho}^2(k)) I_i^2(k), \\ W_i(k) &= \bar{\rho}_1^2(k) I_i^1(k) + (1 - \bar{\rho}_1^2(k)) I_i^2(k), \end{aligned} \tag{1.22}$$

where,

$$\begin{aligned} \bar{\rho}^2(k) &\stackrel{\text{def}}{=} \bar{\rho}^2(k \cdot \tau), \\ \bar{\rho}_1^2(k) &\stackrel{\text{def}}{=} \bar{\rho}^2((k - 0.5) \cdot \tau) \; .y \end{aligned} \tag{1.23}$$

Observe that $V_i(k)$ and $W_i(k)$ are the actual interference powers at time $(t_0 + k \cdot \tau)$ and at time $(t_0 + (k - 0.5) \cdot \tau)$, respectively (which are used by any DCPC variant algorithm). Since they can be measured by the i^{th} transmitter-receiver pair, and $\bar{\rho}^2(t)$'s are known and sharply decreasing with t, $I_i^1(k)$ and $I_i^2(k)$ can be resolved from the last two equations in (1.22). The values of $g_{ii}(t_0)$ and $\tilde{g}_{ii}(t_0)$ are resolved in a similar manner from two equations based on the measurements from the receiver pilot signal at time $(t_0 + k\tau)$ and time $(t_0 + (k - 0.5)\tau)$.

In cases where $\rho(t)$ varies in time, one may apply standard coefficient correlation estimators to adapt to its variation (see, e.g. [19]). To conclude, notice that the FTVPC algorithm applies two essential principles. One is the separation of the interference and the signal powers into two components, one with Rayleigh fading and another without it. The other principle is the construction of a confidence interval which anticipate the SIR values in the near future, with a desirable probability.

NUMERICAL RESULTS

In this section we evaluate the outage probabilities of our FTVPC power control algorithm in a Manhattan-like microcellular system, and compare them to those

of the Constant-Received Power control (CRP). We also reference them to the outage probabilities in a network **without Rayleigh fading** which uses the TVPC power control.

The cellular system under investigation is a typical metropolitan environment consisting of building blocks of a square shape. Each cell size is half a block in each direction, and base stations are placed in intersections with an asymmetric pattern as depicted in Figure 1.3. This cell structure is proposed in [10] and is termed *Asymmetric Half Square (AHS).* We assume that the channel assignment is fixed and divides the cells into 3 different cochannel cells in one case (AHS(1,1,3), see Figure 1.3), and 4 different cochannel cells in the other case (AHS(1,1,4), see Figure 1.4). In AHS(1,1,3) we simulated 48 cochannel cells, and in AHS(1,1,4) 64 cochannel cells.

From the measurement data in [11], we set the standard deviation of the shadow fading to $\sigma = 4$ *dB,* and the correlation distance to $X = 8.3$ *m.* The median *Signal to Noise Ratio (SNR)* at a cell border under the maximum transmission power is calibrated to 82 *dB.* That is, we take a strongly interference limited system.

We confine our examples to the uplink case and synchronous power updates. We assume a constant SIR target, $\gamma_i = \gamma^t$, for all mobiles. Under the Constant-Received Power control, the transmitters update their power according to

$$p_i(t + \Delta t) = \frac{C}{\overline{g}_{ii}(t)} ,$$

where the power target C is determined by a desired SNR of 53 *dB* above γ^t.

The algorithms are compared with respect to their *outage probabilities* using a value of $\beta = 0.05$, and evaluated by simulation. The outage probabilities are evaluated within time windows of 1, 2 and 5 seconds over more than 2,000 windows. Mobiles start from random positions which are independently sampled from a uniform distribution over each cell area, and travel along the streets with a constant speed of 30 *km/h.* At street corners, they turn left or right, or continue straight ahead with equal probabilities. Also, a mobile can hang up with certain probability, in which case it is replaced with a new one at a random location. A mobile transition to another cell is regarded as an hang up. We assume a carrier frequency of 900 *MHz.*

Within every simulated time window, we accumulate the proportion of time where $SIR_i(t) \geq \gamma_i$. If this proportion is greater than $1 - \beta$, then the corresponding mobile is supported, otherwise it is not supported. The proportion of non-supported time windows is defined as the outage probability, and it reflects the bit error rate. In the simulations, we use $\beta = 0.05$.

Findings

Figures 1.5 and 1.6 depict the outage probabilities as a function of the SIR target in the AHS(1,1,3) and AHS(1,1,4) cell plans, respectively. Three power controlled systems are compared, a system with Rayleigh fading under the CRP and the FTVPC power controls, and a system without Rayleigh fading under the TVPC power control. Each power control is represented by four update rates, 1000, 2500, 5000 and 10,000 updates per second. The FTVPC power control uses a convergence horizon parameter $k^* = 3$, for 1, 000 updates per

second, and $k^* = 5$, otherwise. These parameters turned out to best suit our network.

As to the statistical significant of the results, it should be noted that indistinguishable values are obtained for average window sizes of 1, 2 and 5 seconds, as well as for 1/3 of the simulated windows, except AHS(1, 1, 4) and update rate of 10,000. In the latter, we need the entire sample for convergence.

Most notable is the observation that Rayleigh fading is very difficult to counteract. The system without Rayleigh fading experiences significantly less outages compared to the system with Rayleigh fading. Another observation is that FTVPC substantially outperforms CRP. The outage probabilities of FTVPC are much lower than those of CRP. Note that by increasing the update rate, the outage probabilities under FTVPC reach to a level where error correcting codes can tolerate (20 – 25%). In the AHS(1, 1, 3) cell plan (reuse factor of 3), the outage probabilities at SIR targets 9, 10, 11 *dB* are 0.15, 0.175 and 0.225, respectively, when the update rate is 10,000 per second. With 5,000 updates per second the outage probabilities are 0.20, 0.265 and 0.33, respectively.

The outage probabilities become significantly lower when the reuse factor is 4 (in the AHS(1, 1, 4) cell plan). With 10,000 updates per second, the probabilities are 0.055, 0.08, 0.95 and 0.0125 for SIR targets of 9, 10, 11 *dB* and 12 *dB*, respectively. With 5,000 updates per second, the probabilities are 0.055, 0.065, 0.085 and 0.11, respectively. With 2,500 updates per second, the probabilities are 0.055, 0.08, 0.125 and 0.185, respectively.

The CRP fails to reduce the outage probabilities below a 0.25 level in all cases except those with a SIR target less than or equals 8 *dB* in a cell plan with reuse factor of 4. Thus, a potential improvement of 4/3 in the spectrum utilization can be obtained by replacing CRP with FTVPC.

Also observe that at an outage level of 25% in a cell plan with reuse factor of 4 and update rate of 5,000/sec, a SIR gain of about 7 *dB* can be achieved by FTVPC compared to CRP.

References

[1] M. Almgren, H. Andersson and K. Wallstedt. Power Control in a Cellular System. Proc. IEEE Veh. Tech. Conf., VTC-94, 833–837, 1994.

[2] M. Andersin, Z. Rosberg and J. Zander. Distributed Discrete Power Control in Cellular PCS. *Wireless Personal Communications* (to appear). Proc. Workshop on Multiaccess, Mobility and Teletraffic for Personal Communications, MMT'96, Paris, France, May 1996.

[3] M. Andersin and Z. Rosberg. Time Variant Power Control in Cellular Networks. Proc. 7th PIMRC Symposion, Oct. 1996.

[4] M. Andersin, M. Frodigh and K-E. Sunell. Distributed Radio Resource Allocation in Highway Microcellular Systems. Proc. 5th WINLAB Workshop, Apr. 1995.

[5] M. Andersin, N.B. Mandayam and R.D. Yates. Subspace Based Estimation of the Signal to Interference Ratio for TDMA Cellular Systems. Systems. *Wireless Networks* (to appear).

[6] N. Bambos and G. J. Pottie. On Power Control in High Capacity Cellular Radio Networks. Proc. 3rd WINLAB Workshop, Oct., 1992.

[7] R. H. Clarke. A Statistical Theory of Mobile Radio Reception. *Bell Sys. Tech. Journal,* Vol. 47, 957–1000, July, 1968.

[8] G. J. Foschini. A Simple Distributed Autonomous Power Control Algorithm and its Convergence. *IEEE Trans. on Veh. Tech.,* Vol. 42, No. 4, 1993.

[9] S. A. Grandhi, J. Zander and R. Yates. Constrained Power Control. *Wireless Personal Communications,* Vol. 1, No. 4, 1995.

[10] M. Gudmundson. Cell Planning in Manhattan Environments. Proc. IEEE Veh. Tech. Conf., VTC-92, 435–438, 1992.

[11] M. Gudmundson. Correlated Model for Shadow Fading in Mobile Radio Systems. *Electronics Letters,* Vol.27, No. 23, 2145–2146, 1991.

[12] S. V. Hanly. Information Capacity of Radio Networks. Ph.D. Thesis, Cambridge University, Aug. 1993.

[13] W. C. Jakes. *Microwave Mobile Communications.* Wiley, NY 1974.

[14] P. S. Kumar, R. D. Yates and J. Holtzman. Power Control with Bit Error Rates. Proc. MILCOM'95, San Diego, 1995.

[15] D. Mitra. An Asynchronous Distributed Algorithm for Power Control in Cellular Radio Systems. Proc. 4th WINLAB Workshop, 19-20, Oct. 1993.

[16] D. Mitra and J. A. Morrison. A Distributed Power Control Algorithm for Bursty Transmissions on Cellular, Spread Spectrum Wireless Networks. Proc. 5th WINLAB Workshop, Apr. 1995.

[17] J. D. Parsons and A. M. D.Turkmani. Characterization of Mobile Radio Signals: Model Description. IEE Proceeding-I, Vol. 138:549–556, Dec. 1991.

[18] Z. Rosberg and J. Zander. Power Control in Wireless Networks with Random Interferers. Radio Commun. Systems., KTH, Sweden, Dec. 1995.

[19] A. Sampath and J. M. Holtzman. Estimation of Maximum Doppler Frequency for Handoff Decisions. Proc. IEEE Veh. Tech. Conf., VTC-93, 859–862, 1993.

[20] S. Ulukus and R.u Yates. Stochastic Power Control for Cellular Radio Systems. Dept. of EE and CE, WINLAB, Rutgets University, Oct. 1996.

[21] R. Yates. A Framework for Uplink Power Control in Cellular Radio Systems. *IEEE JSAC,* Vol. 13, No. 7, Sept. 1995.

[22] J. Zander. Transmitter Power Control for Co-channel Interference Management in Cellular Radio Systems. Proc. 4th WINLAB Workshop, Oct. 19-20, 1993.

[23] J. Zander. Performance of Optimum Transmitter Power Control in Cellular Radio Systems. *IEEE Trans. on Veh. Tech.,* Vol. 41, No. 1, 1992.

[24] J. Zander. Distributed Cochannel Interference Control in Cellular Radio Systems. *IEEE Trans. on Veh. Tech.,* Vol. 41, No. 3, 1992.

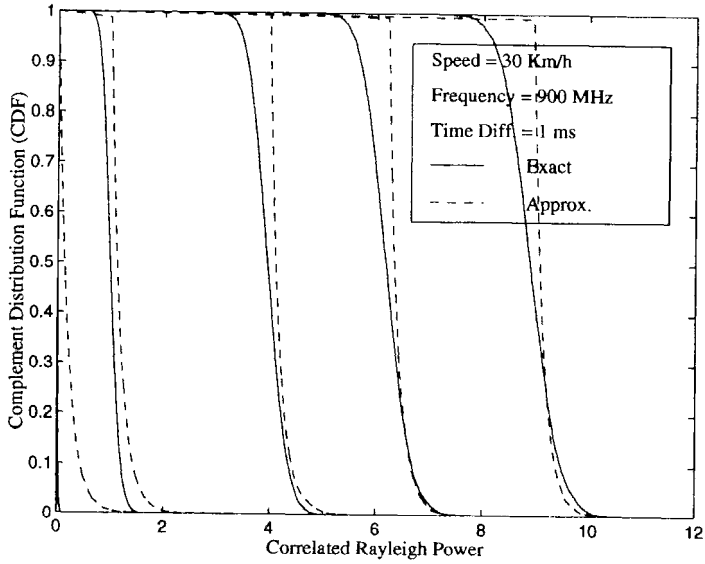

Figure 1.1 The complement distribution function of $R_{ij}^2(t_0 + t)$ and its approximation $\hat{R}_{ij}^2(t_0 + t)$, for various conditional values at time *t*.

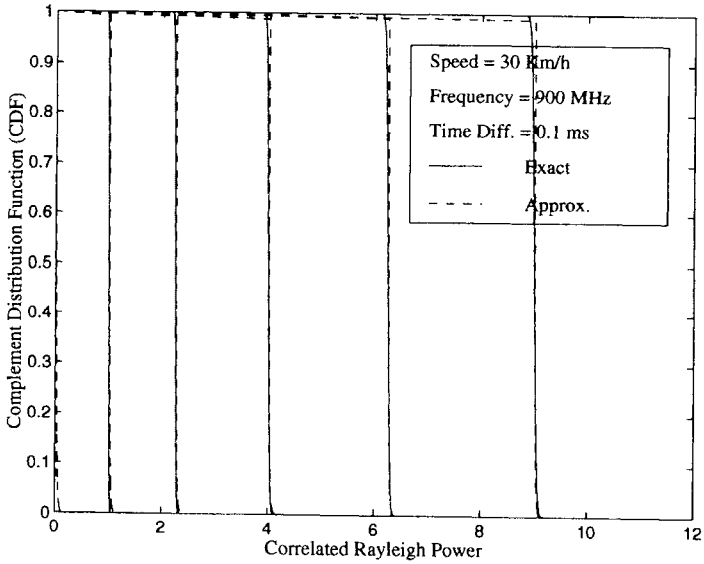

Figure 1.2 The complement distribution function of $R_{ij}^2(t_0 + t)$ and its approximation $\hat{R}_{ij}^2(t_0 + t)$, for various conditional values at time *t*.

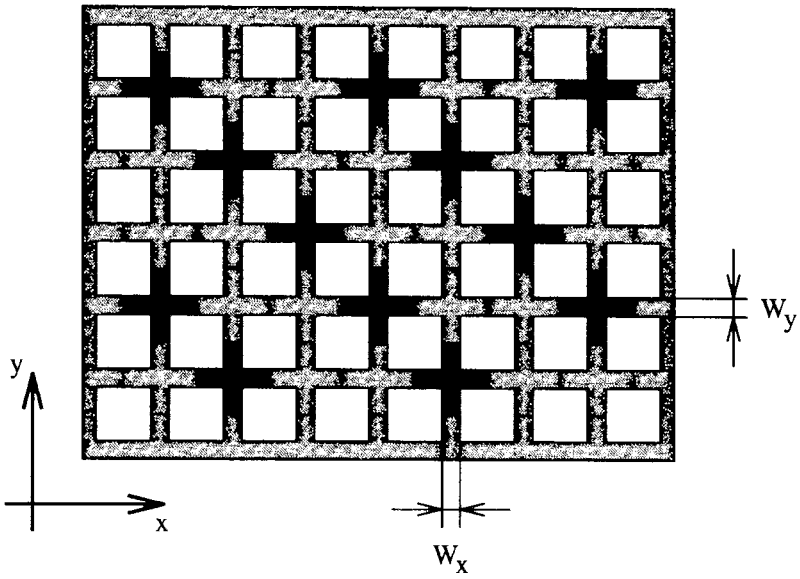

Figure 1.3 The asymmetric AHS(1, 1, 3) cell plan with cluster size N_c = 3. The dark crosses are the cochannel cells and the white squares are the buildings seen from above.

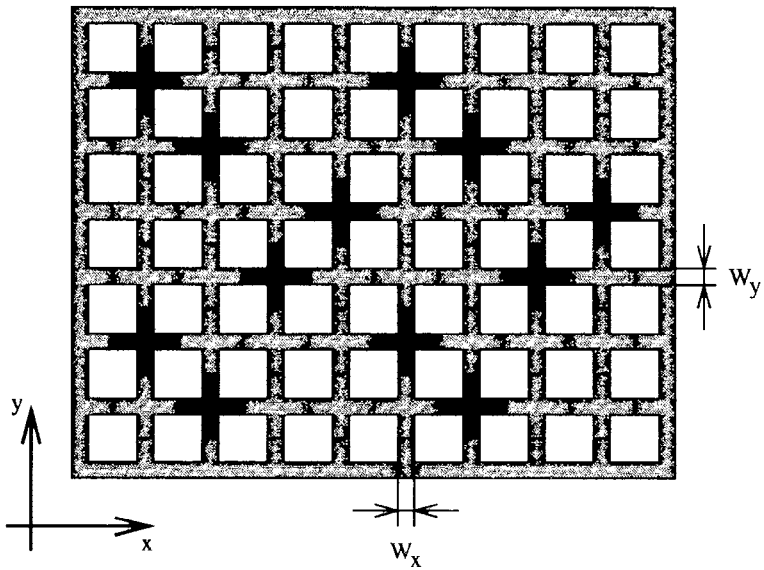

Figure 1.4 The asymmetric AHS(1, 1, 4) cell plan with cluster size N_c = 4. The dark crosses are the cochannel cells and the white squares are the buildings seen from above.

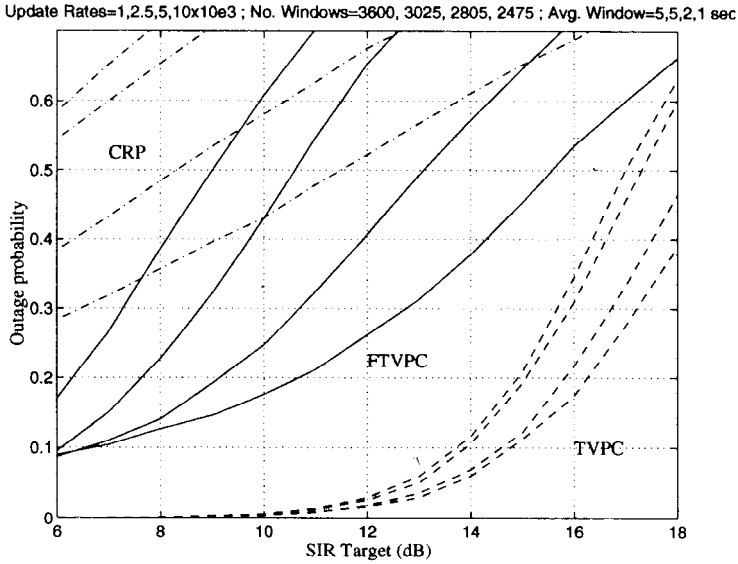

Figure 1.5 Outage probabilities under CRP and FTVPC with Rayleigh fading, and under TVPC without Rayleigh fading using cell plan AHS(1, 1, 3) with 48 cochannel cells.

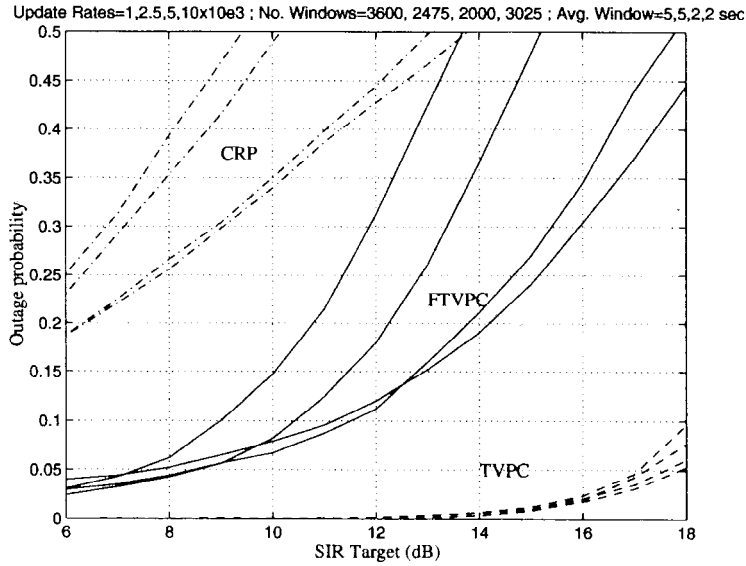

Figure 1.6 The outage probabilities as a function of the SIR target under CRP and FTVPC with Rayleigh fading, and under TVPC without Rayleigh fading using cell plan is AHS(1, 1, 4) with 64 cochannel cells.

This page intentionally left blank.

A SHORT LOOK ON POWER SAVING MECHANISMS IN THE WIRELESS LAN STANDARD IEEE 802.11[1]

Christian Röhl, Hagen Woesner, Adam Wolisz†

Technical University Berlin
Telecommunication Networks Group
roehl, woesner, wolisz@ee.tu-berlin.de

Abstract: This paper describes simulations of the power saving mechanism of the upcoming standard for wireless Local Area Networks IEEE 802.11[1]. They were performed in order to see how typical parameters influence the performance. Simulations were made for a ad-hoc-network with 8 stations. Figures for optimum Beacon intervals and ATIM window sizes were obtained.

1 INTRODUCTION

Wireless Local Area Networks (WLANs) are a rapidly growing area in networking. This is basically due to the upcoming of portable devices like notebooks and mobile phones. A key feature of these devices is the limited battery capacity, which limits their time in action. This results in a need of power saving mechanisms, which prolong the life time of the batteries.
The next chapter describes in short different ways to address the power saving problem. Chapter 2 shows the way power saving is implemented in the IEEE standard 802.11. After that we describe the simulated environment, the source model and the parameter set used for the simulations. In chapter 4 we come to

†also with GMD FOKUS

the simulations and their results. A discussion of the results and conclusions is shown in chapter 5.

2 POWER SAVING IN THE IEEE 802.11 DRAFT STANDARD

In general, the best way to save power for wireless communication devices would be to switch them off. Unfortunately, one can not do this without losing the capability to communicate in both directions, i.e. a station in this kind of a power saving mode would not know of any packets arriving for it at this time. Therefore there are two problems to be addressed in power saving:

1. How does a station in power save mode receive packets from other stations?

2. How does a station send to another station in power save mode?

Within the standard, the general idea is for all stations in PS mode to be synchronized to wake up at the same time. At this time there starts a window in which the sender announces buffered frames for the receiver. A station that received such an announcement frame stays awake until the frame was delivered. This is easy to be done in infrastructure networks, where there is a central access point, which is able to store the packets for stations in doze state and to synchronize all mobile stations. It is more difficult for ad-hoc networks, where the packet store and forward and the timing synchronization has to be done in a distributed manner.
Power Saving in IEEE 802.11 therefore consists of a Timing Synchronization Function (TSF) and the actual power saving mechanism. The TSF for an infrastructure network (the Point Coordination Function - PCF) can be seen in figure 1. The access point (AP) is responsible for generating beacons, which contain a valid time stamp beside other information. Stations within the BSS (Basic Service Set - a wireless cell) adjust their local timers to that time stamp. If the channel is in use after the beacon interval the AP has to defer its transmission until the channel is free again.

The situation is more complicated for an ad-hoc network (the Distributed Coordination Function - DCF, see figure 2). Due to the absence of a trusted authority the timers adjust in a distributed way: Every station is responsible for generating a beacon. After the beacon interval all stations compete for transmission of the beacon using the standard backoff algorithm. The first station ”wins” the competition and all others have to cancel their beacon transmission and to adjust their local timers to the time stamp of the winning beacon.

The power management in the PCF is simple due to the existence of the AP as central buffer for all packets to stations in doze mode. The AP transmits together with the beacon a so-called Traffic Indication Map (TIM). All unicast

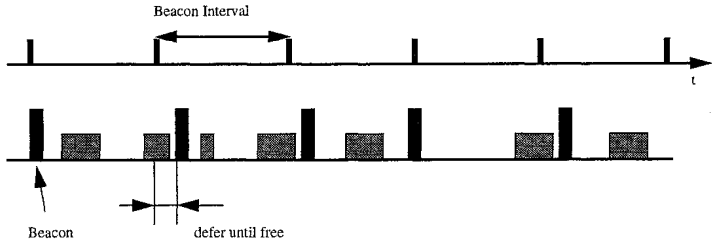

Figure 1: TSF for infrastructure networks in 802.11

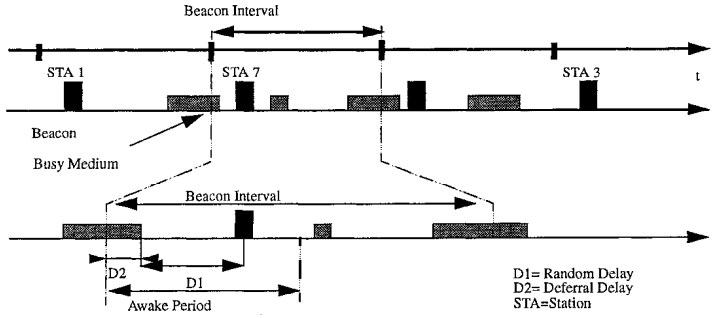

Figure 2: TSF for ad-hoc networks in 802.11

packets for stations in doze mode are announced in the TIM. The mobiles afterwards poll the AP for the packets. If broadcast/multicast frames are to be transmitted, they are announced by a Delivery TIM (DTIM) and sent immediately afterwards. Of course the stations in power save mode have to wake up short before the end of the beacon interval and to stay awake until the beacon transmission is over.

The power management for the DCF is based on the same distributed fashion as it is used for the TSF. Packets for a station in doze state have to be buffered by the sender until the end of the beacon interval. They have to be announced using Ad-hoc TIMs (ATIMs), which are transmitted in a special interval (the ATIM window) directly after the beacon. ATIMs are unicast frames which have to be acknowledged by the receiver. After sending the acknowledgment, the receiver does not fall back into doze state but stays awake and waits for

the announced packet (see figure 3). Both ATIMs and the data packets have to be transmitted using the standard backoff algorithm.

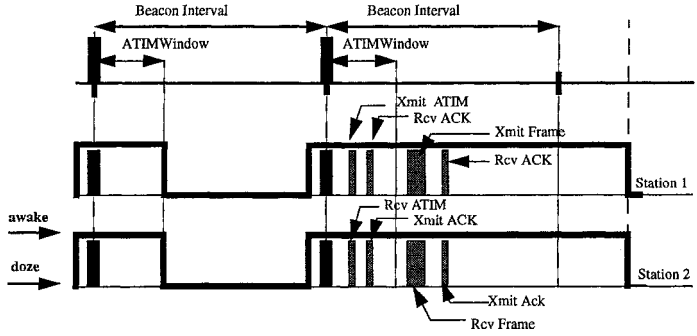

Figure 3: Power Management in the DCF of IEEE 802.11

Our aim was to tune the algorithm to get best values for the throughput of stations in power save mode and on the other hand for a maximum possible time in doze state. We chose the ratio of time in doze state vs. the time in active state as a measure for the quality of the power saving mechanism itself.

3 SIMULATION APPROACH

Our simulations were performed using a PTOLEMY model[2], which is described in greater detail in [3]. We used the appropriate values for the Direct Sequence Spread Spectrum (DSSS) physical layer. The simulation environment consists of 8 stations, which belong to an independent basic service set (IBSS). We did not consider any hidden terminals. Simulations with 1, 2, 4 and all 8 stations in power save mode were performed.

To model realistic traffic on a wireless LAN, we used trace files of an Ethernet [4], which were multiplexed using different start points in the file to lead to different, but predictable load scenarios[2]. We simulated overall offered loads of around 15, 30 and 60% of the raw physical throughput.

We made the assumption that power consumption is proportional to the time in active mode. Any additional effects which are depending on the PHY layer like equalization and on-off switching costs were not taken into account.

4 SIMULATION RESULTS

First we wanted to observe dependencies of the throughput compared to different window sizes for the beacon interval and ATIM window. As it may be expected, higher numbers of stations in power save mode lead to lower throughput. This is because of the overhead for each data packet, which consists of an ATIM and an ACK and two backoff sequences, regardless of the size of the packet to be transmitted[3]. It showed that there is a decrease in throughput for very small and very large ATIM window sizes (see figure 4). An ATIM window which is too small results in less ATIMs and therefore in less packets, which can be announced and transmitted. On the other hand, when the ATIM window is too large, more ATIMs are sent than there is actually time for the packets.

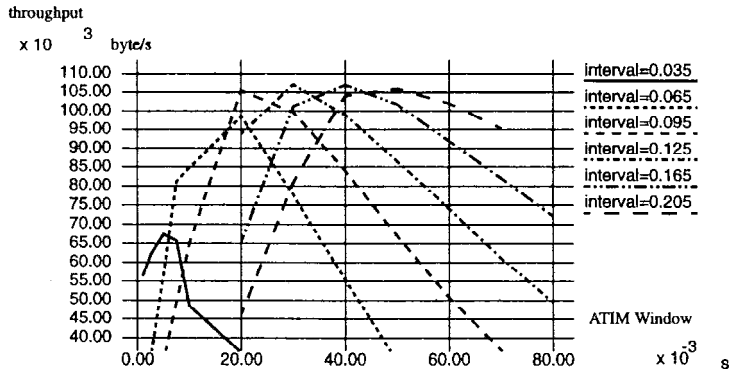

Figure 4: Throughput vs. ATIM window size for different beacon intervals, load=60.76%, 8 stations in power save mode

When we used a lower offered load for the simulations the results were basically the same, though throughput was constant for a broader range of ATIM window sizes. This was due to the fact that the channel could not be saturated any more.

In result it was obvious that the ATIM window size should be proportional to the beacon interval and that it should take 1/4 to 1/3 of the beacon interval. The next question was to determine the time in doze state in relation to the total time. In figure 5 one can see that the time in doze state increases when using shorter beacon intervals.

The simulation shown here was performed at an offered load of about 30%, because at a higher offered load a station would probably not fall into doze state very often. In figure 6 we simulated the same scenario as before, but with

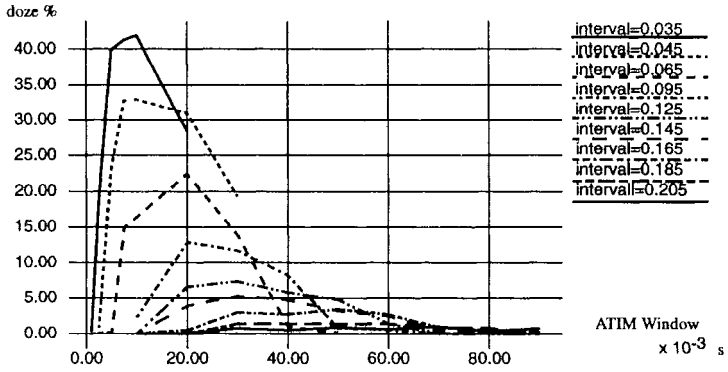

Figure 5: Percentage of time in doze state vs. ATIM window size for different beacon intervals, load=30.72%, 8 stations in PS mode

an offered load of about 15%. It shows that a station can stay in doze mode up to 70% of the time for beacon intervals small enough to allow for a fast transmission of the packet.

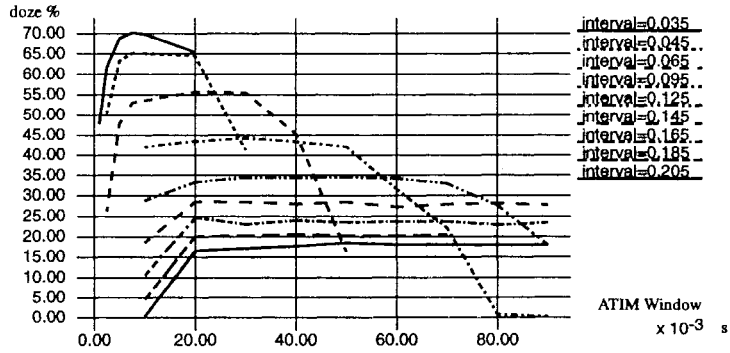

Figure 6: Percentage of time in doze state vs. ATIM window size for different beacon intervals, load=15%

The results can be explained as follows: The bigger the beacon interval the bigger the possibility that a station wishes to send during that time. This means that it has to transmit ATIMs in almost every beacon interval and to stay awake until the transmission is completed. The same applies for a receiving station.

In addition to that, more ATIMs per beacon interval have to be transmitted in bigger beacon intervals, which leads to higher collision rates and longer medium access times.

5 CONCLUSIONS

In this paper we presented simulations of the power saving mechanism in ad-hoc networks using the IEEE 802.11 standard. Work on this simulations started at a time when there was no recommendation for certain values of the parameters in the current version of the draft standard. In the meantime the values of interest in this scope are set to be 100 ms for the beacon interval and only 4 ms for the ATIM window.

Based on this work we can recommend figures for the ATIM window and beacon interval. Generally the mechanism gets less sensitive against the ATIM window size with higher values for the beacon interval. The simulations showed an optimum for the throughput at about 95 ms beacon interval. The ratio between ATIM window and beacon interval should be 1/4 to 1/3. While the first result corresponds to the value in the draft quite well, there is a “slight” difference in our recommended size for the ATIM window. This should be explained as follows: The recommended ATIM window size of only 4 ms (or Kμs, according to the standard) will be too small if there are many stations in power save mode or if the overall load is above 10%. We would definitely recommend for a higher value of the ATIM window parameter. On the other hand, there should be a means to adapt the value of this parameter to the offered load or, to be more exact, to the sum of the offered loads of the stations in PS mode.

The beacon interval should be smaller to lead to longer times in doze state. There has to be a trade-off between power saving and the overhead needed for it. If we would sacrifice about 10% in throughput we can save up to 30% more energy.

Notes

1. THIS WORK HAS BEEN SPONSORED IN PART BY THE DEUTSCHE FORSCHUNGSGEMEINSCHAFT DFG UNDER THE GRADUATE COLLEGE PROGRAM "COMMUNICATION BASED SYSTEMS"

2. The trace files were recorded on our institute-internal 10Base2 Ethernet

3. A comparable overhead applies for the optional RTS/CTS exchange, which in contrast depends on the packet length.

References

[1] The editors of IEEE: *Wireless LAN Media Access Control (MAC) and Physical Layer (PHY) Specification, IEEE 802.11 Draft Version 4.0*; May 1996

[2] *PTOLEMY*; www-site: http://www.ptolemy.eecs.berkeley.edu, Copyright 1990-96 The Regents of the University of California

[3] J.Weinmiller, H. Woesner, A. Wolisz: *Analyzing and Tuning the Distributed Coordination Function in the IEEE 802.11 Draft Standard*; June 1995 MASCOT '96; Feb. 1996, San Jose, California

[4] W. Leland et al.: *On the self-similar nature of Ethernet traffic*; IEEE Transactions on Networking, Vol. 2, No. 1, Feb.'94

[5] R. Caceres, V. Padmababhan: *Fast and Scalable Handoffs for Wireless Internetworks*; Proc. of MobiCom, November 1996

ENERGY MANAGEMENT IN WIRELESS COMMUNICATIONS

Michele Zorzi and Ramesh R. Rao

Center for Wireless Communications, University of California at San Diego
9500 Gilman Drive, La Jolla, CA 92093-0407, USA
fax: +l-619-534-2486 – http://www-cwc.ucsd.edu/
e-mail: {zorzi,rao}@ece.ucsd.edu

Abstract

In this paper, our goal is to study the "bits per joule" efficiency rating of a protocol in the wireless environment. We explore and compare three approaches to evaluating the energy efficiency and assess their accuracy and complexity. Although our technique allows us to accommodate other profiles, for concreteness we model the battery as a device that has the means to support the transmission of a fixed number of packets. We model the fading as a Markov channel, and we present some particular results for link error control protocols. For the particular examples considered, the exact recursive approach and an asymptotic approach were found to predict results that were very close. In addition, a lower bound and an approximation lead to analytical expressions. The quality of the bound and the accuracy of the approximation improves quite quickly as the amount of available energy increases.

1 INTRODUCTION

In the mobile wireless environment, portable communicators can be seen as devices that use a finite energy supply to transfer useful information. Thus there is an interest in understanding the useful "bits per joule" rating of a protocol in a particular environment. Direct measurements are an effective and accurate way of tracking the energy efficiency, but such approaches largely fail to provide insight on how to *design* more energy efficient protocols. Therefore, we turn to analytical approaches. In this regard, one of the key modeling issues is characterizing the energy consumed per packet transmitted. This quantity will depend on the nature of the circuits used to process and transmit the data. Furthermore, saturation effects in the power amplifier imply that the energy consumed per packet transmitted might depend non-linearly on the transmitted power level as well. Similarly, the particular type of battery used might dynamically impact the energy availability characteristic. These and other similar aspects of the modeling span a variety of disciplines including RF circuit design for low power circuits, battery technology, modulation signaling, resource allocation and protocol design.

The problem of error control over a wireless channel has been studied before. In [1], a simple link probing scheme, which slows down the transmission rate when the channel is impaired, was shown to be more energy efficient, in the sense of delivering a greater amount of data per energy unit with a slight loss in instantaneous throughput. Bambos and Rulnick have recently studied a related problem [2, 3]. Their work is concerned with optimizing the power control strategy to maximize the battery life (or, equivalently, to minimize the transmit power) under QoS constraints. Although the optimization problem introduced in [4] and the one considered in [2, 3] are not equivalent (having originated from the perspectives of error control and power control), they lead to similar conclusions about the way in which energy is to be managed.

The thrust of this paper is twofold. First, we extend from an analytical perspective the scope of the metrics that can be tracked in assessing the energy efficiency of the protocols. Second, we explore and compare three approaches to evaluating the energy efficiency, in order to assess their accuracy and complexity.

Although our technique allows us to accommodate other profiles, for concreteness we model the battery as a device that has the means to support the transmission of a fixed number of packets. Specifically, we assume the battery maintains a constant output power until it discharges and ceases to operate. We define the *energy efficiency* of a protocol, λ, as

$$\lambda = \frac{\text{total amount of data delivered}}{\text{total energy consumed}}. \tag{1}$$

We explicitly model the fading as a Markov channel. While we take into account random fluctuations of the channel attenuations, in this paper, we assume that the noise is of constant power.

We derive some particular results by focusing on link error control protocols. Our technique allows us to accommodate other metrics but for presentation purposes the operative metric for measuring consumed energy will be the total number of transmitted packets (successful and unsuccessful). We present results that track the total amount of useful data transferred as a function of the number of transmissions on the error prone wireless channel.

We describe our model for jointly tracking the evolution of a protocol and the available charge in Section 2 and introduce throughput and energy consumption metrics in Section 3. We describe a recursive technique in Section 4. This technique is most broadly applicable but it is also computationally intensive. Therefore, in Section 5 we present a new bounding technique and compare it with the asymptotic results previously described in [4] and the recursive technique of Section 4. In the concluding section we comment on the accuracy and complexity of the various bounding and approximation techniques.

2 THE GENERAL STOCHASTIC MODEL

Consider a discrete-time process which tracks the protocol evolution by means of a state machine. Let $I(n) \in \Omega$ denote the protocol state after the n-th transition, with Ω the (finite) space of all possible states. We will refer to n as "time," since the actual time scale (in slots) will not be explicitly considered in the protocol evolution (its effect is implicitly taken into account in the definition of the transition metrics).

In order to track the energy status, we consider another random process, $C(n)$, which tracks the energy drained from the battery in the time interval $[0, n]$. This variable is useful in tracking energy-dependent transitions as a function of the available charge. For analytical convenience, we assume that $C(n)$ is an integer without loss in generality.

The random process $X(n) = (I(n), C(n))$ jointly tracks the protocol state and the charge evolution and is useful in modeling the interaction between protocol evolution and available charge. We know that the protocol model is Markov. We shall also assume that the battery charge evolution is Markov, thereby allowing the battery to have memory [5, 6]. The assumption of a Markovian behavior for the battery energy process, $C(n)$, results in a Markov behavior for the whole process, $X(n)$.

Let ε be the total energy available, so that after ε energy units are consumed the process $X(n) = (I(n), C(n))$ stops. This might correspond to the complete battery discharge which causes the device to cease operation. Thus the process $X(n)$ is transient, with states (i, e), $i \in \Omega, \tau \geq \varepsilon$, being absorbing states. The need to allow $\tau \geq \varepsilon$, stems from the possibility of overshoot. By excluding such transitions we would underestimate the rate of transitions into the absorbing state. It may be covenient to think of states $(\boldsymbol{i}, \varepsilon)$, $\boldsymbol{i} \in \Omega$ as aggregate states entered by all transitions leading to energy

levels $\tau \geq \varepsilon$. Note that, since ε must be finite in any real system, the state space of $X(n)$, given by $\Omega \times \{0, 1, \ldots, \varepsilon\}$, is finite.

3 THROUGHPUT AND ENERGY CONSUMPTION METRICS

It is possible to define a set of *metrics* associated with the state transitions. In general, a transition metric is a (possibly random) variable $\xi_{ab}(n)$, such that, given that the n-th transition occurs from state a to state b, $\xi_{ab}(n)$ is conditionally independent of the past and future evolution of the process. Also, we consider the cumulative metric (also called *reward* [7]) over transitions 1 to n.

By appropriately defining metrics and by studying the corresponding reward earned throughout the transient evolution of the process (e.g., the "connection lifetime"), we study the throughput performance and the energy efficiency.

Let $a = (i, \tau)$ and $b = (j, v)$ be two states with $P_{ab} = P[X(n) = b \,|X(n-1) = a] > 0$. We define the following metrics on transition ab:

- R_{ab}: the number of successful transmissions
- C_{ab}: the number of consumed energy units.

Note that, according to the definition of the protocol model, R_{ab} only depends on the protocol states so that $R_{ab} = R_{ij}$. (The only care to be taken is that if b is an absorbing state, then the reward on transition ab may not be earned, as the evolution is terminated before the transition itself is completed, and $R_{ab} = 0$. This fact will be neglected in the following for ease of notation.) On the other hand, the consumed energy in general depends on the battery status, as this may affect the output power. As an example, if T_{ij} is the number of packet transmissions associated with the transition from i to j, and if the energy is consumed only due to radiated power, then we have $C_{ab} = T_{ij}\, \omega(\tau)\Delta T$, where $\omega(\tau)$ is the transmit power (which in general depends on the battery state, τ), and ΔT is the slot duration.

Note that the following relationship must be true if $P_{ab} > 0$.

$$\tau + C_{ab} = v. \tag{2}$$

For simplicity, we assume here that C_{ab} is a constant, so that given i, j and τ there is a unique b for which $P_{ab} > 0$. The case in which C_{ab} is a random variable can be incorporated in the model by slightly complicating the transition structure of the process. This extension appears conceptually straightforward, and will not be considered here in detail for ease of notation.

Since v is deterministically related to i, j and τ, C_{ab} and P_{ab} can be alternatively expressed as

$$C_{ab} = C_{ij}(\tau), \qquad P_{ab} = P_{ij}(\tau), \tag{3}$$

which are sometimes easier to deal with, as they explicitly contain the information about the protocol state and the energy level.

Note that, although a and j uniquely determine b, it is not true in general that b and i will uniquely determine a. In fact, if

$$\tau_1 + C_{ij}(\tau_1) = \tau_2 + C_{ij}(\tau_2) = v, \quad \tau_1 \neq \tau_2, \tag{4}$$

both states $a_1 = (i, \tau_1)$ and $a_2 = (i, \tau_2)$ can lead to state $b = (j, v)$ with a single transition. For later use, we define for state $b = (j, v)$ the set of states from which it can be reached with a single transition:

$$\begin{aligned} \Omega(j,v) &= \{(i,\tau) : P[X(n) = (j,v) | X(n-1) = (i,\tau)] > 0\} \\ &= \{(i,\tau) : C_{ij}(\tau) = v - \tau \text{ and } P_{ij}(\tau) > 0\}. \end{aligned} \tag{5}$$

We remark that, in the case where the battery output power can be made independent of the energy status, all the above quantities are independent of τ, and $\Omega(j,v) = \{(i, v - C_{ij}), i \in \Omega\}$.

4 DYNAMIC EVOLUTION OF X

In order to study the process behavior during the transient evolution, we track jointly the reward process and the energy consumption process. Define the process $R(n)$ as the total reward earned in transitions 1 through n. Define

$$\phi_{ij}(\rho, v, n) = P[R(n) = \rho, X(n) = (j,v) | R(0) = 0, X(0) = (i,0)]. \tag{6}$$

Then, the following recursive relationship can be established by conditioning on $X(n-1)$ and applying the theorem of total probability:

$$\phi_{ij}(\rho, v, n) = \sum_{(m,\tau)\in\Omega(j,v)} P_{mj}(\tau)\phi_{im}(\rho - R_{mj}, \tau, n-1) + \delta_{ij}\delta(n)\delta(\rho)\delta(v), n \geq 0. \tag{7}$$

where δ_{ij} is the Kronecker symbol, $\delta(x) = \delta_{x0}$, and where it is understood that $\phi_{ij}(\rho, v, n) = 0$ if $v < 0$ or $\rho < 0$ (this will also apply to all functions throughout the paper, unless otherwise indicated).

The recursion (7) can be used to find the probability distribution of the reward earned throughout the whole evolution, i.e., from time 0 to when an absorbing state (i.e., a state with $C(n) = \varepsilon$) is entered. In fact, if we let R be this total reward, we have

$$P[\mathcal{R} = \rho] = \sum_{n=0}^{\infty} P[R(n) = \rho, C(n) = \mathcal{E}] = \sum_{i,j\in\Omega} \sum_{n=0}^{\infty} \pi_i^{(0)} \phi_{ij}(\rho, \mathcal{E}, n), \tag{8}$$

where $\pi_i^{(0)}, i \in \Omega$, is the probability distribution according to which the initial protocol state is chosen at time zero, i.e.,

$$\pi_i^{(0)} = P[I(0) = i] = P[X(0) = (i,0)]. \tag{9}$$

Closed-form solution of the recursion (7), e.g., by transform techniques, appears to be a rather difficult task in general, due to the multiple variables involved. On the other hand, if we are only interested in the statistics of R, a simplified recursion can be found by summing (7) over n. Further, if we are interested in the moments of R, we can also eliminate ρ by multiplying both sides of (7) by the appropriate power of ρ and summing it from 0 to ∞. As an example, consider the mean $E[R]$. Define

$$\theta_j(v) = \sum_{n=0}^{\infty} \sum_{\rho=0}^{\infty} \rho \sum_{i \in \Omega} \pi_i^{(0)} \phi_{ij}(\rho, v, n). \tag{10}$$

Then, we obtain from (7) and (10):

$$\begin{aligned} \theta_j(v) &= \sum_{n=0}^{\infty} \sum_{\rho=0}^{\infty} \rho \sum_{i \in \Omega} \pi_i^{(0)} \left[\sum_{(m,\tau) \in \Omega(j,v)} P_{mj}(\tau) \phi_{im}(\rho - R_{mj}, \tau, n-1) \right. \\ &\qquad \left. + \delta_{ij} \delta(n) \delta(\rho) \delta(v) \right] \\ &= \sum_{n=0}^{\infty} \sum_{(m,\tau) \in \Omega(j,v)} P_{mj}(\tau) \sum_{\rho=0}^{\infty} (\rho - R_{mj} + R_{mj}) \\ &\qquad \times \sum_{i \in \Omega} \pi_i^{(0)} \phi_{im}(\rho - R_{mj}, \tau, n-1) \\ &= \sum_{(m,\tau) \in \Omega(j,v)} P_{mj}(\tau) \left[\theta_m(\tau) + R_{mj} \chi_m(\tau) \right], v \geq 0, \end{aligned} \tag{11}$$

where

$$\chi_m(v) = \sum_{i \in \Omega} \sum_{n=0}^{\infty} \pi_i^{(0)} \sum_{\rho=0}^{\infty} \phi_{im}(\rho, v, n), \tag{12}$$

is the probability that state (m, v) will be visited given that the initial state of the protocol is chosen according to the probability distribution $\pi^{(0)}$.

Finally, the average reward can be found from (8), (10) as

$$E[\mathcal{R}] = \sum_{j \in \Omega} \theta_j(\mathcal{E}), \tag{13}$$

where all "overshoot" states (j, v), $v \geq \varepsilon$ are collapsed into (j, ε).

Similar relationships can be established for the variance of R and in principle for any moment. This approach via recursive technique is therefore very powerful in that it allows to easily find the statistics of R, while taking into account a possible detailed dependence of the parameters on the energy status, and is proposed here as a general

tool for the exact evaluation of the energy performance. However, even running the simplified recursion (11) requires a number of operations which increases linearly with ε (as opposed to ε^2 for (7)). In the following, we will study some asymptotic properties of the process $X(n)$, which will help us find simple approximations for the quantities of interest in the situations where direct solution of the recursion is impractical, i.e., for large ε.

5 SPECIAL CASE: FLAT PROFILE

As a first step, we will restrict our analysis to the case of *flat power profile*, in which the transition probabilities and metrics do not depend on the energy state, τ. This case is an idealization of a regulation mechanism which tries to keep the battery output power constant while the battery discharges, and can be taken as a fair approximation of real devices. Also, it can be considered as a good model if the focus is on a single connection which is allocated a small portion of the total battery energy, so that the battery behavior can be considered ideal throughout its duration. Furthermore, the results obtained in this case will provide a useful tool for approximate analysis of the general case.

The first obvious simplification which is obtained in this case is that the recursion coefficients are constant numbers, i.e., $P_{ij}(\tau) = P_{ij}$ and $C_{ij}(\tau) = C_{ij}$ for all values of τ, except for those transitions leading to the absorbing states, where some boundary conditions must be satisfied. This simplifies the description of the evolution, but does not decrease the compexity of the recursion. More significant simplifications are discussed next.

5.1 Asymptotic result

Due to the independence of the metrics of the energy dimension, we can simplify the analysis by collapsing all states (i, τ) into a single state i. The process $I(n)$ is now a homogeneous Markov chain (it is non homogeneous in the general case), and can be studied more easily. In particular, let us arbitrarily choose one state, $l \in \Omega$. Then, the visits of the process $I(n)$ to state l are renewal instants, and the system evolution repeats itself (in a statistical sense).

More specifically, consider the renewal instants, i.e., of the sequence of integers $\{n_k, k = 0,1, \ldots\}$, such that $n_k < n_{k+1}$, $I(n_k) = l$ and $I(n) \neq l$ for $n_k, < n < n_{k+1}$, with $I(n_0) = I(0) = l$ by definition (extension to random $I(0)$ is also possible). Let R_k and C_k be the reward earned and the energy consumed between the $(k-1)$-st and the k-th visit, i.e., in $(n_{k-1}, n_k]$ (also called the *k-th interrenewal time*), and consider the vector process

$$\mathcal{X}(k) = (\mathcal{R}(k), \mathcal{C}(k)) = \left(\sum_{i=1}^{k} \mathcal{R}_i, \sum_{i=1}^{k} \mathcal{C}_i \right). \tag{14}$$

In an attempt to avoid confusion, we will use the index k for renewals, as opposed to n which was used for transitions. Also, calligraphic letters will be used for variables in this renewal process setting (e.g., $R(k) = R(n_k)$, and R_k and R_{ij} denote the reward earned during the k-th interrenewal interval and during transition ij, respectively).

The problem of finding the total reward is equivalent to finding the value of $R(k)$ given that $C(k-1) < \varepsilon \leq C(k)$. In this context, $X(k)$ is a stopped random walk [8]. If ε is the total amount of energy available, it is possible to show, using renewal theory, that

$$\lambda_\infty = \lim_{\mathcal{E}\to\infty} \frac{\mathcal{R}(k)}{\mathcal{C}(k)} = \frac{\sum_{j,k\in\Omega} \pi_j R_{jk}}{\sum_{j,k\in\Omega} \pi_j C_{jk}}, \tag{15}$$

where π_j, $j \in \Omega$, is the steady-state distribution of the chain $I(n)$. Therefore, from simple Markov analysis, it is possible to find the asymptotic energy efficiency of the scheme, λ_∞. The total reward is in this case given by

$$\mathcal{R} = \lambda_\infty \mathcal{E}, \tag{16}$$

and is of course infinite for $\varepsilon \to \infty$. However, the result (16), which is rigorously true only in the limit, holds asymptotically as ε is large. In particular, it is expected to hold in practice, where the batteries are supposed to be able to transmit a large number of data units. Some experiments we made comparing (16) with the results obtained by running the recursion show that even for moderate values of ε the difference is negligible.

5.2 Finite values of ε

To study the behavior of the process in further detail, we consider here the rate of convergence to the above asymptotic result. Note that the value of λ_∞ obtained in (15) is *not* a random variable. On the other hand, for every finite value of ε, the total reward R and the energy efficiency λ are random variables. In this subsection, we are interested in the statistics of R for finite ε.

Let L denote the random variable that represents the renewal cycle in which the total consumed energy reaches the limit of ε. From the renewal property, the random vectors (R_i, C_i) and (R_j, C_j) are independent and identically distributed for $i \neq j$ and, in particular, $E[R_i] = r$, $E[C_i] = c$ for all i. However, R_i and C_i are allowed to be (and often are) statistically dependent. Since

$$\{L = m\} = \{\mathcal{C}(m-1) < \mathcal{E}\} \bigcap \{\mathcal{C}(m) \geq \mathcal{E}\}, \tag{17}$$

L is a stopping time [7] relative to the family $\{(R_i, C_i), i \geq 1\}$. Note that the exact value of $C(L)$ is difficult to determine, due to the fact that the L-th renewal interval is

not "typical." Therefore, from (17) we can only conclude that $E[C(L)] \geq \varepsilon$, so that

$$\mathcal{E} \leq E\left[\sum_{i=1}^{L} \mathcal{C}_i\right] \leq E[L]c, \tag{18}$$

$$E[L] \geq \frac{\mathcal{E}}{c}. \tag{19}$$

Now we are interested in finding the total number of rewards, R, earned during the time the energy ε was consumed. To this aim, let us again consider the vector process (R_i, C_i), $i \geq 1$. Since L is a stopping time, we have from Wald's equation that [7]

$$E[\mathcal{R}(L)] = E\left[\sum_{i=1}^{L} \mathcal{R}_i\right] = E[L]r \geq \frac{\mathcal{E}r}{c}. \tag{20}$$

Consider the Markov chain which tracks the protocol evolution, $I(n)$ (n denotes here the number of actual transitions, not renewals). This chain visits state l at renewal instants. Define ν_{il} as the first passage time from state i to state l, i.e.,

$$\nu_{i\ell} = \min\{n > 0 : I(n) = \ell | I(0) = i\}. \tag{21}$$

Also, let F_{il} and $\overline{\mathcal{F}}_{il}$ be the reward earned during ν_{il} and its expected value, respectively, i.e.,

$$\{\mathcal{F}_{i\ell} = \alpha\} = \{\mathcal{R}(n) - \mathcal{R}(0) = \alpha | I(0) = i, \nu_{i\ell} = n\} \tag{22}$$

$$\overline{\mathcal{F}}_{i\ell} = \sum_{\alpha=0}^{\infty} \alpha P[\mathcal{F}_{i\ell} = \alpha]. \tag{23}$$

Note that, if we interpret the reward as time, F_{il} can itself be seen as a "first passage time" from state i to l. The statistics of F_{il}, and in particular $\overline{\mathcal{F}}_{il}$, can be readily computed from [9, Ch. 10].

Let γ_i be the probability that the chain is in state i when the total energy consumed reaches ε. Then, the additional reward earned until the next renewal is given by F_{il}, so that

$$\mathcal{R}(L) = \mathcal{R} + \mathcal{F}_{i\ell}, \tag{24}$$

and

$$\begin{aligned} E[\mathcal{R}] &= E[\mathcal{R}(L)] - E[\mathcal{F}_{i\ell}] \geq E[L]r - \sum_{i\in\Omega} \gamma_i \overline{\mathcal{F}}_{i\ell} \\ &\geq E[L]r - \sum_{i\in\Omega} \gamma_i \max_i \overline{\mathcal{F}}_{i\ell} = E[L]r - \max_i \overline{\mathcal{F}}_{i\ell}. \end{aligned} \tag{25}$$

Since the renewal state l is chosen arbitrarily, (25) holds for any l, and the tightest bound is given by

$$E[\mathcal{R}] \geq E[L]r - \min_{l} \max_{i} \overline{\mathcal{F}}_{il}. \tag{26}$$

Another approach to the computation of the rewards earned can be developed by considering an auxiliary renewal process whose interrenewals are the random variables R_i. Inspecting this process at the time that the battery is discharged, one finds the atypical interrenewal interval R_L whose length can be deduced from the age and residual life results of renewal theory [7]. Assuming steady state operation allows us to use the asympototically true analytical representation of the mean residual life as an approximation. This approach results in the estimate

$$E[\mathcal{R}] \simeq E[L]r - \frac{E[\mathcal{R}_1^2]}{2r}. \tag{27}$$

Figure 1 shows the results obtained for the energy efficiency using the various computational techniques illistrated so far, i.e., the recursive technique of Section 4, the asymptotic result (15), the lower bound (26) and the approximation (27). The protocol considered is ARQ GBN, whose description in terms of Markov structure and reward/energy parameters is given in [1] (the round-trip delay is 3 slots in this example). Different sets of curves correspond to different noise levels, with F the corresponding fading margin. It can be seen that the asymptotic result is an excellent approximation of the exact value obtained via the recursive technique. The bound (26) and the approximation (27) also converge very quickly to the exact value.

6 GENERAL PROFILE

The results of the previous section were derived assuming a flat profile, where all the transition metrics and probabilities were independent of the energy state. This assumption, although useful for analytical purposes, may not always be true in practice. In fact it is known that in general the constant power discharge regime is less efficient than constant load or constant current, and degrades cell capacity [5, 6]. In this situation, it is not unreasonable to give up on the perfect power regulation (flat profile), in favor of an increased battery efficiency.

Therefore the flat power profile, besides being an idealization of what a power regulator can actually do, may not correspond to the way in which the battery is operated anyway, and consideration of more general profiles is of great practical relevance. Nevertheless, the asymptotic results of the previous section can be used to obtain an accurate approximation when the power profile varies slowly as the battery energy decreases.

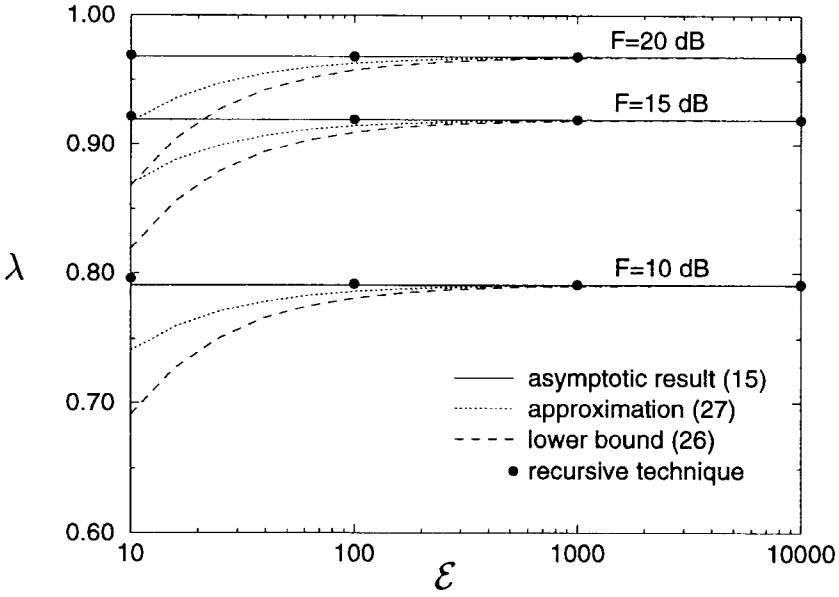

Figure 1: Energy efficiency, λ, vs. the energy allocation, ε, for ARQ Go-Back-N with round-trip delay 3 slots. Comparison of asymptotic result (15), approximation (27), lower bound (26) and recursive technique. Various values of the fading margin.

Let us consider the complete flow graph for $X(n)$, composed of stages 0, . . . , ε. Let us assume that stage 0 (corresponding to no energy consumed) is at the left end of the graph and that stage ε (all available energy consumed) is at the right end. Let us divide the flow graph into a number of consecutive segments in such a way that the transition parameters within each one of them exhibit negligible variations. Let the sequence $\{\tau_k, k = 0, 1, \ldots, T\}$ be chosen so that $\tau_k < \tau_{k+1}$, with $\tau_0 = 0$ and $\tau_T = \varepsilon$. Also, let the transition metrics in $[\tau_k, \tau_{k+1})$ exhibit negligible variation, so that

$$P_{ij}(\tau) \simeq P_{ij}(\tau_k), \qquad C_{ij}(\tau) \simeq C_{ij}(\tau_k), \qquad \tau \in [\tau_k, \tau_{k+1}). \tag{28}$$

A profile will be said *smooth* if the number of stages contained in each segment is sufficiently large as to allow application of the asymptotic result, e.g., if $\tau_{k+1} - \tau_k \gg C_{\max}$, with $C_{\max}$ the maximum amount of energy consumed during a transition. The bound given in (26) holds for any finite value of ε, and therefore can be used to bound the total reward in this case as well.

Note that the process will largely visit the segments in order, from left to right when the profile is smooth. Assuming that this is always the case makes it possible to deduce

the total reward earned in a simple way. Let i_k be the protocol state when the process $X(n)$ enters the k-th. segment for the first time. Then, the total reward can be written as

$$\mathcal{R} = \sum_{k=0}^{T-1} \Delta\mathcal{R}_k, \tag{29}$$

where ΔR_k is the reward earned while the process was flowing through the k-th segment of the graph, and depends in general on both i_k and τ_k. Under the assumption of smooth profile, we can assume that in each segment the steady-state is reached, so that

$$\Delta\mathcal{R}_k \simeq \lambda(\tau_k)(\tau_{k+1} - \tau_k), \tag{30}$$

and the dependence on the initial protocol state i_k vanishes. We can then rewrite (29) as

$$\mathcal{R} = \sum_{k=0}^{T-1} \lambda(\tau_k)(\tau_{k+1} - \tau_k) \simeq \int_0^{\mathcal{E}} \lambda(\tau) d\tau, \tag{31}$$

which makes it possible to compute the total reward by integrating the energy efficiency curve with respect to the energy state of the battery. This curve can be either computed via Markov analysis (as was done in [4]) or directly measured in an experiment.

6.1 Non-smooth power profiles

There are situations where the smooth profile model is not applicable. As an example, consider the results in [5], where pulsed discharge of a battery is considered. It is shown that as soon as current is drained from the battery, the voltage exhibits a sudden drop and then decreases more gracefully. Similarly, during the times in which current is not drained, the voltage recovers up to some level, due to chemical relaxation processes.

Therefore, a more accurate model at a finer time scale could take into account this "charge recovery" mechanism. In this latter case, it is clear that the evolution of the battery state $C(n)$ is not graceful and smooth, but can exhibit rather large fluctuations (up to about 33% of the maximum voltage in the example reported in [5, Fig. 15]). Therefore, the evolution of $X(n)$ through the flow graph, although showing a slow general trend from left to right, will fluctuate back and forth throughout the graph itself.

Such a behavior clearly violates the assumption of smooth profile but the recursive approach, which does not rely on any major assumption, will provide the correct solution in this case as well.

7 CONCLUSIONS

In this paper, we studied and compared a collection of results pertaining to the energy efficiency of link error control protocols. The most comprehensive of them is the recursive approach of Section 4. This approach can track arbitrary power profiles but can be computationally burdensome. At the other extreme is the asymptotic techique which predicts the efficiency accurately in the limit as $\varepsilon \to \infty$. For the particular examples considered the two approaches were found to agree very closely. Furthermore, the recursive technique is computational and does not result in analytical expressions. In contrast, the lower bound and approximations of Sections 4 and 5 lead to analytical expressions. The quality of the bound and the accuracy of the approximation improves as $\varepsilon \to \infty$. In other examples perhaps the accuracy and quality of the bounds will be different but our choice of the renewal instant appears to recur frequently enough to render the asymptotic results first presented in [4] quite accurate.

As a final remark, we note that the unavailability of data regarding the battery behavior under dynamic discharge conditions, especially at the time scales involved in packet data transmission (miliseconds), makes it difficult to accurately assess the effect of the battery behavior on the energy efficiency. It seems that such studies would be of great help to the protocol designer, and should be performed and made available. The techniques presented here do not depend on such data, and can be applied using results computed by analysis or generated by simulation, as well as experimental measurements.

References

[l] M. Zorzi, R.R. Rao, "Error Control and Energy Consumption in Communications for Nomadic Computing," *IEEE Trans. Computers* (special issue on Mobile Computing), vol. 46, pp. 279–289, Mar. 1997.

[2] N. Bambos, J.M. Rulnick, "Mobile power management for maximum battery life in wireless communication networks," in *Proc. IEEE INFOCOM'96*, pp. 443–50, Mar. 1996.

[3] N. Bambos, J.M. Rulnick, "Performance evaluation of power-managed mobile communication devices," in *Proc. IEEE ICC'96*, pp. 1477–81, Mar. 1996.

[4] M. Zorzi, R.R. Rao, "Energy Constrained Error Control for Wireless Channels," in *Proc. IEEE GLOBECOM'96*, Nov. 1996.

[5] E.J. Podlaha, H.Y. Cheh, "Modeling of cylindrical alkaline cells. VI: variable discharge conditions," *J. Electrochem. Soc.*, vol. 141, pp. 28-35, Jan. 1994.

[6] D. Linden, ed., *Handbook of Batteries*, 2nd edition, New York: McGraw-Hill, 1995.

[7] S.H. Ross, *Stochastic processes*, John Wiley & Sons, 1983.

[8] A. Gut, *Stopped Random Walks: limit theorems and applications*, New York: Springer-Verlag, 1988.

[9] R.A. Howard, *Dynamic probabilistic systems*, New York: John Wiley & Sons, 1971.

AN ACCESS SCHEME FOR HIGH SPEED PACKET DATA SERVICE ON IS-95 BASED CDMA

Sarath Kumar and Sanjiv Nanda
Bell Labs, Lucent Technologies, Holmdel NJ 07733-3030, USA
{sarath,nanda}@lucent.com

Abstract:

In this work, we study a simple admission control scheme for high speed packet data service over the reverse link of an IS-95 based CDMA air interface. A preliminary version of this scheme was originally proposed in [3] and admits data users on a burst level. In this work, we study the performance and design tradeoffs through simulations. The basic tradeoff is the coverage of high speed packet data service and its impact on outage experienced by the voice users. The results indicate that through limiting the region where high speed data service is available, the performance impact on voice users can be made small. It is possible to extend the region of coverage of high speed data by using more complex schemes proposed in [3]. These schemes are subject of future work.

1 Introduction

IS-95 based systems operate in circuit mode, assume a homogeneous user population, and limit each user to a rate which is a very small fraction of the system capacity. Users at higher data rates in a system with mixed traffic cause adjacent cell interference variations which can degrade system capacity. High speed data needs strict admission controls as compared to the voice calls because of the interference generated by these users at adjacent cells. The idea behind this scheme is to control the tails of the out-of-cell interference caused by the data user through burst admission. In this scheme, a data user burst is admitted if the interference caused at the adjacent cells is below a predetermined threshold. This can be estimated indirectly through the pilot strength measurements at the mobile.

Consider the reverse link of a system in which distributed high data rate users in a cellular environment share the cellular CDMA band with conventional

mobile cellular voice and circuit-mode data users. Burst admission schemes can be designed using:

1. The load information in the cell and its neighbors,
2. The pilot strength measurements provided by the mobile, and
3. Coordination of the burst rate, burst length and burst starting time between neighbor cells.

In this work, we study the utilization of the radio resources assuming that network resources are always available at the sectors with which the mobile is in soft-handoff. Our strategy is to assume that the interference caused by the mobile at adjacent sectors is acceptable if the corresponding pilot is below a threshold parameterized by T_{burst}. We study the interference and outage probability for this strategy, as a function of T_{burst}. The rest of the document is organized as follows. In Section 2, we describe the burst access algorithm. This algorithm is studied through detailed simulations in the sequel. Section 3 describes the simulation environment used, followed by a discussion of the bandwidth and coverage tradeoffs in Sections 4 and 5. Finally, we conclude in Section 6.

2 Burst Access Mechanism

2.1 Quiescent Mode

As shown in Figure 1, an active high rate mobile is assigned a basic "full" rate channel on origination. Parameters of the high data rate service are negotiated at that point. The mobile then goes into a quiescent mode if it has no data to transmit. When a user is quiescent, a very low rate (say) eighth rate (sub-rate) signaling channel is maintained using its primary code. This sub-rate channel helps in maintaining synchronization and coarse power control. It is maintained whether the user is "connected" to one base station or is in soft handoff with multiple cells. This basic channel may be considered a signaling channel for high data rate service.

2.2 Burst Access Request

Since the transmission during eighth rate frames is intermittent, both the synchronization and the power control are inadequate if the quiescent period is long. Hence, any transmission from the mobile after a long quiescent period may be lost. This can be easily fixed by requiring the mobile to transmit one (or more) idle basic rate frame at the end of a "long" quiescent period. Following the idle frame(s) that give the receiver time to synchronize and provide power control feedback, the mobile signals a request for data transmission using signaling messages over the basic (full) rate channel. Alternately, instead of the idle frames, the mobile could be required to transmit the request multiple times.

The access request from the mobile contains the data rate requested and the burst length requested. The maximum burst length that may be requested is

Flow Diagram: Service Option Z

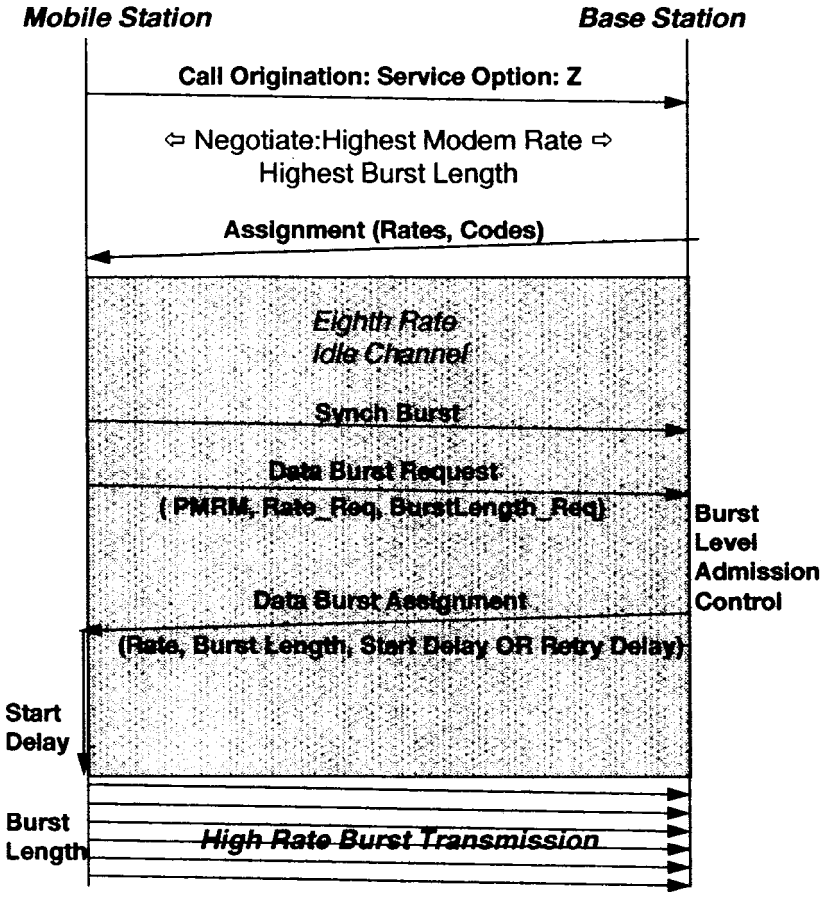

Figure 1: Burst Access Mechanism

specified by the system (and is chosen to best coordinate shared access between users). In addition, to provide interference information to the base station, the mobile includes *pilot strength information for cells in its neighbor list within the access request* [1]. This indicates the interference levels that will be seen at the neighboring basestations due to transmissions from the mobile.

2.3 Access Control Mechanisms

Once the burst request is received at the base station, the network must coordinate access with other requests as well as with the voice load already offered to the system. Based on all information available, the network makes a decision to either accept, delay or deny the request. It then creates the assignment message and transmits it to the mobile.

The following section describes the simulation environment used to study one such access control mechanism.

3 Simulation Description

The access schemes use load information about neighbor cells to determine what data rate is to be permitted. Simple thresholds on load can be compared with reported pilot measurements to determine the permitted data rate for a burst. Alternatively, the voice load can be reduced to design an access control that shares bandwidth in a more static manner.

Here we study the effectiveness of such a scheme through simulations. In particular, we study the performance of this scheme as a function of thresholds used for admission control and the length of the burst admitted.

The simulation environment consists of a hexagonal grid with base stations located at the center of every hexagon. A *single* mobile's operation is simulated in detail, accounting for the following system features:

1. realistic antenna patterns for three-sector cells,
2. IS-95 handoff algorithm details,
3. handoff processing and messaging delays.

Details of the simulation environment follow.

Propagation Loss The propagation model has two components: distance loss and shadow fading. The distance loss is taken as $d^{-\alpha}$, where α is the propagation constant. Hence the attenuation in dB is given as

$$a(d) = K_1 + 10\,\alpha \log(d) + u(d) \tag{1}$$

where the first two components denote the distance loss, K_1 being the path loss at unit distance and $u(d)$ is the loss due to shadowing.

[1] The inclusion of the pilot strength measurements within the access request is independent of (and in addition to) any such reports used for handling soft handoffs. We assume that the soft handoff procedures defined in IS95 are left unchanged.

Fading Model We assume suffficient averaging of the signal strength measurements to smooth out Fast fading. For slow or shadow fading, two models are simulated: the first of them is lognormal with exponential correlation function as suggested in [4]. That is, $u(d)$ is zero mean stationary Gaussian process with correlation

$$E\{u(d_1), u(d_2)\} = \sigma_s^2 \exp(-|d_1 - d_2|/d_0) \tag{2}$$

where σ_s is the standard deviation of shadow fading and d_0 determines how fast the correlation decays with distance. For a vehicle moving at speed v meters/second, the spatial correlation translates into time correlation so that

$$E\{u(t_1), u(t_2)\} = \sigma_s^2 \exp(-|v(t_1 - t_2)|/d_0) \tag{3}$$

The second model used for Shadow fading is the Mawira model [5], which simulates urban environment. It is the sum of two independent lognormal components v and w, each with correlation function given in (3), but with different correlation distance and variance. The first component, v has large variance and short correlation distance, and second component a small variance with large correlation length. Thus, according to this model, the shadow fading in dB domain is given by

$$u(d) = v(d) + w(d) \tag{4}$$

where $v(d)$ and $w(d)$ are zero mean Gaussian processes.

Mobility Model The mobile chooses a random starting location and direction of motion at the beginning of the call. It traverses in a straight line until the call is completed. Starting location is uniformly distributed over the entire area of the cell; direction of motion is uniformly distributed in $[0, 2\pi]$. Call holding time and the speed of the mobile are taken to be constant.

Power Control Since our interest is in studying the coverage aspects, we focus on the pilot which is transmitted at a fixed power level. We assume that the reverse link power control completely compensates for the distance loss and shadow fading. The interference generated at the adjacent cells is determined by the best active pilot. The mobile is assumed to transmit at a power level that is needed so that an adequate signal level is received at this cell. The other cells see interference corresponding to this transmit power level.

Handoff Algorithm The mobile and the network follow the IS-95 handoff algorithms. During the length of the call, the mobile keeps monitoring the pilot signal to interference ratio or E_c/I_0, where E_c is the chip energy and I_0 is the total received power spectral density. Handoff decisions are made at $\{k\,T_m\}$ where k is an integer and T_m is the pilot strength measurement interval.

If, after going through the handoff evaluation sequence, the mobile finds all its active pilots dropped, the call is assumed dropped due to pilot loss.

Burst Access In the burst access mode, the data mobile makes a request to send a burst every R_{burst} seconds. The mobile request is admitted after a delay of D_{burst} seconds if the E_c/N_o of the largest non-active pilot is below a threshold T_{burst} and the mobile is not in the handoff processing mode, that is, in the process of adding or dropping a leg. Otherwise, the burst admission is denied. A burst transmission is interrupted when a soft handoff add trigger comes in.

The algorithm is studied as a function of T_{burst}. The metrics that are of importance in studying the performance of the algorithm are: fraction of the burst requests that are admitted, average burst length that determines the througput and the distribution of interference caused at the adjacent sectors.

4 Bandwidth Tradeoffs

The access schemes provide a simple mechanism to tradeoff bandwidth between circuit-mode voice users and packet mode access by high rate data users. Thus a single access scheme permits dynamic sharing of bandwidth between voice and data users. The bandwidth that a data user can be permitted to transmit depends on (i) the number of voice users already admitted in the cell and its neighbors, and (ii) the position of the high rate data user requesting burst access.

This simulation can be used to study the effectiveness of the burst admission strategies in terms of the fraction of burst requests admitted, the burst throughput achieved and the interference caused at the adjacent sectors. The probability density function (pdf) of the interference generated at each cells by this single mobile is collected from the simulation. These pdfs are computed for each of the following scenarios:

1. a full rate circuit-mode mobile, $p_{c,j}$, where j denotes the cell index,
2. a voice mobile, $p_{v,j}$,
3. a mobile following the burst admission strategy described here, $p_{d,j}$.

We now use symmetry to obtain the distribution of *total signal power generated* at base station 0 when there is *one mobile per cell.* For example, the distribution of received power due to one voice mobile in each of the cells is given

$$p_v = p_{v,1} \otimes p_{v,2} \otimes \cdots \otimes p_{v,M} \tag{5}$$

where $\otimes$ denotes convolution and M is the number of adjacent cells which contribute reverse link interference.

These pdfs are then further convolved to obtain the distribution of interference due to a given mix of voice and data users. That is, if there are N_v voice users and N_d packet-mode data users, then the resulting interference distribution on the uplink is obtained as

$$\begin{aligned} p(x/N_v, N_d) = (p_v(x) \otimes p_v(x) \otimes \cdots p_v(x))_{N_v} \otimes \\ (p_d(x) \otimes p_d(x) \otimes \cdots p_d(x))_{N_d} \end{aligned} \tag{6}$$

The expression for outage, defined as the probability of the Signal to Interference Ratio falling below a threshold I_T can be computed from (6) as

$$P_{out}(N_v, N_d) = \int_0^{I_T} p(x/N_v, N_d)\, dx \tag{7}$$

I_T is assumed to be the same for both data and voice users and is computed using the methodology described in [1].

The parameter values used in obtaining P_{out} are: $T_{Add} = -12dB$, $T_{Drop} = -16dB$, $T_{T_{drop}} = 3$ sec. and path loss exponent is 4. Shadow fading is simulated using the Mawira model with standard deviations: $5.4dB$ and $4.2dB$ and correlation lengths $100m$ and $1000m$ respectively. In Figs. 2 and 3, the mobile speed is assumed to be 35 miles per hour.

Fig. 2 shows the outage of the system as a function of the number of Rate Set 1 voice users. The voice activity factor is taken to be 3/8. In idle mode, we assume that they are transmitting eighth rate frames. It can be seen from this plot that the capacity of the system is 22 voice users per cell at 5% outage and 18.5 voice users at 1% outage. The outage is computed assuming an Eb/No requirement of 7 dB. Within the approximations of the simulation, these results are reasonable.

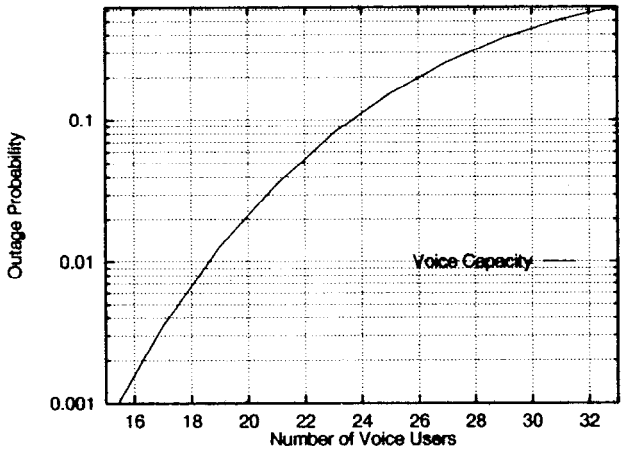

Figure 2: Outage of the system with only voice users

In Figs. 3 and 4, we plot the outage experienced when there is one data user per sector operating in either circuit or packet mode at 57.6 kbps (4-codes aggregated, each at 14.4 kbps). The circuit mode data user is active with activity factor 1. We also assume that there is one packet-mode data user active at any given time. The access mechanism could be designed to ensure this. The packet mode data user follows the burst access scheme described in Section 3, with D_{burst} = 1sec. Fig. 3 is parameterized with the burst threshold T_{burst}.

Some salients points from simulation:

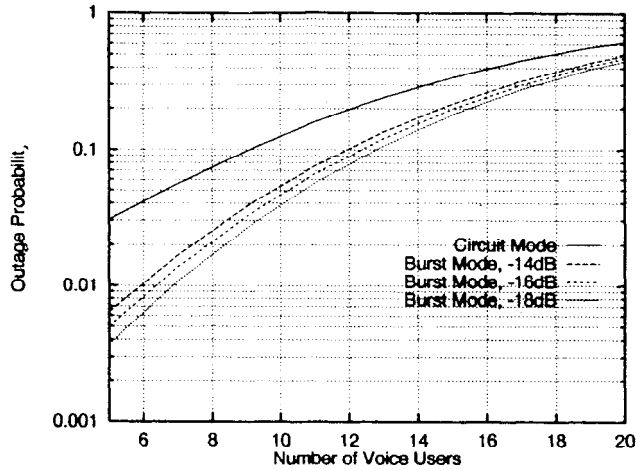

Figure 3: Outage with one high speed data channel at 57.6 kbps, and activity factor=1: Impact of threshold on voice capacity

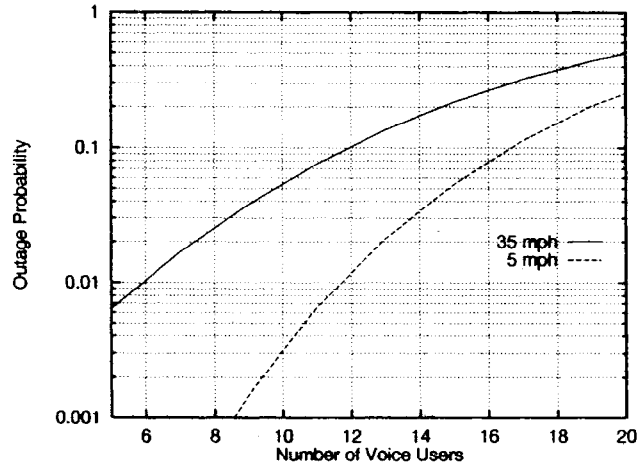

Figure 4: Outage with one high speed data channel at 57.6 kbps, and activity factor=1: Impact of mobile speed on voice capacity

- From Fig. 3, we observe that when the data user is operating in circuit mode, the system can support 7 voice users at 5% outage. Since the 6 codes corresponding to the data user are active all the time, the data throughput on each of those codes is approximately 1/voice_activity or 2.2 times higher than that of a voice user. This gives an equivalent total throughput of $7 + 2.2 \times 6 = 20.2$ voice users when there is one high-rate circuit mode data user active. It translates to a loss of 8% in capacity.
- In packet mode data, the burst admission control strategy ensures that the data users that cause excessive interference at the neighbors are not admitted. This results in higher capacities than is the case with circuit mode data. For the range of parameters considered in Fig. 3, we see that the capacity penalty can be completely eliminated. The penalty paid for this improved capacity is the lower coverage for the high speed data service described in Section 5.
- Fig. 4 shows the capacity plots as a function of the mobile speed. Assuming an Eb/No requirement of 5 dB for mobiles at 5 mph results in 45% higher capacity for the voice users.

5 Coverage for High Speed Data

We use static simulations to obtain the coverage plots for the high speed data. These are obtained using the following steps:

1. At each point, generate shadow fading to all the adjacent sectors and also compute the corresponding path loss. Use this to estimate the Ec/Io corresponding to all the pilots. The mobile is assumed to be in soft-handoff with all the base stations whose pilot strengths exceed the threshold T_{Add}.
2. If the best non-active pilot strength falls below T_{burst}, burst is admitted.
3. Repeat steps 1–2 above N_{iter} times keeping track of the fraction of iterations over which the burst is admitted.

Fig. 5 shows the the burst admission probability obtained in step 3 above for $T_{burst} = -14dB$ as a function of normanized distance from the cell site. These simulations assume that the mobile does not send a burst admission request when the mobile is in the handoff processing mode. We make the following observations for the coverage of the high speed data users:

- Burst admission probability gets smaller as we approach the cell boundary.
- if $T_{burst} = T_{Add}$, the burst request is never denied. The penalty paid here is the outage experienced by the voice users. This can be traded off by taking a capacity hit on the voice users. The tradeoffs are quantified in Table 1.
- For a given burst length, lowering T_{burst} reduces the outage experienced by the voice users. But this also limits the region over which high speed data service is availablie. For instance, with $R_{burst} - D_{burst} = 3$ sec.,

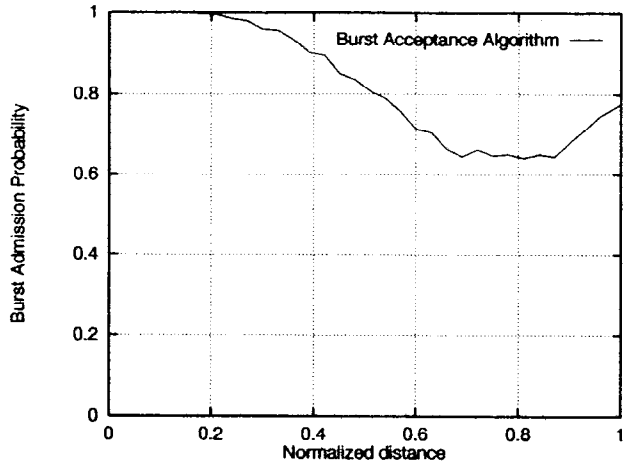

Figure 5: Admission Probabilities (Reverse Link)

the area over which the service is available changes from 68% to 35% on reducing T_{burst} from $-14dB$ to $-18dB$ (Table 1).

Threshold (dB)	Admission Prob.
-14.0	0.68
-16.0	0.50
-18.0	0.35

Table 1: Performance trade-offs for different threshold values

6 Conclusions

A simple burst admission control scheme that limits access based on the pilot strength measurements from the mobile is studied. The burst allocation is done for a fixed maximum duration. The mobile terminates the burst transmission when a handoff add trigger is generated.

The simulation results indicate that high speed data service can be provided over 65% of the area through a simple scheme that only looks at the local measurements made by the mobile. Larger burst lengths lead to higher probability of being interrupted, which may mean that the mobile will retry for the high speed service once the handoff processing is completed. Even better coverage over area can be achieved through using higher values for T_{burst} and taking a hit on the voice capacity.

Note that all the discussion in this paper corresponds to the coverage for *high-speed* data. Better coverage can be provided by trading off data rates for

outage.

References

[1] K. S. Gilhousen *et al*, "On the Capacity of a Cellular CDMA System," *IEEE Trans. Veh. Technol.*, Vol. VT-40, No. 2, pp. 303-312, May 1991.

[2] N. D. Wilson, R. Ganesh, K. Joseph, and D. Raychaudhuri, "Packet CDMA Versus Dynamic TDMA for Multiple Access in an Integrated Voice/Data PCN," *IEEE J. Selected Area in Communications*, Vol. JSAC-11, No. 6, pp. 870-884, August 1993.

[3] C-L I and S. Nanda, "Load and Interference Based Demand Assignment for Wireless CDMA Networks," in *Proc. IEEE Globecom 96.*

[4] M. Gudmundson, "Correlation Model for Shadow Fading in Mobile Radio Systems," *Electronics Letters*, vol. 27, No. 23, 1991.

[5] A. Mawira, "Models for the spatial correlation functions of the (Log)-Normal component of the variability of VHF/UHF field strength in Urban environment", *3rd IEEE International Conference on Personal, Indoor and Mobile Radio Communications*, pp. 436-440, 1992.

This page intentionally left blank.

CAPACITY WHEN USING DIVERSITY AT TRANSMIT AND RECEIVE SITES AND THE RAYLEIGH-FADED MATRIX CHANNEL IS UNKNOWN AT THE TRANSMITTER

G. J. Foschini and M. J. Gans
Lucent Technologies, Bell Laboratories Innovations
Crawford Hill Laboratory
Holmdel, N. J. 07733-0400

ABSTRACT

We report on theoretical studies for higher bit-rates for burst mode communication in wireless LANs and wireless interbuilding links. We discuss an information theoretic investigation of the benefits of multi-element arrays (MEAs) when communicating in a narrowband Rayleigh fading environment. While the channel is unknown at the transmitter, it is assumed known (tracked) at the receiver. Say that the same number of antennas, n, is used at both sites and that beside bandwidth, *total* radiated power is fixed. We show that for large n capacity increases linearly with n. The signal components launched from distinct transmit elements interfere at the receiver but in the large n, large ρ, realm, capacity is the same as if there were no interference but SNR was reduced by $10 \cdot \log_{10} e \approx 4.343$ dB ($e \approx 2.7183...$). The derivations hint at the existence of an architecture with extraordinary capacity that can be approached using one dimensional codecs as building blocks. We exhibit this means of communication, which is termed the *layered space-time architecture* because of the way signals are stratified along diagonals in space-time.

1. INTRODUCTION

We report on theoretical studies of higher bit-rates for burst mode communication in wireless LANs and wireless interbuilding links. We present an information theoretic investigation of the benefits of multi-element arrays (MEAs) for communicating in a narrowband Rayleigh fading environment. While the channel is assumed unknown

at the transmitter, it is known (tracked) at the receiver. We also discuss an architecture aimed at delivering a significant fraction of the great capacities uncovered

Assume that the volumes to be occupied by the transmitter and receiver are very far apart on the scale of a wavelength. Also assume that each of the volumes is ample on the scale of a wavelength (at least several wavelengths in the largest dimension). Fix the following:

- Carrier (.8 to 60 GHZ)
- Total radiated power
- (n_T, n_R) = (no. transmit antennas, no. receive antennas)

We address the questions: *How many error free bits/sec/Hz can be delivered from the space the transmitter occupies to the distant space that the receiver occupies — no limit on processing complexity? Then, accounting for complexity: For a specified low BER, how can the ultimate bit rate be approached?* We will discuss some recent progress toward answering these questions.

The scope here is limited to a single point-to-point channel. We take the perspective of discrete time complex baseband involving a fixed linear matrix channel with additive white gaussian noise (AWGN). Although fixed, the channel can vary randomly from burst to burst. We assume that the burst is long enough that the standard information theoretic infinite time horizon view provides a meaningful idealization. First we need to specify the framework of our analysis.

2. MATHEMATICAL MODEL FOR THE WIRELESS CHANNEL

Following [1-2] we list some basic assumptions. We use m–D in describing a vector as m dimensional, meaning that it has m *complex* components (2m components totaling real plus imaginary).

- TRANSMITTED SIGNAL s(t): The total power is constrained to $\hat{P}$ regardless of the value of n_T (the dimension of s(t)). The bandwidth is narrow enough that we can treat the channel frequency characteristic as flat over frequency.
- NOISE AT RECEIVER $v(t)$: complex n_R – D AWGN with statistically independent components of identical power N at each of the n_R receiver branches.
- RECEIVED SIGNAL r(t): n_R–D received signal so that at each point in time there is one complex vector component per receive antenna. When there is only one transmit antenna, it radiates power $\hat{P}$ and we denote the average power at the output of each of the receiver branches by P.
- AVERAGE SNR AT EACH RECEIVER BRANCH: $\rho = P/N$ independent of n_T.
- MATRIX CHANNEL IMPULSE RESPONSE: g(t) has n_T columns and n_R rows. We use G(f) for the Fourier transform of g(t). Consistent with the narrowband assumption, we treat this (matrix) transform as constant over the band of interest, writing G, suppressing the frequency dependence. So, except for g(O), g(t) is the zero matrix. As we will see in the following sections, it will

be convenient to represent the matrix channel response in normalized form, h(t). Specifically, related to G, we have the matrix H, where the equation $\hat{P}^{½} \cdot G = P^{½} \cdot H$ defines the relationship so, $g(t) = (P/\hat{P})^{½} \cdot h(t)$.

The following standard notation will be needed: ' for vector transpose, † for transpose conjugate, I_n for the n×n identity matrix, E{.} for expectation and * for convolution.

The equation for communicating over the channel is

$$r(t) = g(t) * s(t) + v(t). \tag{1}$$

The added vectors are complex n_R–D vectors. Using the narrowband assumption, we simplify, replacing convolution by product and write

$$r(t) = (P/(\hat{P} \cdot n_T))^{½} \cdot h(0) \cdot s(t) + v(t). \tag{2}$$

3. GENERALIZED CAPACITY FORMULA AND SOME EXAMPLES

It is essential to have the n_T transmit elements convey different signal constituents. Starting from fundamentals in [3], the following generalized formula for capacity, C, can be derived [1]:

$$C = \log_2 \det [I_{n_R} + (\rho/n_T) \cdot HH^{\dagger}] \text{ bps/Hz}. \tag{3}$$

Here H is an $n_T \times n_R$ matrix, ρ is the spatial average SNR at each of the received array elements and det means determinant.

We will often deal with H as a random matrix. Our idealized representation of Rayleigh fading has it that the entries of H are complex iid Gaussians of zero mean and unit variance. The real and imaginary part of each entry are independent of each other and of variance ½. As mentioned in Section 1, the view in many of our applications is that the channel, though random, is fixed during a communication burst. Then we take a "quasi-static" perspective treating capacity as a random variable.

Since the available surface areas of the transmit and receive regions can be paved with a lattice (spacing of the order of λ) of approximately independent omni-antenna elements there is an opportunity for an enormous number of useful elements [4-5]. Eg, at 7.5 GHz λ is only ≈ 4 cm.

Modifying the array location causes the channel to change. Later we will give some plots of the complementary cumulative distribution functions (CCDFs) of capacity. Typically we are interested the the small outage tail and will hold the Probability [outage] to under 1 to 2% or so.

It is important to note that, remarkably, it is possible to code over an ensemble of channels even though the channel is unknown to the transmitter [6-7].

Next, for reference, we use (3) to list some capacity formulas for the simplest cases, namely when $n_T = 1$ and $n_R = n$. These simple architectures are well known architectures. First we give the capacity of what is referred to as the maximum ratio combining structure. Since the n received signals are optimally combined we will abbreviate this case as OC(n) This optimum receive diversity formula is:

$$C = \log_2[1 + \rho \cdot \sum_{i=0}^{n_R} |H_i|^2]. \tag{4}$$

For the Rayleigh environment this formula becomes

$$C = \log_2[1 + \rho \cdot \chi^2_{2n}]. \tag{4a}$$

For expressiveness in the presentation of this formula (and in several other formulas in this paper) we take a notational liberty writing χ^2_{2n} to *directly* convey a chi-squared variate with 2n degrees of freedom (DOF). Since we have assumed that the underlying real and imaginary Gaussian variates have variance ½ it is also nonstandard that $E\chi^2_{2n} = n$ instead of the standard expectation of 2n for χ^2_{2n}.

Selection diversity is a form of diversity that is often contrasted with OC(n). Selection diversity capacity is the capacity offered by the best of the nR channels. In the Rayleigh case we have:

$$C = \log_2[1 + \rho \cdot \max\{n \text{ independent } \chi^2_2 s\}]. \tag{5}$$

We also include the (n, 1) case assuming Rayleigh fading. This transmit diversity capacity formula is

$$C = \log_2[1 + (\rho/n) \cdot \chi^2_{2n}]. \tag{6}$$

Note that for large n, that C converges in distribution to $\log_2(1 + \rho)$.

4. NUMERICAL EXAMPLES

Figures 1, 2 and 3 show some complementary cumulative distribution functions (ccdfs) of capacity.

Figure 1 below shows the case of (3, 3) for various SNRs. in increments of 3 dB. These are shown in bold. For contrast the corresponding (1, 1) cases are shown in regular line curves. Note that for 21 dB P_{out} = 1% the (3, 3) capacity exceeds about 13 bits/cycle while for n = 1 there is only about one bit/cycle. While 13 bits/cycle may seem very large it amounts to about 4.33 bits/cycle/dimension.

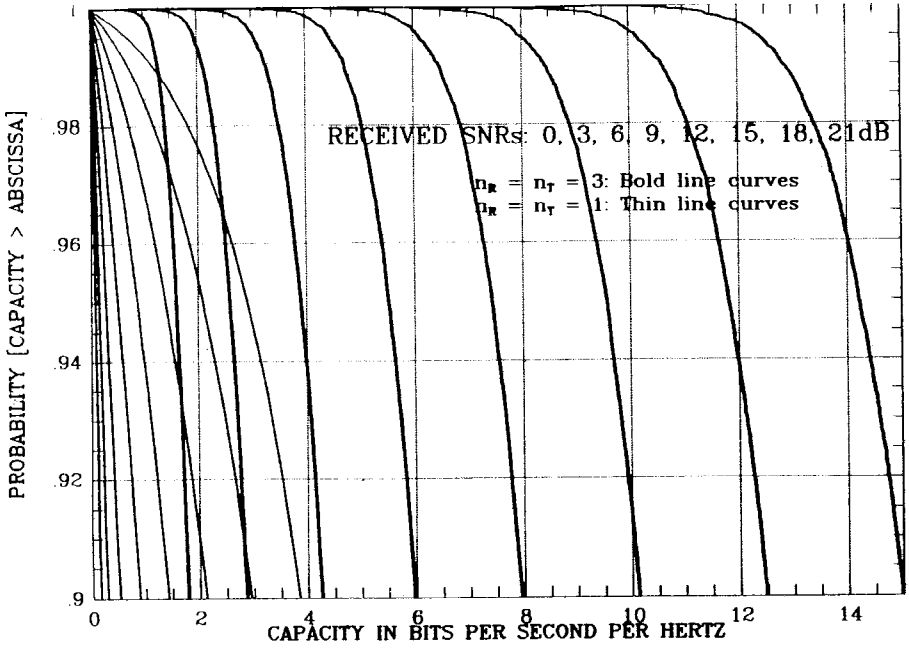

FIGURE 1: CAPACITY CCDFs TAILS FOR (1,1) AND (3.3) CASES

Figure 2 shows some comparisons with a (4, 4) system for an SNR of 21 dB. We see the (4, 4) capacity is significantly greater than the OC(4) capacity: at P_{out} = 1% we get 19 bits/cycle versus 6.7 bits/cycle. OC(n) has the implementation advantage of being based on 1–D codec technology. However, while as n increases, OC(n) is significantly better than a (1,1) system, capacity only scales logarithmically with n. In Figure 2 selection diversity is seen to perform somewhat worse than OC(4) as it must, and transmit diversity (4, 1) diversity is still worse.

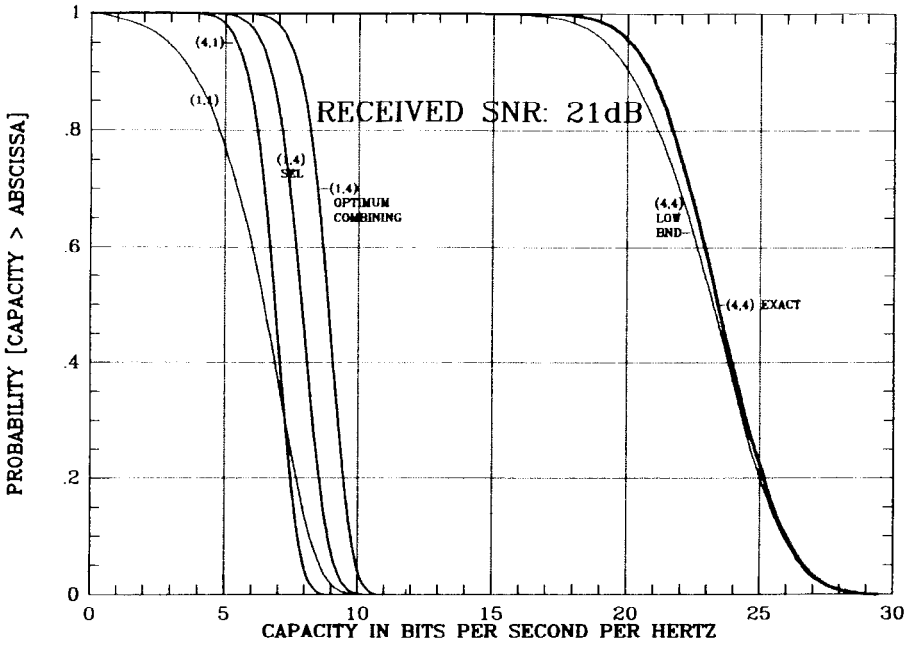

FIGURE 2: CAPACITY CCDFs: CONTRASTS WITH (4, 4) CASE

Figure 3 shows some comparisons for a (32, 32) system for an SNR of 21 dB. Now the departure in performance between OC(n) and (n, n) at say the 99%–tile is seen to be enormous. Eg, for (32, 32), at 1% outage and 21 dB average SNR at each of the receiving antenna elements, we get 182 bps/Hz which is over 16 times the capacity for OC(32). From a signal constellation standpoint it may seem unreasonable to strive for a significant fraction of such a huge capacity. Moreover, 182 bps/Hz is merely for n = 32, in some applications n could easily be larger. However, later we point out that (n, n) systems can be designed for which it is n lower component capacities, one for each of the n signaling dimensions, that is relevant from the standpoint of signal constellations.

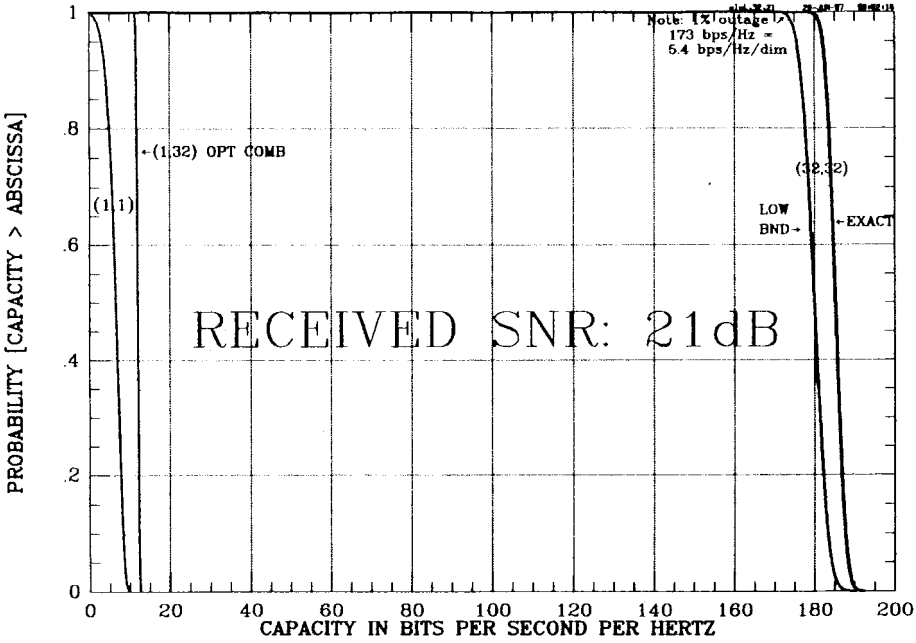

FIGURE 3: CAPACITY CCDFs, CONTRASTS RELATED TO (32.32) CASE

5. LOWER BOUND ON CAPACITY

Toward the goal of understanding how the great bit rates that are possible at low outages might be approached, an especially expressive lower bound on the capacity of combined transmit-receive diversity is very useful. In the case $n_R = n_T = n$, this bound, which holds with probability one is

$$C > \sum_{k=1}^{n} \log_2 [1 + (\rho/n)\cdot\chi^2_{2k}] \text{ bps/Hz.} \quad (7)$$

Above the χ^2_{2k} represent a finite sequence of independent chi-squared variates with 2k DOF and mean k.

5.1 Derivation of the Lower Bound on Capacity

A detailed derivation of the lower bound is now given. Use m for n_T and n for n_R. We say two that two m×n matrices A and B are unitarily equivalent [8] if there exists two matrices U and V such that A = UBV† We derive a lower bound on capacity using the following result on random matrices from [9-10]:

The matrix H is unitarily equivalent to the m by n matrix given below:

$$\begin{pmatrix} x_{2m} & & & & 0 & \cdots & 0 \\ y_{2(n-1)} & x_{2(m-1)} & & & \vdots & & \vdots \\ & \ddots & \ddots & & \vdots & & \vdots \\ & & y_2 & x_{2(m-(n-1))} & 0 & \cdots & 0 \end{pmatrix} \quad (8)$$

where x_j^2 and y_j^2 are distributed as χ^2 variables with j DOF (i.e. χ_j^2). The x and y elements here are all non-negative and independent. All matrix entries below the y stripe and to the right of the x stripe are understood to be zero.

We use the representation (8) to rewrite $I_{n_R} + (\rho/n_T)HH^\dagger$. First define $X_{2j} = (\rho/n_T)^{1/2} x_{2j}$ and $Y_{2j} = (\rho/n_T)^{1/2} y_{2j}$. Then $I_{n_R} + (\rho/n_T)HH^\dagger$ is the $n_R \times n_R$ matrix of the form

$$\begin{pmatrix} 1 + X_{2m}^2 & X_{2m}Y_{2(n-1)} & & & \\ Y_{2(n-1)}X_{2m} & 1 + X_{2(m-1)}^2 + Y_{2(n-1)}^2 & X_{2(m-1)}Y_{2(n-2)} & & \\ & Y_{2(n-2)}X_{2(m-1)} & 1 + X_{2(m-2)}^2 + Y_{2(n-2)}^2 & X_{2(m-2)}Y_{2(n-3)} & \\ & & & \ddots & \\ & & & Y_2X_{2(m-(n-2))} & 1 + X_{2(m-(n-1))}^2 + Y_2^2 \end{pmatrix}. \quad (9)$$

We have displayed all nonzero entries in the first three and last rows.

In what follows let DET = det $(I_{n_R} + (\rho/n_T)HH^\dagger)$. Expanding the product of the n main diagonal entries we see that we get L + Q where

$$L = \prod_{j=m-(n-1)}^{m} (1 + X_{2j}^2) \quad (10)$$

and Q is a positive random variable obtained by summing over only positive terms. The special symmetric tridiagonal structure exhibited in (9) will now enable us to establish L as the lower bond.

The standard definition of determinant involves a signed sum over n! n–fold products. Each n–fold product is the product of the n terms indicated by a permutation matrix. Let σ_k k = 1, 2, · · · n! denote these signed summands comprising DET each one of which is an n–fold product. We remind the reader that every permutation is a product of transpositions and the negatively signed contributions to a determinant are those that involve an odd number of transpositions [8]. We will look at the structure of each of the nondiagonal n–fold products.

Note that an above diagonal term, say the i (i + 1) entry, has the possibility of appearing as a multiplier in a nonzero σ_k only if its counterpart in the (i + 1)i position below the diagonal also appears as a multiplier. (Since these two entries are equal the i(i + 1) multiplier appears squared in σ_k.) Suppose the (i + 1)i entry was not also included. We need to include exactly one entry from each row and column. So the only other option among the (n–1)! terms that involve the i(i+1) entry that have at least n – 1 nonzero multiplicands is easily seen to be that σ_k forced to include the entire superdiagonal in the n–fold product. However, even that σ_k must be zero. The reason such a $\sigma_k = 0$ is that including the entire superdiagonal forces the inclusion of the lower left corner entry among the n multipliers. The reason that the lower left corner entry, the n1 entry zero, is forced to be a multiplicand is again that exactly one entry from each row and column is required to qualify a permutation matrix.

Now that we have detailed the form of the negative σ_k we can see why it is that their sum does not undermine the bound. Indeed, we shall see that each such negative σ_k is cancelled. To see the cancellation, look at the structure of the terms in the product of the ii and (i + 1)(i + 1) entries, specifically, look at the perfect square terms apart from those that give rise to L. It is immediate that each of the negative σ_k is cancelled by a distinct positive contribution to Q. Since Q contains more terms than those needed to cancel all the negative σ_k, we get that C > L with probability one. See [2] for a discussion of the large n, large ρ, asymptotic equivalence of C and L.

5.2 Capacity Lower Bound Analysis For A Large Number of Antennas

Using L to denote this lower bound, it easily follows that, in the limit of large n, L/n becomes

$$L(n)/n \rightarrow (1 + \rho^{-1}) \cdot \log_2(1+\rho) - \log_2 e \quad \text{as } n \rightarrow \infty. \tag{11}$$

To get this result simply notice that the left hand side of (11) is a discretization of the integral of $\log_2(1 + \rho x)$ with respect to x over [0,1]. In the limit of large ρ

$$L(n)/n \rightarrow \log_2 \rho/e \quad \text{as } n \rightarrow \infty. \tag{12}$$

We see that for this (n, n) case, despite the fact that the n received waves interfere randomly, capacity grows linearly with n. What is the cost of the interference between the waves from different transmitters? That is an easy question to answer. Just compare (12) with what one gets with n distinct OC(n) systems each transmitter transmitting power $\hat{P}/n$, but each having its own separate set of n receivers. In other words compare with the capacity, $\mathbf{C}(n)$, that would be if there were no interference. We need to sum over n instances of (4a), except that, because of the reduced transmit power, we need to replace ρ by ρ/n in (4a). In such a case the $\mathbf{C}(n)/n$ ratio goes to $\log_2(1 + \rho)$. So the cost of the interference is that ρ gets replaced by ρ/e in the capacity formula, a cost of about 4.343 dB in average SNR.

In most 1-D information theory analyses one learns *what* ultimate rates are possible but not *how* to obtain the rates. However, as regards the *spatial aspect* of communication system design, our multidimensional MEA analysis hints *how* to design an architecture approaching capacity. Specifically, interpreting the aforementioned lower bound, L, in terms of OC(k) systems (k = 1, 2, ... n) is very informative to the design process as we will now show.

6. THE LAYERED SPACE-TIME ARCHITECTURE

It is encouraging to know that the (n, n) system has capacity linear in n. A goal is to synthesize an (n, n) architecture of great capacity using 1–D codecs as building blocks avoiding complexity that is exponential in n. The lower bound suggests that, at least mathematically, asymptotically bound to the Shannon capacity of an (n, n) system is a system comprised of n OC(k) systems with k = 1, 2, ...n. Recently a layered space-time architecture associated with the asymptotic lower bound was found. The construction uses only 1–D codecs as building blocks. The expressive lower bound (7) played a key role in discovering this architecture yet the system is realized in terms of n systems of equal capacity.

We give a concise explanation of this new structure which is illustrated in Figure 4 for n = 6. The general n case will be understandable from this special case. Within a baseband equivalent context, we begin with a discussion of transmission. First the primitive data stream is demultiplexed and then each of the six equal rate substreams goes through an encoding/modulation stage. Each of the six encoders/demodulators can proceed independently of the others. Instead of a fixed association of these substreams with antennas, the association is periodically cycled. The dwell at each association is τ seconds. One full cycle takes 6τ seconds ($n\tau$ in the general case). As will eventually be clear, the architecture that we are describing attains the capacity lower bound in the limit of large τ.

Next we describe the receiver. The lower part Figure of 4 shows space-time as one large rectangle. The rectangle base is time where the basic tick is the time to send one symbol. The duration τ comprises many ticks, in the limit, infinitely many. Space is also discrete: a six point space one for each transmit antenna. Signals are preprocessed iteratively along space-time diagonals, ie, we "peel off" one diagonal after another. We will discuss the reception of the highlighted *a* diagonal. There are

two stages to reception: preprocessing and final processing. We start by describing the preprocessing.

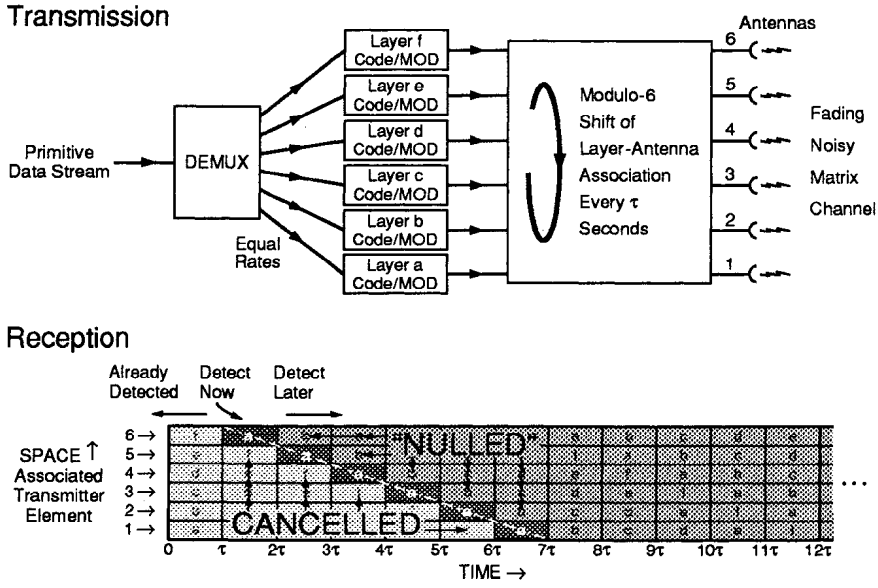

FIGURE 4: The layered space-time communication architecture. Transmit and receive methods for a (6,6) system.

The preprocessing has two parts: cancellation and nulling. Below the highlighted diagonal all detection is assumed completed and interference subtracted off. It is assumed that this interference cancellation is perfect - just as in decision feedback systems. Above the highlighted diagonal no detection has taken place yet, so interference is nulled. Resolve the received vector into two orthogonal components, one in the space of the interfering vectors. By nulling, we mean the transformation of the 6–D received vector to the other of the two component vectors, the one disposed orthogonal to the space spanned by the uncancelled interferers. So we are nulling out the component of what is received that suffers interference. Analysis of the six successive nulling processes enables us to see how the lower bound is approached. We discuss the nulling process counting *complex* DOF. (In other words, half the DOF compared to totaling real plus imaginary DOF.) Look at the six rectangles comprising the *a* layer from NW to SE. The NW rectangle involves no nulling so it is received as OC(6). For the next rectangle the nulling projects the signal into a 5-D space so it is processed as OC(5), the next rectangle as OC(4) and so on until we get to the last rectangle. This last rectangle is received OC(2) since all five of the overlying layers must be nulled. We see the six OC(6) systems k = 1, 2, · · · 6 just as the bound hinted. From a mathematical standpoint the six systems can be seen to involve statistically independent chi-squared variates [2].

The final processing is the decoding process applied to the preprocessed signal. We leave the coding/decoding details open. As technology progresses this will, of course, evolve. However, it is important that the detection of the *a* layer be completed and subtraction take place before the next layer is detected. Then we can assume that the interference effect from this layer has been cancelled.

The reader may wonder: What about error propagation? The answer is to invest the huge capacity available to take advantage of powerful error protection. What is important is that failure of a burst due to error propagation be say one order of magnitude below failure due to outage. Take the view that whether the outage is 1% or 1.1% is inconsequential, so error propagation is inconsequential. So powerful 1 – D codes (whose existence is guaranteed by information theory) will be essential to efficient use of bandwidth. We stress that Figures 1-3 show the relative closeness of the lower bound to the (n, n) limit improving as n increases.

7. CLOSING REMARKS

Using information theory we have shown how it is theoretically possible to achieve great capacities through judicious use of MEAs at both transmitter and receiver. For (n, n) systems we have seen how in the large ρ large n realm that capacity increases linearly with n as if the waves transmitted from different antennas did not interfere except to reduce ρ to ρ/e. The capacity lower bound we derived suggested that an architecture of enormous capacity existed. We went on to explain this architecture which utilized signals stratified along diagonals in space-time.

In assuming lattices at both transmit and receive sites with spacing of antenna elements of the order of λ we have tacitly avoided the onset of supergain MEAs in our analysis. In the appendix we discuss the theoretical possibility for even significantly greater capacity than uncovered so far – by impinging on supergain.

APPENDIX

SUPERGAIN: SOME CURRENTLY IMPRACTICAL ANTENNA POSSIBILITIES

In considering the ultimate capacity limits achievable with antenna arrays, the notion of supergain offers tremendous possibilities. Supergain is defined as the use of amplitude and phase combinations along with antenna element positioning in an antenna array causing the array to have an overall gain, G, higher than that associated with its projected aperture area, A, in square wavelengths (i.e., $G > 4\pi \cdot A$).

Although we are mainly concerned here with local communications such as wireless indoor systems, to clearly introduce the supergain concept, we first examine a free space thought experiment. For example, consider a transmitting horn antenna, with an aperture about 10 wavelengths on a side, located in outer space roughly aimed at the earth. With a one wavelength diameter supergain receiving antenna on the earth it is possible to receive virtually all of the power radiated by the horn antenna. The surface of the supergain antenna is covered with microscopic elements connected through a lossless matching network to a receiver. The phases and amplitudes of all the elements are adjusted so that, when operating as a transmitter, the earth antenna would produce at the horn antenna aperture an exact duplicate of its desired receiving field and would produce zero field in all other directions. In other words, on the sphere, centered on the earth antenna, we have specified the tangential electric field. This tangential electric field can then be expanded in spherical harmonics (see [11] Section 7.3) for outgoing waves and extrapolated back to the earth antenna to determine the required amplitude and phase to be produced over the surface of the earth antenna. The existence of such a solution depends only on the completeness of the expansion [op. cit.]. The mode field produced at the aperture of the horn antenna by the earth antenna is completely received by the horn without reflection; i.e., the horn antenna receives all the power radiated by the earth antenna. By reciprocity, the earth antenna must receive essentially all the power radiated by the horn antenna. Thus, instead of hundreds of dB loss due to the great distances involved, there is 0 dB loss in transmission achieved by the supergain of the earth antenna. Thus, the academic answer to the ultimate limit of achievable capacity is ridiculously large.

Now let us change the thought experiment to the case of indoor wireless communications, where the transmitting horn now appears to the receiver as a myriad of horns formed by the imaging property of the reflecting walls and objects in the building. Instead of matching the supergain antenna far field to a single horn aperture, we now match it to all of the horn apertures of the original horn and all of its images. Although many of the images may be pointing away from the receiver, and the reflecting walls affect the mutual coupling of the supergain antenna, we can still receive a sizable fraction of the transmitted power, so that our transmission loss to the supergain antenna would be on the order of 10 dB instead of about 100 dB without supergain.

The only problem with supergain is that it is impractical beyond about a 3 dB increase. Antenna current fluctuations with position on the supergain antenna are extreme and require extreme precision. Any loss in the antenna material reduces the antenna efficiency to virtually zero. The precision requirement typically reduces achievable bandwidth to infinitesimal values. These properties are a result of the very high order spherical harmonics required to match the desired far field. Extrapolation of the high order harmonics towards the center of the far field sphere quickly encounters exponential growth. Thus, although supergain is an academic possibility, it is totally impractical for realizable systems.

A possible item for future consideration would be to refine the very preliminary ideas in this appendix to bring in practical considerations like the effects of imperfections and also impose design constraints that preclude inordinately high currents. Also the possible impact of continuing advances in superconductivity would seem worthwhile exploring. A prime objective would be to translate practical effects and constraints into mathematical form so one could probe analytically antenna constructs offering greater capacities while at the same time insuring practicality.

REFERENCES

[1] G. J. Foschini and M. J. Gans, On The Limits Of Wireless In A Fading Environment When Using Multiple Antennas, Accepted for publication in Wireless Personal Communications.

[2] G. J. Foschini, Layered Space-Time Architecture for Wireless Communication In A Fading Environment When Using Multi-Element Antennas, Bell Labs Technical Journal, scheduled for Vol. 1, No. 2, Winter 1996.

[3] M. S. Pinsker, "Information and Information Stability of Random Processes", San Francisco: Holden Bay, 1964, Chapter 10.

[4] W. C. Jakes, Jr., Microwave Mobile Communications, John Wiley and Sons, New York, 1974, Chapters 1 and 5.

[5] G. J. Foschini and R. A. Valenzuela, Initial Estimation of communication Efficiency for Indoor Wireless Channels, Accepted for Publication in Wireless Networks.

[6] J. Wolfowitz, Coding Theorems of Information Theory, Springer-Verlag, New York, 1978.

[7] E. Csiszar and J. Korner, Information Theory: Coding Theorems for Discrete Memoryless Systems, Academic Press, New York, 1981.

[8] P. Lancaster and M. Tismenetsky, The Theory of Matrices, Academic Press, 1985, pg 46.

[9] A. Edelman, Eigenvalues and Condition Numbers of Random Matrices, M. I. T. Doctoral Dissertation, Mathematics Department, May 1989.

[10] T. W. Anderson editor, S. S. Wilks: Collected Papers Contributions to Mathematical Statistics, John Wiley and Sons, New York, 1967.

[11] J. A. Stratton, "Electromagnetic Theory", McGraw-Hill Book Company, Inc., New York, 1941.

This page intentionally left blank.

ON THE PERFORMANCE OF A MEDIUM ACCESS CONTROL SCHEME FOR THE RECONFIGURABLE WIRELESS NETWORKS

Zygmunt J. Haas

School of Electrical Engineering
Cornell University, Ithaca, NY 14853
http://www.ee.cornell.edu/~haas/wnl.html

Abstract

In this paper, we propose a MAC protocol for an ad-hoc network architecture, termed by us the Reconfigurable Wireless Networks, *and investigate its performance. We discuss the access synchronization and the construction of network connections features of the MAC protocol. Using event-driven simulation, we evaluate the throughput and the blocking probability of the scheme, demonstrating its behavior for different parameters of the mobility, traffic, and propagation models.*

Introduction

A *Reconfigurable Wireless Network* (RWN), which was introduced in [1], is an ad-hoc network architecture that can be rapidly deployed, without relying on a preexisting fixed network infrastructure. The nodes in a RWN can dynamically join and leave the network, frequently, often without warning, and without disruption to other nodes' communication. RWNs are a special case of the ad-hoc network architecture and distinguish themselves in the following features: large network span, large number of nodes, and highly mobile nodes. In particular, the node

Portions reprinted, with permission, from *Proceedings of MILCOM'97, Monterey, CA, November 2-5, 1997.*

constellation can change rapidly, and so is the presence or absence of links. Examples of the use of the RWNs are:

- military (tactical communication) - for fast establishment of a communication infrastructure during deployment of forces in a foreign and hostile terrain,
- rescue missions - for communication in areas without adequate wireless coverage,
- national security - for communication in times of national crisis, where the existing communication infrastructure is non-operational due to a natural disaster or a global war,
- law enforcement - similar to tactical communication,
- commercial use - for setting up communication in exhibitions, conferences, or sale presentations,
- education - for operation of virtual classrooms, and
- sensor networks - for communication between intelligent sensors (e.g., MEMS[1]) mounted on mobile platforms.

In contrast with the conventional cellular networks, in the RWNs, direct mobile-to-mobile (**peer-to-peer**) communication is allowed. Mobiles that cannot be reached directly, use other mobiles to relay their transmissions using **multi-hop** routing.

There is a number of central issues that need to be addressed to make the RWN a workable system: the mobility management, the routing protocol, and the Medium Access Control (MAC) scheme. In this paper, we deal with the MAC scheme. A routing protocol applicable to the RWN architecture – the *Zone Routing Protocol* - was introduced in [2] and further evaluated in [3]. Finally, we will report on the RWN mobility management scheme in a future publication.

In general, the MAC for wireless networks has been a subject of extensive studies, mostly concentrating on random access mechanisms. The traditional MAC schemes for radio networks, such as the Carrier Sense Multiple Access (CSMA), do not fully address the "hidden terminal problem." Newer schemes, such as *MACA* [4], *MACAW* [5], and *FAMA* [6] rely on low-level reservation mechanism. In these protocols, the RTS/CTS (Request-To-Send/Clear-To-Send) dialog precedes any communication between nodes. The MACAW protocol includes an additional feature of acknowledging, on the link layer, the received transmissions and, thus, provides fast recovery from collisions. In the FAMA protocol, the non-persistent CSMA scheme is used, which avoids repetitious collisions.

It should be noted that all these protocols do not resolve the hidden terminal problem and, therefore, do not prevent collisions. The notorious example is when a node, which did not hear the RTS/CTS dialog, migrates to within reception range of the already communicating nodes. A similar situation occurs when a node (due to transmission errors, for example) did not hear the CTS reply. In such cases, any transmission attempt of this node will most probably result in a collision. Additionally, because of the random access nature of these schemes, reservations of network capacity for real-time multimedia traffic cannot be easily accommodated.

[1] Micro Electro Mechanical Systems

As opposed to random access schemes, controlled-access schemes do not suffer from the above shortcomings. Controlled-access schemes have been considered before for multi-hop network environment; for example, the *Cluster TDMA* scheme [7]. The difficulty with these schemes, however, is that they require network-wide synchronization. This appears to be a significant problem, especially in a fast changing communication environment, such as the RWNs.

We propose here a hybrid of random- and controlled-access schemes, based on a local polling mechanism. It exhibits controlled-access behavior among neighbor nodes due to polling. Yet, the polls originating at different nodes are random and unsynchronized, and, thus, may lead to collisions. However, as the length of the polls is short compared with the actual data transmission and as the number of competing nodes is restricted to a subset of the neighbors only, the loss of capacity due to the random-access process is relatively small. The salient features of our scheme are that it requires local synchronization only (i.e., among the neighbors) and that it eliminates collisions of payload information. Additionally, it reduces the access time variance and inherently allows capacity reservation. Thus, it is applicable to multimedia traffic.

Medium Access Control (MAC) Protocol

The purpose of a MAC protocol is to synchronize access of mobiles to the shared communication medium. In the RWN, the span of the network is considerably larger than the transmission range of a single node transceiver. This necessitates multi-hop routing.

A MAC can be based on a single shared-channel[2] or multiple channels. Use of a single shared-channel MAC scheme requires each transceiver to operate at the maximum channel bit-rate. In addition, single shared-channel can lead to inefficient operation or to the need for network-wide access synchronization. Thus, a multi-channel solution, in which nodes are assigned individual channels, is a more viable option. Of course, in the multi-channel approach, channels can be reused in spatially separated areas and can be based on different multiple access schemes, such as TDMA or CDMA. Our work is independent of the actual multiple access scheme used. The determination of the actual multiple access scheme, as well as the channel/code allocation algorithm, are outside the scope of this work.

MAC-layer connections are established between neighbors.[3] With time, mobiles can become separated by large enough distance or by a radio propagation obstacle, so that connections will be broken. Likewise, mobiles can migrate close enough to each other to become neighbors and to establish new connections.

We assume the following elements of the MAC scheme:

- The transmission spectrum is divided into channels. For the sake of simplicity, we assume here that channels are created in time and frequency (FDMA/TDMA), rather than in code (i.e., CDMA). Nevertheless, the presented-here MAC scheme can be easily modified for CDMA-based transmission.

[2] In this work, we refer to a *channel* as a TDMA slot assignment on a particular frequency channel, or, in CDMA, as an assignment of one of the orthogonal codes. Note that this is not the traditional definition of a wireless channel.

[3] A mobile that can be reached directly by a mobile in question is referred to as the mobile's *neighbors.*

- Each channel consists of two parts, which we call the *payload* and the *poll* portions of the channel. The bandwidth of the payload portion is larger than the bandwidth of the poll portion. The capacities of the payload and the poll depend on several factors and performance requirements. We label the payload portion of the k^{th} channel as Θ_k, and the poll portion as φ_k.
- Nodes are dynamically assigned channels. The reuse of channels is such that it corresponds to an acceptable Signal-to-Interference (SIR) ratio at the assigned nodes. Channel assignment schemes will not be discussed here.
- Each node maintains an updated list of who are its neighbors and what are their assigned channels (*neighbor-table*). This list is created and maintained as part of the *Neighbor Discovery* process, which will be explained later.
- Every node is equipped with two half-duplex transceivers. Although, this is not an essential assumption, as our MAC protocol will also operate with a single transceiver per node, our study indicates that a considerable advantage in reliability and network capacity can be achieved with two half-duplex transceivers. We term the two transceivers: an *in-transceiver* and an *out-transceiver.*

In reference to this last point: inclusion of two transceivers per station increases the reliability of the network. In a peer-to-peer routed network, such as the RWN, failures of a single-transceiver node can lead to network partitioning and to loss of connectivity with some network nodes. On the other hand, our MAC protocol is designed in such a way that a node can still continue to operate even when a failure of one of the two transceivers occurs (although the node's capacity is reduced).

Access Synchronization

Access synchronization is based on the operation of the two node's transceivers. Ordinarily, one node's transceiver (referred to here as *out-transceiver*) is assigned to traffic from the node in question to other nodes, while the other transceiver (the *in-transceiver)* is used for communication in the other direction.

The operation of the access synchronization is based on every node continuously polling its neighbors to determine whether they have any communication to send. The polling is performed on the poll portions of the channels. For illustration, assume that node-*k* is assigned the channel (Θ_k, φ_k) and its two neighbors, node-*i* and node-*j*, are assigned frequencies (Θ_i, φ_i) and (Θ_j, φ_j), respectively. Node-*k* tunes its in-transceiver to φ_i, polling the node-*i* for any data and immediately retunes its in-transceiver for reception on Θ_k. As the out-transceiver of node–*i* is ordinarily tuned to φ_i, it receives the poll and, assuming it has nothing to send to node-*k*, it responds with End-Of-Transmission (EOT) message on Θ_k. Node-*k* proceeds now

to poll node-*j* on φ_j. Now assume that node-*j* has a number of packets to send to node-*k*. Node-*j*, after receiving the poll on its out-transceiver (ordinarily tuned to φ_j), tunes it to Θ_k and transmits the packets, ending with the EOT message. Node-*k* then starts a new poll cycle. The procedure is depicted in Figure 1, where the oval indicates tuning of the particular transceiver to the specified frequency.

Note that all the nodes perform the same procedure. Thus, at the same time as node-*k* polls its neighbors, the neighbor nodes (node-*i* and node-*j*) both poll node-*k*, which has his out-transceivers ordinarily tuned to its polling frequency, φ_k. In

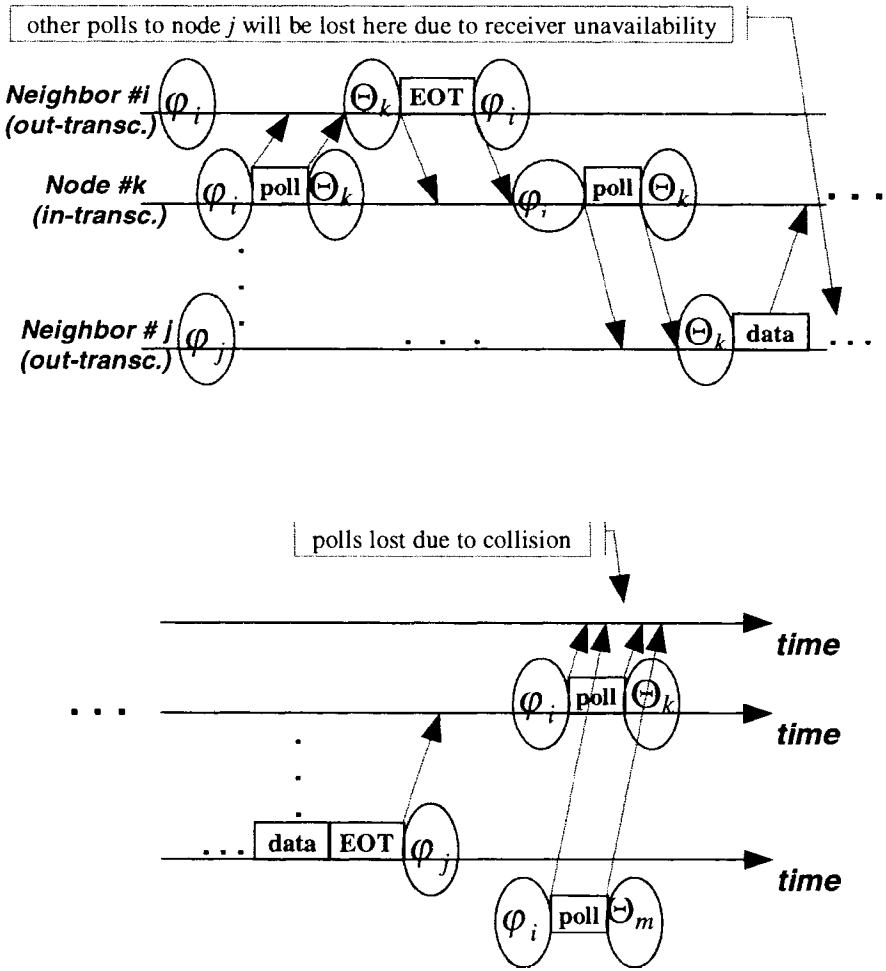

Figure 1: Access Synchronization Mechanism

summary, the process is as follows: In general, a node's out-transceiver is tuned to its own polling frequency. When a poll arrives, the node answers the poll (by either sending packets and/or EOT message) on the polling node's payload frequency. A

node's in-transceiver is used for polling its neighbors by tuning to their polling frequency, sending a poll message and retuning to its own payload frequency to hear the polled node answer.

Of course, as the polling channels are unsynchronized, there is a possibility of polls collisions and of some of the polls being lost. (See Figure 1 for an example.) If a polling node gets no response to its poll within the specified time-out interval (*poll_time-out),* it repeats the poll in the next cycles for up to maximum number of attempts (*max_poll_attempts).* If all the attempts fail, then the neighbor is placed on a watch-list. If no polls are received from the neighbor within *max_poll_cycle,* the link to the neighbor is considered lost and the neighbor identity is removed from the local *neighbor-table.*

A number of comments regarding the access synchronization scheme:

- As a node has full control over access to its payload portion of the channel, there are no collisions on the payload portion of the channel and, consequently, no packet transmission is lost due to collisions.
- While responding to a poll, a node can send only a limited number of packets. This number determines the maximum time a node can spend away from its frequency, *max_transmission_time.* The goal is to ensure that other nodes have a chance to communicate as well and to allow bandwidth reservations for multimedia traffic. However, when only few nodes want to communicate, the poll will return to the ready station sooner, allowing the station to capture larger portion of the channel.
- A possible modification is to allow a two-way traffic exchange, once a poll is answered. Such a modification increases reliability by preserving connectivity while loss of a single transceiver occurs.

Construction of network's connections

As the network nodes migrate, existing connections between adjacent nodes are broken and new connections are made possible. The *Network Construction* procedure is responsible for updating the connections' status. To accomplish this, each node maintains a list (*neighbor_table)* of all the nodes with which it can establish a direct link (referred to here as *neighbors). Neighbor Discovery* is the process by which a node learns about its new neighbors. (Note that the reverse process, removal of out-of-reach nodes from the *neighbor_table*, is triggered by missing poll responses, as explained above.) The *Neighbor Discovery* process is performed by the in-transceivers, scanning the available channels and discovering polling messages from adjacent nodes.[4] (Note that when a node has no neighbors, it still periodically sends null polling messages.) In particular, when a message is received with "good enough quality" (e.g., sufficient Signal to Interference Ratio), the node starts the neighbor acquisition process by including the new neighbor in its next polling cycle. After one successful polling response from the neighbor, the

[4] This process is inspired by the search of control channels for handoff purpose in a cellular network. For example, this process is implemented by the "idle slot" in GSM.

neighbor is added to the *neighbor_table.* The principle of the process is shown in Figure 2.

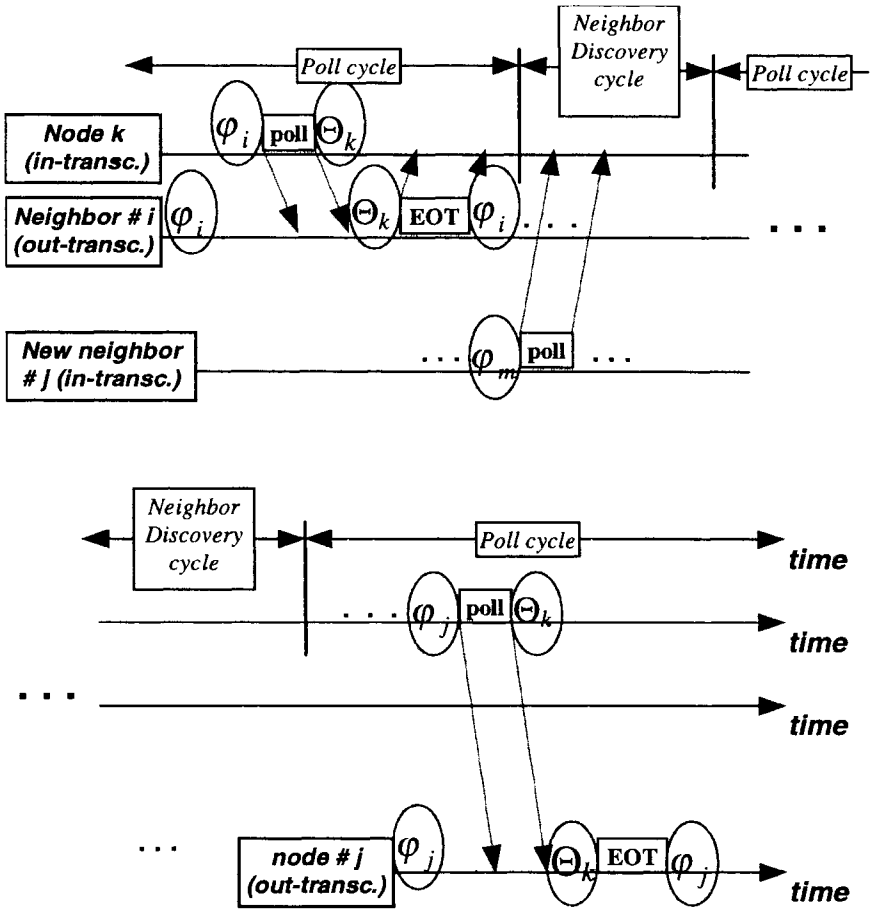

Fig 2: The Neighbor Discovery Mechanism

The Network, Traffic, and Radio Models

The coverage area of our network is a $1[km]\times 1[km]$ closed area (see [2] for the definition of a closed area). Initially, the mobiles are randomly distributed throughout this area. Each node is equipped with a transmitter/receiver pair, operating in the half-duplex mode. We assumed a narrowband system, in which the total spectrum is divided into channels through the FDM/TDMA technique. To make sure that the blocking is not spectrum- but system-limited, the system has large enough number of channels. The maximum transmission power of a mobile limits its connectivity to within an area of the radius $r_{\max}$.

Each node can be in one of the four possible states: *idle, receive, transmit,* and *transmit/receive.* The transitions between the states are shown in Figure 3.

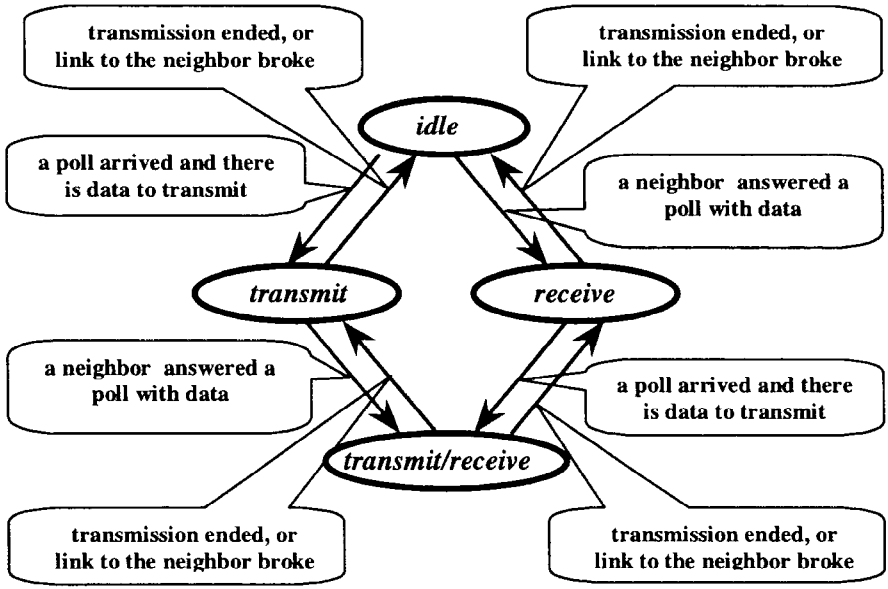

Fig 3: State transition diagram

Packets arrive to mobiles in a random manner with total inter-arrival time and the packet transmission time modeled as negative exponential distributions with parameters $1/\lambda$[sec/packet] and $1/\mu$ [sec/packet], respectively. There is no distinction between the local and the to-be-forwarded traffic. Packets have to be fully received before they are forwarded to a neighbor node. Packets that arrive to a mobile while the mobile is already transmitting are considered blocked. Upon arrival of a packet, a destination is randomly chosen from the set of neighbors. If the node has no neighbors, the packet is considered blocked. If the destination neighbor is already receiving from another node, the packet is again blocked. Note that due to the dual-transceiver feature, the transmission and the receiving processes are decoupled. In other words and for example, if a packet is destined to a transmitting, but not receiving node, the packet is not blocked. A packet is considered *dropped* if during its transmission the link between the communicating nodes is broken due to mobility.

Blocked packets can be treated in several ways. The appropriate action depends on the reason for blocking the packet. In general, one possibility is to queue blocked packets. Another approach would be to discard blocked packets and alert the sending node (i.e., the previous node) or the local application (if the packet is locally generated) that the packet cannot be forwarded. The second choice might make sense in the highly versatile RWN environment, since a route to the destination might be broken while the packet is queued, effectively discarding the

packet without notification. In such a case, a timeout mechanism and end-to-end retransmission would be necessary, at the expense of increased delay.[5] Discarding a packet with notification, on the other hand, might allow rerouting the packet through a different, available route. Possibly, a limited queue would be a better solution. In this paper, we assumed that a single buffer exists at each node and that blocked packets are discarded with notification. The treatment of blocked packets is still an open issue and we will not pursue it any further here.

The Mobility Model

Mobiles roam freely throughout the closed coverage area. At any point in time, mobile movement is characterized by two parameters, its velocity, $\vec{v}$, (value, v, and direction ϕ) and its current position, (x, y). The velocity is updated every time interval, Δt, according to the following formula:

$$v(t+\Delta t) = \min[\max(v(t)+\Delta v, 0), v_{\max}], \quad (1)$$

where Δv is a uniformly distributed random variable in $[-\alpha \cdot \Delta t, \alpha \cdot \Delta t]$, $v_{\max}$ is the maximal mobile velocity, and α is the maximal mobile node acceleration/deceleration (assumed here to be of equal magnitude). The movement direction, ϕ, is also updated every Δt:

$$\phi(t+\Delta t) = \phi(t) + \Delta\phi, \quad (2)$$

where $\Delta\phi$ is a uniformly distributed random variable in $[-\Delta\phi_{\max} \cdot \Delta t, \Delta\phi_{\max} \cdot \Delta t]$, and $\Delta\phi_{\max}$ represents the maximal angular change per unit time. This mobility model is an improvement on the previously used in the literature Brownian Motion model. The initial velocity $\vec{v}$ of the mobiles is assumed random, within the above-mentioned constraints.

The Simulation

We have simulated our MAC protocol with the above models by an event-driven simulation. The purpose of the simulation was to determine the effect of different parameter setting on the performance of the network. In particular, we evaluated the average system blocking and the total system throughput as a function of the number of mobiles in the system (M), maximal transmission radius ($r_{\max}$), maximal velocity ($v_{\max}$), the length of the Discovery Cycle (τ_{disc}), and the system loading (λ^{-1}).

[5] Of course, retransmission is not an option for real-time traffic and, thus, discarding packets without a notification to a previous node would result in reduced communications quality.

Numerical Results

Our simulation was tested for the range of parameters' values, as shown in the following Table. When the effect of a parameter was considered, all other parameters were assigned their respective default values.

Parameter	Symbol	Default Values (Range)
Coverage area	$S \times S$	1[km] X 1[km]
Discovery Cycle duration	τ_{disc}	100 [msec] (100[msec] - 300[msec])
Discovery Cycle deviation	$\Delta\tau_{disc}$	10 [msec]
Poll Cycle duration	τ_{poll}	900 [msec]
Poll Cycle deviation	$\Delta\tau_{poll}$	50 [msec]
Number of mobiles	M	40 [mobiles] (20 - 60)
Time increment	Δt	100 [msec]
Maximal transmission radius	$r_{\max}$	500 [meter] (100[m] - 300[m])
Maximal angular change	$\Delta\phi_{\max}$	$0.2 \cdot \pi$ [rad/sec]
Maximal velocity	$v_{\max}$	65 [mph] (15[mph] - 195[mph])
Acceleration/Deceleration	α	0.5 [m/sec^2]
Average packet transmission duration	μ^{-1}	100 [msec]
Average call inter-arrival time	λ^{-1}	0.5 – 10 [packets/sec]

The Discovery Cycle and Poll Cycle deviations were introduced to eliminate synchronized blocking phenomenon. Furthermore, in our simulations, we observed that the value of the time increment (Δt), has very little effect on the results of the simulation.

The results of the simulation are shown in Figures 4 - 11, The "Packet Arrival Rate" is the per-node total arrival rate. The "Throughput" is the aggregated per-node throughputs of all the network nodes[6]. Of course, to learn the effect of the various network's parameters, the relative, rather than the actual values of the throughput, are of most interest. The "Blocking probability" is the probability that an arrival packet cannot be immediately forwarded to the next node on its path. Note that this does not imply that these packets are necessarily lost, but it does indicate on how efficient the MAC is in forwarding packets.

[6] Note that this is not the traditional definition of network throughput. This definition was adopted because we are interested in the performance of the MAC scheme, rather than other network algorithms, such as the routing algorithm.

Figure 4 and 5 show the effect of changing the number of network nodes. With some number of nodes, the increase in the packet arrival rate results in smaller than linear increase in the throughput, due to the fact that the blocking probability increases as well with the increase in the packet arrival rate. From these figures, the throughput increases linearly with the number of mobiles and the blocking probability is not significantly affected by the number of users. This is the result of the fact that these figures were calculated under the assumption that the packet arrival rate per node is independent of the number of nodes. In general, increasing the number of nodes increases the per node arrival rate and this increase depends on the routing algorithm and on the average path length. That is, given the routing algorithm (and average path length), one could use these graphs to estimate the actual nodal throughput and blocking probability.

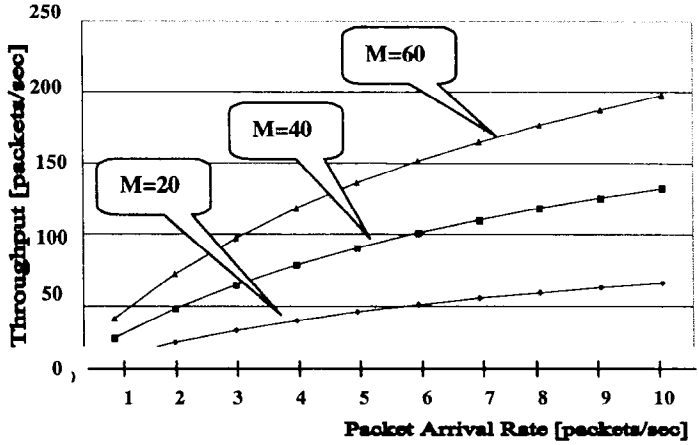

Fig 4: The effect of the number of users on throughput

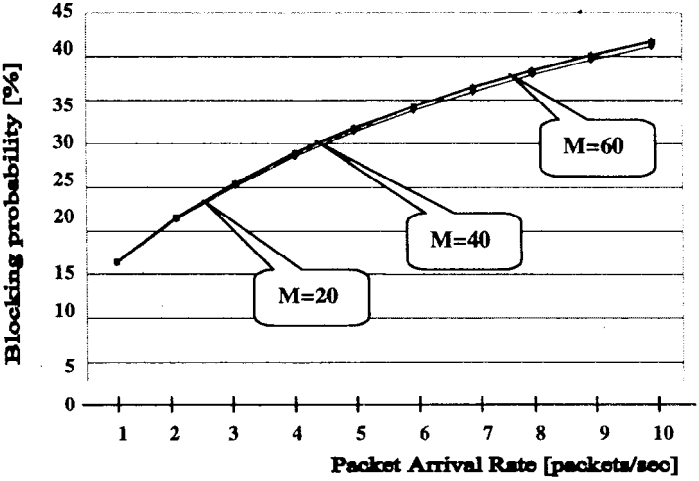

Fig 5: The effect of the number of users on blocking probability

Our conclusion from Figures 4 and 5 is that, with fixed per-node arrival rate, the nodal throughput and the blocking probability are independent of the number of network nodes.

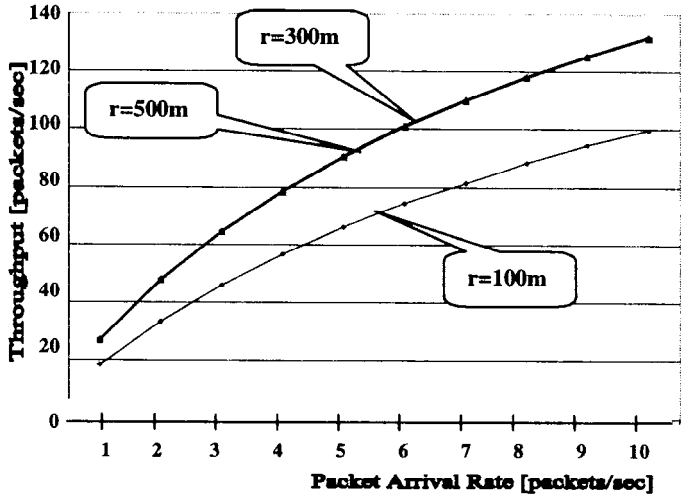

Fig 6: The effect of the transmission radius on throughput

Figures 6 and 7 show the effect of sizing the mobiles' transmission radius. The larger is the radius, the more neighbors each node has. Thus, the chance that a node has no neighbors (and, thus, blocks every packet) is minimized. Also, occupancy of

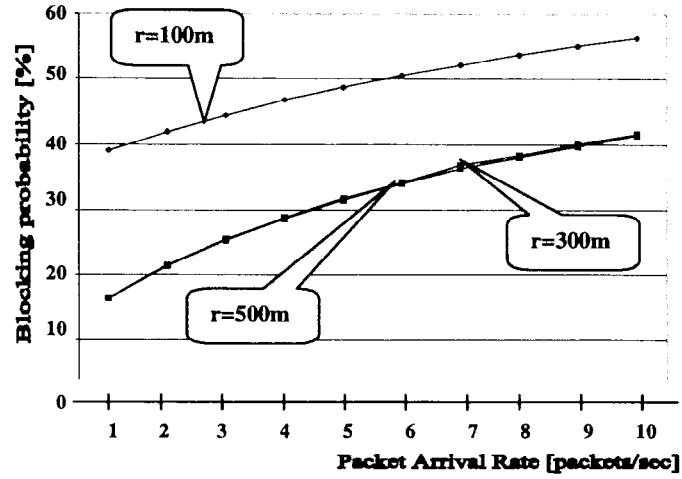

Fig 7: The effect of the transmission radius on blocking probability

the neighbor nodes tends to be more uniform. Consequently, the blocking probability decreases and the throughout increases. However, note that increasing the radius beyond the 100-200 [m] range has very little effect on the system performance. This reinforces and further qualifies our conclusion from Figures 4 and 5 that the number of network nodes has little effect on nodal throughput and blocking probability, as long as the number of nodes is larger than some minimum. (Or in other words, as long as each mobile has at least one neighbors that can be used to forward the received packets.)

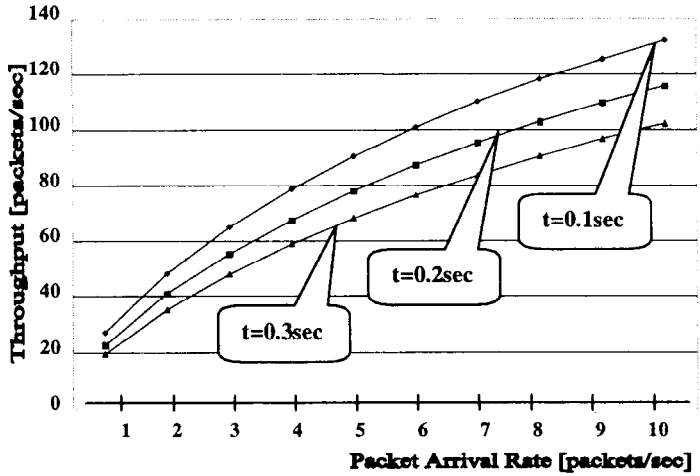

Fig 8: The effect of the Discovery Cycle length on throughput

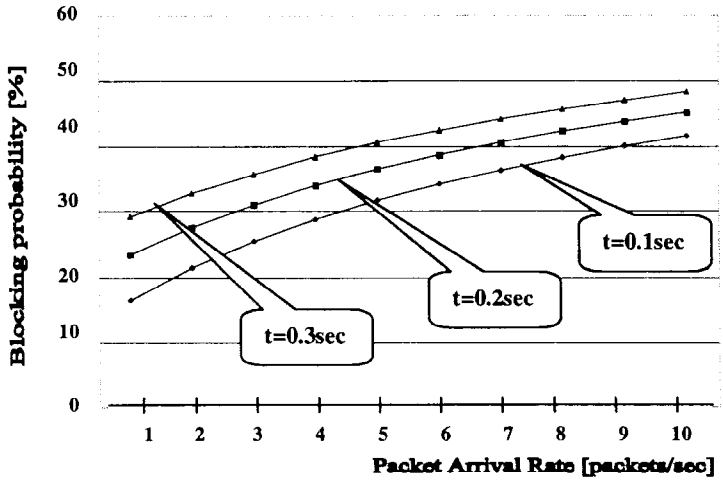

Fig 9: The effect of the Discovery Cycle length on blocking probability

Figures 8 and 9 depict the effect of the Discovery Cycle length. As anticipated, the longer the Discovery Cycle is, the lower is the throughput and the higher is the

probability of blocking. However, longer Discovery Cycle does not correspond to proportionally smaller throughput. The blocking probability exhibits similar behavior. The reason is that the Discovery Cycle is responsible for only part of the blocked packets; other packets are blocked due to unavailability of the neighboring node transmitting or receiving from other nodes. This accounts for larger blocking probability, especially at higher loads, that could have been otherwise predicted from the portion of the total time that a node spends in the Discovery Cycle. The results in Figures 8 and 9 suggest that, if kept to a reasonable length, the Discovery Cycle does not carry prohibitively large overhead.

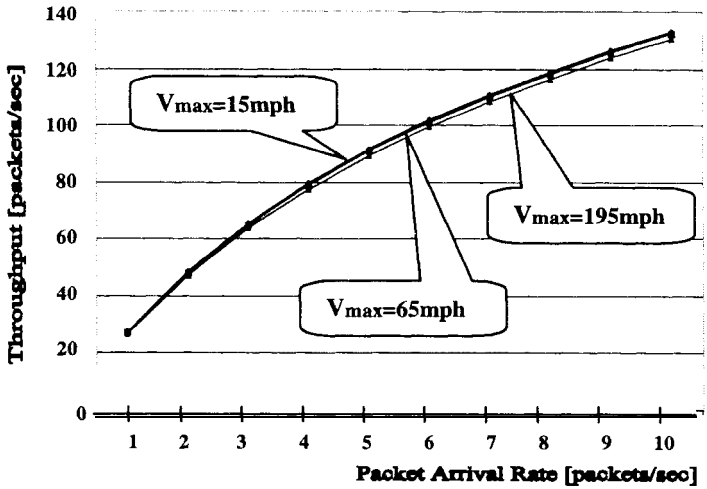

Fig 10: The effect of mobile velocity on throughput

Varying the user mobility (Figures 10 and 11) has only a small effect on the system performance. This can be explained by the fact that as the transmission time of the packets is so short, loosing the link in the middle of a packet transmission is rather a low-probability event. Connections do stay long enough not to affect majority of transmitted packets. Comparing this result with a connection-oriented transport ([1]) provides a strong indication that the preferred mode of operation of the RWNs is in connectionless communication.

Concluding Remarks

We have introduced here a novel MAC scheme, which is based on polling mechanism with local synchronization requirements only. Operation of the scheme relies on partitioning the channels (TDMA or CDMA) into two parts: the poll and the payload portions. The poll portions of the channel use random-access mechanism to perform the polling operation and are, thus, subject to collisions. However, because of the restricted number of neighbors and the short duration of

the poll messages, the frequency of collisions is quite small in a practical network scenario. Transmission on the payload portion will not undergo collisions.

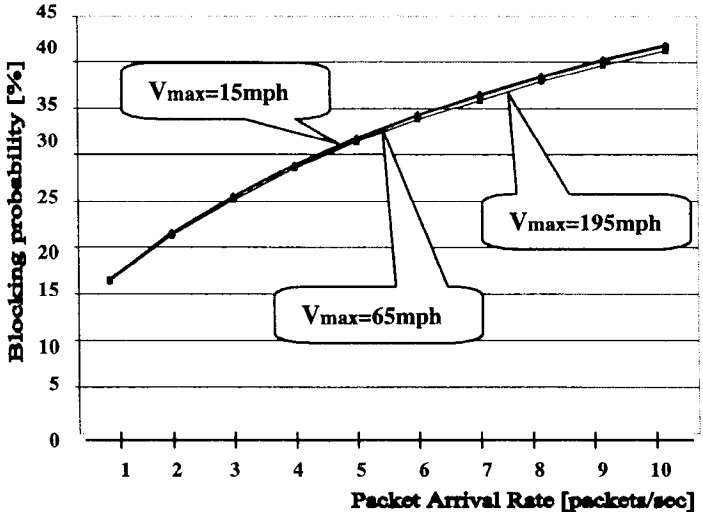

Fig 11: The effect of mobile velocity on blocking probability

Our evaluation of the proposed MAC scheme reveals that, for the set of parameter values considered, its performance in terms of nodal throughout and blocking probability is independent of the mobile population. This is correct as long as some minimal number of nodes are present, so that the probability that a node has no neighbors in small. This behavior translates to a minimum transmission radius, given some mobile density. Note that this conclusion addresses the MAC performance only and not the routing scheme. In fact, in another study, we have shown that the density of user population has a crucial effect on the probability of finding a route within a network [3]. However, from the MAC point of view, as long as there is at least one neighbor to which the packet can be forwarded, the performance of the MAC scheme is relatively invariant with the actual number of such neighbors.

Another conclusion that can be drawn from our study is that the Neighbor Discovery cycle is not a major source of blocking, as long as it is not excessively long. Further study is required to determine the required duration of the Neighbor Discovery cycle, which depends on the number of channels and mobile density.

Finally, we have demonstrated that the mobile's speed has only minor effect on the nodal throughput and blocking probability. This is an encouraging result as it indicates that, under the range of parameters considered, the dropping probability is not strongly affected by the mobility. It also suggests, as we have suspected, that connectionless mode of communication is the proper scheme for the RWN environment.

REFERENCES

[1] Z.J. Haas, "On the Relaying Capability of the Reconfigurable Wireless Networks," *VTC'97*, Phoenix, AZ, May 5-7, 1997.

[2] Z.J. Haas, "A New Routing Protocol for the Reconfigurable Wireless Networks," *ICUPC'97*, San-Diego, CA, Oct. 12-16, 1997.

[3] Z.J. Haas, "Providing Ad-Hoc Connectivity with the Reconfigurable Wireless Networks," in preparation.

[4] P. Karn, "MACA – a New Channel Access Method for Packet Radio," *ARRL/CRRL Amateur Radio 9th Computer Networking Conf.* ARRL, 1990

[5] V. Bharghavan *et al*, "MACAW: A Media Access Protocol for Wireless LANs," *ACM SIGCOMM'94.*

[6] C.L Fullmer and J.J. Garcia-Luna-Aceves, "Floor Equisition Multiple Access (FAMA) for Packet-Radio Networks," *ACM SIGCOMM'95.*

[7] M. Gerla and J.T.-C. Tsai, "Multicluster, Mobile, Multimedia Radio Network," *ACM/Baltzer Journal of Wireless Networks*, vol. 1, no.3, 1995.

CODING AND NETWORKING TECHNIQUES FOR RADIO NETWORKS

Fulvio Babich
and Francesca Vatta

DEEI, Università di Trieste,
Via Valerio 10, 34127 Trieste, ITALY
babich,vatta@uts.univ.trieste.it

Abstract: We study the interaction of source and channel coding, and of coding and multiplexing techniques, given a multi-mode, Variable Bit Rate (VBR), and embedded source coding scheme. We estimate the coded bit-stream resistance to channel impairments and frame losses. Our aim is to increase network capacity while maintaining a smooth quality degradation at high loads and heavy interference. We show that, by exploiting the peculiar characteristics of the chosen coding technique, which allows scalability of the signal bandwidth and of the delivered bit-rate, it is possible to achieve a strong interaction between the system and the bit-stream. Both narrowband and wideband speech services are evaluated.

1 INTRODUCTION

Nowadays there is an evolution towards the creation of a single network infrastructure where different kinds of information can coexist. A key element is represented by the wireless network, that may make it possible to transfer any kind of information between any desired locations, but whose capacity and reliability are strictly limited.

Our attention is addressed towards speech/audio coding and networking techniques for wireless networks, which are suitable to increase network capacity while maintaining a smooth degradation of quality at high loads and heavy interference. In the paper we show that embedded and variable bit-rate (VBR) multimode source encoding should be used to fully exploit the capacity of the system. Indeed, such coding schemes

solve the problem of having the output bit-rate vary in an optimal sense, in response to source entropy, channel quality, and network congestion.

In the paper, the interaction of source and channel coding, and of coding and multiplexing techniques is investigated. Both narrow band and wide band audio services are examined, considering different switching techniques. First, circuit switching applications are taken into consideration, to study the interaction between quality of service and transmission reliability. A possible trade-off between audio information bandwidth and channel quality is investigated [1]. Then, packet switching is considered, to assess the potential benefits we can obtain by using the coding scheme under investigation.

The paper is organized as follows. Section 2 describes the used audio encoder. Section 3 presents the bandwidth requirements and the tolerable discarding rates, for speech transmission. Section 4 deals with circuit switching applications, while Section 5 covers packet switching applications. Section 6 summarizes the main conclusions.

2 AUDIO CODER

The Mobile Audio Visual Terminal (MAVT) audio encoder-decoder we use, adopts a flexible, object oriented coding scheme. It consists of a low band core algorithm which can contain any speech or music compression scheme, and of an arbitrary number of low band and high band enhancement stages. The number of enhancement stages allocated to a certain band depends on the result of energy and Signal to Noise Ratio (SNR) evaluations and on the available bit-rate. Core bits identify the minimum amount of information needed to provide an output signal. Enhancement bits are those relevant to the additional information which, if delivered to the destination, will provide the maximum quality, but which can be discarded with no need to change the core information. This coding scheme allows complexity scalability and interoperability with other standards, since a plurality of core low-rate algorithms/tools can be used by the algorithm.

The speech core algorithm we use is a Fast Variable Rate - Code Excited Linear Prediction (FVR-CELP) encoder, developed by CSELT S.p.A. [2]. The algorithm is designed for speech sampled with a sampling frequency of 8 kHz. The signal is segmented in frames of 10 ms, that is 80 samples. Comfort noise insertion is used to replace silence. There are several different coding rates, corresponding to different operating modes, as shown in Table 1. The bit-rate varies between 400 bit/s and 16 kbit/s: the lower two rates are used to represent noise, while the higher five ones are meant for speech. The average rate depends on speaker characteristics and on background noise. The typical average rate is 6 kbit/s. Current MAVT implementation adopts a slightly modified FVR-CELP algorithm, in which the signal is segmented in frames of 32 ms, that is 256 samples. The 32 ms frame is split into two 16 ms sub-frames. The core algorithm is applied twice per frame.

Mode	1	2	3	4	5	6	7
Rate [kbit/s]	0.4	4.1	8.5	12.5	7.2	12	16

Table 1 Coding modes and bit rates

The quality achieved in the core stages can be improved by adding a certain number of enhancement stages. The signal can be split in subbands of 1 kHz each (a narrow band input signal, with a 4 kHz bandwidth, can be split in 4 subbands). An arbitrary number of enhancement stages can be added for each sub-band: from 1 to 7 in the low band (i.e. 0-4 kHz) for a music signal, from 1 to 3 in the low band for a speech signal, from 1 to 3 in the high band (i.e. 4-20 kHz) for all types of signals. The number of enhancement stages allocated to a certain band depends on its energy, on the Signal to Noise Ratio (SNR) of the reproduced signal, and on the available bit-rate. Bit-rate scalability with a graceful quality reduction is one of the main features of the algorithm. Scalability is achieved by using the *bit-manipulation* unit. Scalability can be done at frame level, so that the bit-rate can be lowered only for a short time (i.e. 32 ms), and then raised again. The rate reduction is achieved by a specific algorithm which determines the number of enhancement stages to be discarded and then identifies those among them which can be discarded with the minimum impact on the output quality.

3 BANDWIDTH REQUIREMENTS AND DISCARDING THRESHOLDS

Bandwidth Requirements

In this section we present some preliminary results on relationship between bit-rate and output quality for wide band signals, whose sampling frequency is 16 kHz.

A large set of wide band speech signals has been processed and evaluated, by selecting a range of bit-rates that is upper bounded by the maximum bit-rate, corresponding to the best achievable quality. To quantify the objective speech degradation, an objective distance measure is employed, namely, the log-spectral distance measure (SD) [3], which quantifies the spectral degradation due to processing. The evaluation of the objective speech quality degradation, due to a bit-rate reduction, is performed, first, considering a bit-rate reduction applied to the whole voice signal (*all modes* curve in Fig. 1), then, lowering the bit-rate only in the frames corresponding to a given mode. Only modes representing activity are considered, as modes representing inactivity can be suppressed during transmission. The degradation is evaluated with respect to the signal coded with the maximum bit-rate, that is about 50 kbit/s, which is selected automatically by the encoder on the basis of a SNR saturation.

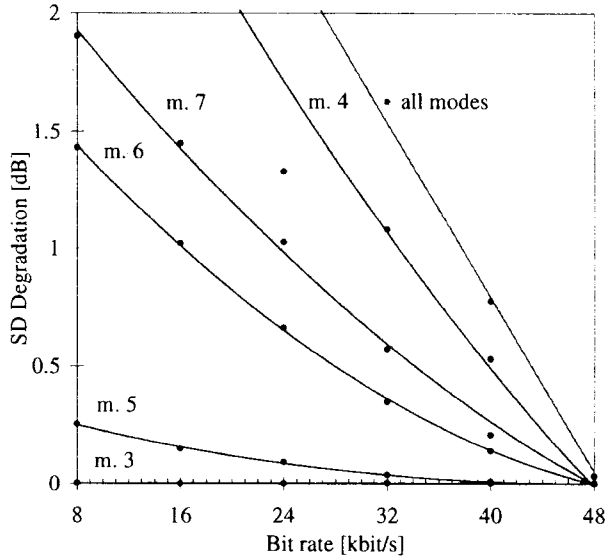

Figure 1 Wideband speech: Spectral distance degradation vs. selected bit rate

From Fig. 1 we observe that mode 4, which represents unvoiced speech segments with the highest energy, suffers a more evident quality degradation, when the required output bit-rate is decreased. By a comparison between objective measurements and informal subjective measurements, done by using the Absolute Category Rating (ACR) method [4], we can infer that the acceptable spectral distance degradation threshold is about 1 dB: with such a threshold, the different bit-rates needed by the different modes are, approximately, 32 kbit/s for mode 4, 24 kbit/s for modes 6 and 7, <8 kbit/s for modes 3 and 5.

Frame Discarding Sensitivity

For narrow-band signals, the basic information is represented by core bits, because they assure toll quality achievement by themselves. In such a case, quality reduction can be originated by frame discarding due to poor channel quality or network congestion. The FVR-CELP algorithm's bits are classifiable in groups of different importance. Selective discarding of less important bits permits a smooth quality degradation, in case of congestion or severe channel conditions. Moreover, each group of the coded bit-stream can have a particular allocation of bits between source and channel coding stages, leading to an unequal Forward Error Correction (FEC) coding scheme, which

can match its protection power with the source sensitivity. In this way a higher robustness of the coded bit-stream to channel impairments can be obtained.

In [5], four bit-groups of different importance have been identified and their sensitivity to a range of selective group discarding percentages has been evaluated. The subdivision of the coded bit-stream in four groups is summarized in Table 2.

Mode	Number of bits	Bit division in groups			
		Group number			
		1	2	3	4
7	160	24	39	39	57
6	120	24	45	51	-
5	72	30	45	-	-
4	125	30	45	-	51
3	85	30	-	-	57

Table 2 Unequal bit division

N	*k*	*t*
63	57	1
63	51	2
63	45	3
63	39	4
63	30	6
63	24	7

Table 3 Error Correction Capabilities

The chosen group sizes permit to adopt, for error detection and correction, the set of Bose-Chaudhury-Hocquenghem(BCH) block codecs with parameters (63, *k*, *t*), where $N = 63$ represents the length of the block, *k* is the number of actual user bits in a block, and *t* is the number of errors that can be corrected in one block. The parameters of the adopted BCH codes are summarized in Table 3.

From the results presented in [5] for partial concealment, the different frame loss percentages tolerated by the different bit-groups with an acceptable quality degradation are 1% for groups 1 and 2, and 5% for groups 3 and 4.

4 CIRCUIT SWITCHING OPERATION

Two existing low-tier radio air interface protocols, namely the Digital European Cordless Telephone (DECT) and the Personal Access Communication System (PACS), provide a raw bit rate of 32 kbit/s for supporting circuit switched speech communication [6]. Both systems employ a Time Division Multiplexing (TDM) scheme. DECT operates in Time Division Duplexing (TDD) mode, while PACS has a TDD mode and a Frequency Division Duplexing (FDD) mode. A DECT frame is 10 ms in duration. Each frame includes 12 slots for the up-link communication and 12 slots for the down-link communication. A slot, which carries 320 bits for user information transmission, can be sub-divided into two half-slots, each of which can carry 160 user bits. A PACS frame is 2.5 ms long. Each frame includes 8 slots, 7 of which can be employed for information transmission. A slot carries 80 bits of user data. PACS supports sub-rate channels of 16 kbit/s and 8 kbit/s, by using one slot every two or every four frames respectively.

In Section 3 we have shown that the different modes of the used MAVT encoder require at most 32 kbit/s for a qualitatively acceptable wide-band speech encoding, while narrow-band speech encoding requires at most 16 kbit/s. Thus, DECT half-slot channels and PACS 16 kbit/s sub-rate channels are adequate for transmitting uncoded narrow-band speech, while DECT full-slot and PACS 32 kbit/s channels can be used for transmitting uncoded wide-band speech. Moreover, the channel encoding scheme proposed in Section 3 can be easily accommodated both in DECT and in PACS frame structure, using the full rate channels, permitting a trade-off between transmission quality and transmission reliability.

Error Probability

To assess the range of applicability of the proposed encoding scheme, it is necessary to determine the error probability, ε, that is the probability that a block correct reception fails due to excessive interference and noise power at the receiver input. Using a BCH code with parameters (N, k, t), reception fails when the received block does not lie within distance t from the transmitted codeword [7]. The methodology for evaluating the error probability is outlined in [5]. The modulation scheme adopted for the analysis is a BPSK coherent modulation scheme. Slow multipath fading is assumed, that is the mobile terminal is assumed to move very slowly so that the signal-to-noise ratio and the signal-to-interference ratio for all N bits within a block or codeword are the same.

We suppose that errors are due to interference. A microcellular mobile radio system, with cluster sizes $C = 3$ and $C = 7$, is considered. The system is statistically modeled

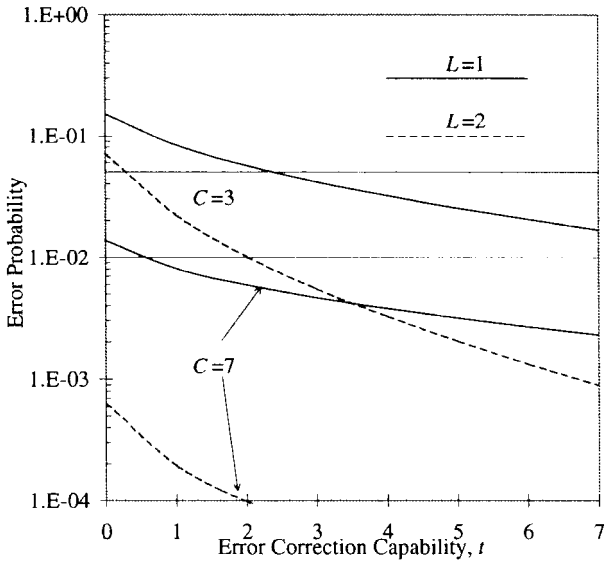

Figure 2 Error Probability vs. Error Correction Capability, fading only

by one Rician-distributed desired signal and several uncorrelated Rayleigh plus log-normally shadowed interfering signals, propagating according to a dual path loss law with a turning point [8]. The received area mean power $\bar{\bar{p}}_{rx}$ is given by:

$$\bar{\bar{p}}_{rx} = \bar{\bar{p}}_{tx} r^{-a} (1 + r/g)^{-b} \tag{1}$$

where $\bar{\bar{p}}_{tx}$ represents the transmitted area-mean power and r the normalized propagation distance between transmitter and receiver (normalized with respect to the cell radius). The exponent a is the propagation loss exponent for short distances, while the exponent b accounts for the additional propagation loss exponent for distances beyond the turning point g of the attenuation curve. In the paper, the following values are used: $a = 2$, $b = 2$, $g = 0.67$. We suppose to have a symmetrical microcellular system, with 6 active interferers. The composite p.d.f. for the total interference sum is determined according to the methodology proposed by Prasad and Kegel [9]. Examples with pure Rayleigh faded and combined faded and shadowed interferers are evaluated. For the useful signal, a Rice factor $K = 4$ (6 dB) is used. Performances with L diversity branches are evaluated, that is assuming L independent, identically distributed useful signals, all having a Rician p.d.f. The results are presented in Figs. 2 and 3. In circuit switching applications, errors cause frame erasures. The allowed erasure thresholds are shown in the figures.

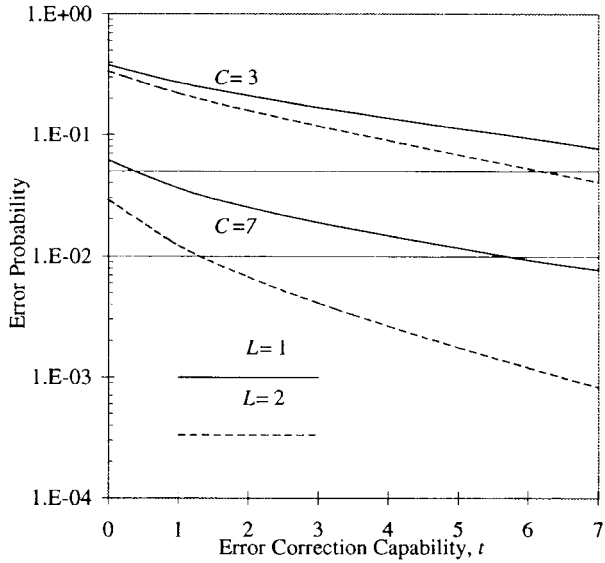

Figure 3 Error Probability vs. Error Correction Capability, fading and shadowing, σ_s = 6 dB

From Tables 2 and 3 we observe that, with the selected codes, groups 1 and 2 can use a correction capability $3 \leq t \leq 7$, while groups 3 and 4 can use a correction capability $1 \leq t \leq 4$. From these values and from Figs. 2 and 3 we derive that the erasure thresholds can be met, assuming $C = 7$, in almost all channel conditions, but $\sigma_s = 6$ dB and $L = 1$ (i.e. with shadowing and no diversity), while, for $C = 3$ the same limiting conditions can be satisfied only with no shadowing and using diversity ($L = 2$). From Figs. 2 and 3 we also can infer that, for the transmission of not channel encoded wide-band speech, the erasure requirements can be satisfied only assuming $C = 7$, $L = 2$ and no shadowing.

5 PACKET SWITCHING OPERATION

Packet switching algorithms, performing statistical multiplexing, may lead to greater efficiency than basic TDM, by exploiting the bursty character of speech traffic. Moreover, centralized access techniques take advantage of the small propagation delay of the micro cellular environment [10], [11]. These techniques assign a central role to the base station that schedules the transmissions of the speech packets. The base station reacts to transmission errors by rescheduling the transmission of failed packets.

The maximum number of users, $U_{\max}$, that can be accommodated in the system, fulfilling the tolerable discarding thresholds, depend on the speaker statistic, on the error probability, ε, and on the multiplexing technique. A Markov model has been proposed in [12], for representing a typical VBR-CELP encoded speech signal. The state probabilities, p_i, are summarized in Table 4.

First, to determine the achievable multiplexing gain, that is the number of users per slot which can be accommodated in the system, we do not take into account the multiple access protocol. The adopted methodology is outlined in Appendix.

Mode	1	2	3	4	5	6	7
Prob.	0.36	0.05	0.06	0.26	0.06	0.14	0.07

Table 4 State probabilities of coding modes

We hypothesize that bits of different priority are grouped into different packets. At the beginning of a frame the transmission of high priority packets is requested. Errors are recovered by re-transmissions. Then, the process is repeated for the other priorities. We suppose that at the end of the frame all the non transmitted packets have to be dropped. Three different situations are taken into consideration.

1. Single packet per frame. Silences are suppressed (modes 1 and 2), but the VBR character of the code is not exploited.
2. Up to two minipackets per frame (the slots are subdivided in half slots, as in DECT system). The VBR character is partly exploited, as modes 3 and 5 generate only one half packet per frame.
3. Up to four minipackets per frame (a full packet requires 4 minislots, as in PACS system). The VBR character is fully exploited, and the number of minipackets can be derived from Table 2 (the higher priority packets are those corresponding to groups 1 and 2, whereas the lower priority packets are those corresponding to groups 3 and 4).

The number of users which can be accommodated in the system is shown in Fig. 4. Although the actual numbers are integers, continuous fitting lines are added, for clarity purposes. From the figure we derive that a system with 12 slots per frame (DECT equivalent) could accept up to 16 users in single packet mode (1 packet requires 1 full slot), and up to 19 users in the two-minipacket mode (1 minipacket uses 1 half slot), while a system with 7 slots per frame (PACS equivalent, excluding the broadcast channel) could accept up to 12 users in the four-minipacket mode (1 minipacket uses 1 burst).

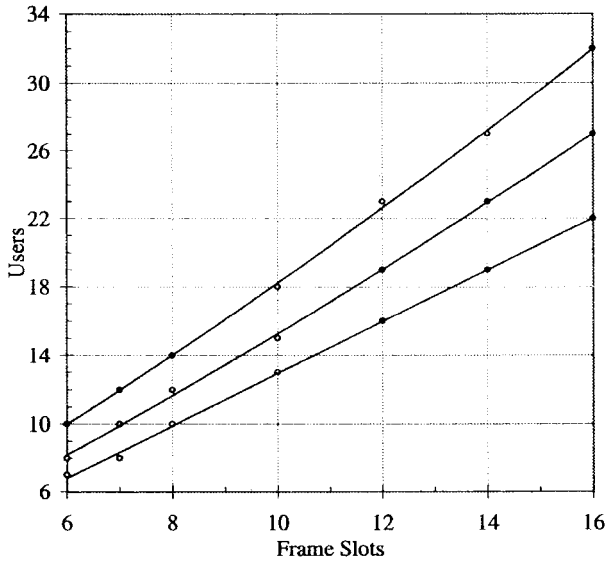

Figure 4 Maximum active users vs. frame slots; error-free channel; single mini-packet(bottom), 2 minipackets (middle), 4 minipackets (top)

Effects of channel errors

The results shown in Fig. 4 represent an upper bound on system performances. Actual performances depend on channel errors which reduce the number of useful transmissions. For our analysis we use a discrete memoryless error model. The methodology for determining the packet error probability, ε, is shown in [13]. Error probability depends on traffic in co-channel cells, so it is assumed that all the surrounding cells are equally loaded. Rayleigh fading is assumed for all the interferers, and the joint interference power is approximated by a gamma distribution. Up to 18 interfering users are considered, belonging to three first groups of surrounding cells. Rice fading is assumed for the useful signal. The average error probability is determined by assuming uniform placement of both the useful and the interfering signals. A transmission is considered successful if the Signal to Interference Ratio (SIR) exceeds a given threshold, Z_0, namely the capture ratio. The chosen value is $Z_0 = 10$ dB. This value represents a reasonable compromise between the capture ratio of high priority packets, with strong error protection, and low priority packets. However, the chosen value could be considered optimistic when channel encoding is not used (wide-band services). The channel model used in Section 4 is assumed, that is $a = 2$, $b = 2$, $g = 0.67$, $K = 4$. Both cluster sizes $C = 3$ and $C = 7$ are considered. For this

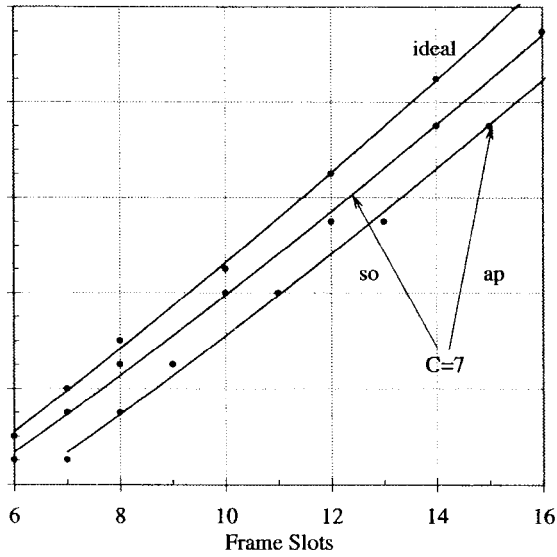

Figure 5 Maximum active users vs. frame slots; 2 minipackets; cluster size: $C = 7$; speech only (so) and with access protocol (ap)

analysis, the two-minipacket mode is considered. The results are presented in Figs. 5 and 6.

In the figures, the *speech only* curves do not take into account the access protocol, while the *access protocol* curves are evaluated assuming to reserve one slot per frame to the access protocol. The performances of the system with $C = 7$ are close to ideal performances. However, the normalized number of users, $U_{\max}/C$, is higher for the system with $C = 3$. To assess the correctness of the analysis, a simulation has been carried out, assuming that the access protocol is used to inform the base station that a user enters an active mode (silence to talkspurt transition). Then, the base station is informed about all the state transitions between active modes by using some spare capacity of the information channel. Thus, the signaling traffic due to the multiple access is rather low, so that all access protocols perform reasonably well. The simulation has been performed by using a Centralized Collision Resolution Algorithm (C-CRA) [11]. The simulation results are in good agreement with the analysis results.

6 DISCUSSION AND CONCLUSIONS

We have investigated the interaction of source and channel coding, and of coding and networking techniques. An object oriented (multimode), VBR, and embedded source

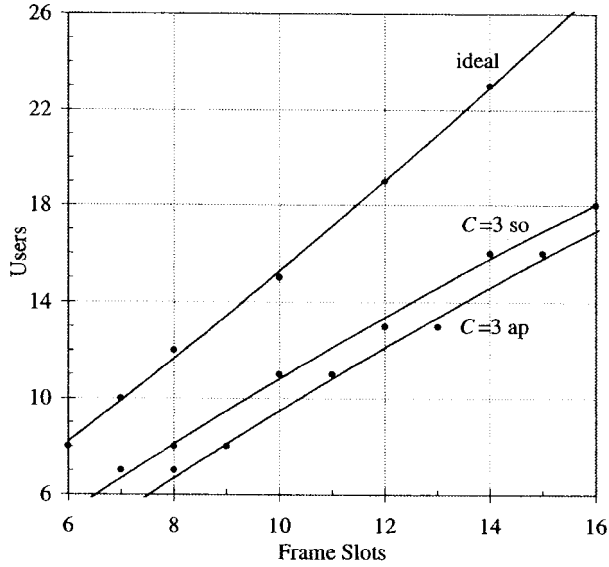

Figure 6 Maximum active users vs. frame slots; 2 minipackets; cluster size: $C = 3$; speech only (so) and with access protocol (ap)

coding scheme has been utilized. Both, narrow band and wide band audio services have been considered. First, circuit switching operation has been analyzed. Two existing low-tier radio air interface protocols, namely DECT and PACS have been considered. It has been shown that DECT half-slot channels and PACS 16 kbit/s sub-rate channels are adequate for transmitting uncoded narrow-band speech, while DECT full-slot and PACS 32 kbit/s channels can be used for transmitting uncoded wide-band speech. We have proposed a channel encoding scheme that can be easily accommodated both in DECT and in PACS frame structure, using the full rate channels. The choice between the use of uncoded wide-band transmission and encoded narrow-band transmission permits a trade-off between transmission quality and transmission reliability. By evaluating the error probability for a micro-cellular system, we have shown that uncoded wide-band could be feasible only under some special conditions (e.g. no shadowing and adoption of diversity techniques), while, also for encoded narrow-band services a sufficiently high reuse factor is required.

Then, centralized packet switching techniques have been investigated. We have shown that system capacity can be increased, by exploiting the variable rate character of the coding technique. The presented results apply to encoded transmission of narrow-band services, because the chosen capture threshold is slightly optimistic for uncoded wide-band services.

We recall that our results rely on to the memoryless channel model assumption. The study of systems with channel memory is for further work.

Acknowledgments

This work has been partially supported by MURST Italy, “ex quota 40%”.
The authors wish to thank CSELT S.p.A. for the provision of the FVR-CELP and the MAVT encoders.

Appendix: Dropping Probability Evaluation

The dropping probability, d_i of packets of priority i, can be determined as follows. Let us introduce the following quantities.

$P\ (N_s = n)$: Probability of transmitting successfully n packets in s slots.

$P\,(S_n = s)$: Probability that successful transmission of n packets requires s slots.

N_{ij} : Packets of priority i generated by a user in mode j.

P_i : Average number of generated packets having priority i.

T_i : average number of packets successfully transmitted having priority i.

A Markov model is used for representing the single user mode evolution, being p_j the probability of mode j. We name system state the number, u_j, of users in mode j; $U = \sum_{j=1}^{7} u_j$ represents the total number of users in the system. We can now define the following quantities.

$\pi(u_j) = U! \prod_{j=1}^{7} \frac{p_j^{u_j}}{u_j!}$: state probability.

$N_i(u_j) = \sum_{j=1}^{7} u_j N_{ij}$: Packets of priority i.

$M_i(u_j) = \sum_{h=1}^{i-1} N_h(u_j)$: Packets of priority greater than priority i.

We have:

$$P\,(N_s = n) = \begin{cases} 1 & n = 0, s = 0 \\ \varepsilon^s & n = 0, s > 0 \\ \varepsilon P\,(N_{s-1} = n) + (1-\varepsilon)\,P\,(N_{s-1} = n-1) & n > 0, s > 0 \end{cases} \tag{A.1}$$

$$P(S_n = s) \begin{cases} 1 & n = 0, s = 0 \\ (1-\varepsilon)\, P(N_{s-1} = n-1) & n > 0, s > 0 \\ 0 & \text{otherwise} \end{cases} . \qquad \text{(A.2)}$$

Suppose that S indicates the number of slots per frame. The average number of successful transmissions of packets having priority i is given by:

$$T_i = \sum_{u_j} \pi(u_j) \sum_{s=M_i(u_j)}^{S} P\left(S_{M_i(u_j)} = s\right) \qquad \text{(A.3)}$$
$$\left(\sum_{n=1}^{N_i(u_j)-1} nP\left(N_{S-s} = n\right) + N_i(u_j) \sum_{r=N_i(u_j)}^{S-s} P\left(S_{N_i(u_j)} = r\right) \right).$$

The average number of packets having priority i is given by:

$$P_i = U \sum_{j=1}^{7} p_j N_{ij}. \qquad \text{(A.4)}$$

The dropping probability, d_i , is given by:

$$d_i = (P_i - T_i) / P_i. \qquad \text{(A.5)}$$

References

[1] L. Hanzo and J. P. Woodard, "An Intelligent Multimode Voice Communications System for Indoor Communications", IEEE Transactions on Vehicular Technology, Vol. 44, N. 4, November 1995, pp. 735 - 748.

[2] L. Cellario and D. Sereno, "CELP Coding at Variable Rate", ETT., Vol. 5, No. 5, Sept. 1994, pp. 603-613.

[3] S. Dimolitsas, "Objective speech distortion measures and their relevance to speech quality assessments", IEE Proceedings-I, Vol. 136, No. 5, October 1989, pp. 317-324.

[4] S. Dimolitsas, ”Subjective quality quantification of digital voice communication systems”, IEE Proc.-I, Vol. 138, No. 6, Dec. 1991, pp. 585-595.

[5] F. Babich, F. Vatta, ”A Multimode Voice Communications System with Source-Matched Error Protection for Mobile Communications”, 1997 IEEE Vehicular Technology Conference (VTC ’97), Phoenix, AZ, May 5-7, 1997.

[6] A.R. Noerpel, Y. Lin, H. Sherry, ”PACS: Personal Access Communication System - A Tutorial”, IEEE Personal Communications, Vol. 3, No. 3, June 1996, pp. 32-43.

[7] J.-P.M.G. Linnartz, A.J.T Jong and R. Prasad, ”Effect of coding in digital microcellular personal communication systems with co-channel interference, fading, shadowing and noise”, IEEE Vol. SAC-11, No. 6, Aug. 1993, pp. 901-910.

[8] J.-P. Linnartz, ”Narrowband Land-Mobile Radio Networks”, Artech House, Inc., Boston, 1993.

[9] R. Prasad and A. Kegel, ”Improved Assessment of Interference Limits in Cellular Radio Performance”, IEEE Vol. VT-40, No. 2, May 1991, pp. 412-419.

[10] G. Bianchi, F. Borgonovo, L. Fratta, L. Musumeci, M. Zorzi, ”C-PRMA: the Centralized Packet Reservation Multiple Access for Local Wireless Communications”, in Proc. GLOBECOM 1994, S. Francisco, Nov. 27 - Dec. 1, 1995.

[11] F. Babich, ”Analysis of Frame Based Reservation Random Access Protocols for Microcellular Radio Networks”, IEEE Transactions on Vehicular Technology (to appear).

[12] ”VR-CELP speech coder: evaluation of the state model probabilities”, CSELT Internal Technical Report.

[13] F. Babich, ”Free Access Stack Algorithms for Microcellular Radio Systems”, IEEE Transactions on Vehicular Technology, submitted for publication.

This page intentionally left blank.

Multilevel Channel Assignment (MCA): A Performance Analysis

Farooq Khan, Djamal Zeghlache

Institut National des Télécommunications
9 rue Charles Fourier, 91011 Evry, France
Email: Farooq.Khan(Djamal.Zeghlache)@int-evry.fr

Abstract

A channel allocation technique to respect different Carrier to Interference (C/I) ratio constraints in wireless networks supporting multiple services is proposed and analyzed through a system supporting two services, voice and data. Since these services require different Bit Error Rate (BER) performance, different C/I values should be used to meet these requirements. To achieve this goal a cell is divided into concentric zones with the objective of trading off C/I performance of users according to their service class and geographical location. The underlying principle is to offer the minimum required C/I performance for voice users in the inner zone having more than adequate transmission quality in order to improve C/I performance for data users which require very low BER. A Markov chain model of the proposed scheme is identified and results obtained through a comprehensive computer simulation for voice call blocking probability and data access delay are provided.

1. Introduction

The next generation Personal Communications Systems (PCS) promise to provide a wide range of services to users, including high quality voice, variable rate data, full motion video, high resolution image, etc. These services requiring different Bit Error Rate (BER) and hence different Carrier-to-Interference (C/I) ratio will share a common group of channels (frequencies, slots or spreading codes). In order to guarantee each service its quality of service, we must identify resource allocation and sharing schemes capable of statistically multiplexing services with considerably different characteristics. The problem of multiplexing services with different transmission speeds dealt with in [1] is somehow re-addressed to take into account the fact that different services have different C/I (or BER) requirements. In this

paper, the problem of multiplexing these services while guaranteeing each service its required BER is addressed.

In the proposed Multilevel Channel Assignment (MCA) for two service classes, a cell is divided into two concentric zones: an inner zone and an outer ring. A similar cell structure is also used in reuse partitioning [2] (or overlay-underlay) [3] and in Channel Borrowing Without Locking (CBWL) [4]. In reuse partitioning, the available channels are split among several reuse patterns with different reuse factors. Mobile units with the best received signal quality are preferentially assigned to the group of channels having the smallest reuse factor value, while those with the poorest received signal quality will be only assigned to the group of channels having the largest reuse factor value. By using two reuse factors of 3 and 9, an increase in channel capacity of 30\% can be achieved compared to that obtained by a single reuse factor of 7 for a system with objective C/I of 17dB. CBWL, which uses only one reuse pattern, allows real-time borrowing of channels from adjacent cells without the need for channel locking in co-channel cells. Like reuse partition this scheme also provides enhanced traffic performance. These schemes are proposed for a voice only system with a single minimum C/I constraint.

The scheme proposed in this paper allows different services to operate at different C/I levels. In MCA, voice users maintain the same C/I ratio as in the other schemes. The data traffic which is more sensitive to Bit Error Rate (BER) is allowed to use channels at a higher C/I level than voice traffic. The underlying principle is to operate at the minimum required C/I level for voice users (in the inner zone) where voice users experience more than adequate transmission quality in favor of better C/I performance for data users (in the outer zone) which require very low BER. This may be achieved by allowing data users in the outer ring to use channels at a higher power than other users. However, this increased radiated power on the outer data channels may cause a significantly higher interference in the co-channel cells. Hence, these channels should be reused carefully in the co-channel cells. A thorough analysis of the C/I performance of the proposed scheme shows that the outer data channels can still be used by a certain population of voice users in the inner zone of the co-channel cells. Thus, in MCA the outer data channels are locked in the co-channel cells for all users except for the inner zone voice users

In MCA, mobile units located further away from the cell site will in general have fewer channels at their disposal. The area around the base station has some preferential treatment compared to the outer cell area. This will give to mobiles in different parts of the system different grade of service in terms of the blocking probability or experienced access delay. A cut-off priority scheme can be used in favor of voice users at the periphery of the cell where voice calls originating in the cell inner zone are denied access to the resource even if some channels are free. Data requests from the inner and outer zone wait in a buffer and are served according to FCFS discipline. Using this queuing discipline, if an outer zone data request is waiting in the buffer an arriving inner zone data call will not be allocated

a channel even if a channel is accessible to this call. This helps distribute access delay uniformly among the inner and outer zone data users.

2. System Model

2.1 Analysis of Co-Channel Interference

The cellular mobile system studied in this paper consists of a finite number of cells. Base stations use omni-directional antennas and are located at the center of the cells. Mobile Stations (MS) are assumed to be uniformly distributed over the cell area. The interference caused by adjacent channels is assumed to be much smaller than co-channel interference and is neglected. To simplify analysis, the effects of fading are not included either.

Since a cell in MCA is divided into two concentric zones: an inner cell and an outer ring, we identify 4 different types of new call originations for the two service system:

- I_v - type calls: voice calls arising in the inner cell
- I_d - type calls: data calls arising in the inner cell
- O_v - type calls: voice calls arising in the outer ring
- O_d - type calls: data calls arising in the outer ring

Let R be the radius of a cell and D the distance between two co-channel cells. Then the C/I of a cellular system can be approximated by [3]:

$$\frac{C}{I} = \frac{1}{M}\left(\frac{D}{R}\right)^n \qquad (1)$$

Where M is the number of co-channel interfering cells and n is a path loss exponent that ranges between 2 and 4 in urban cellular systems. M is equal to 6 for the case of hexagonal cell layout if the interference is considered from only first tier cells as shown in Figure 1. We take $n = 4$ in our study. Note that we are using completely inadequate propagation models to conduct the study. A thorough effort requires the use of a true reference scenario with geographical data bases and appropriate propagation models (Okumura-Hata, Walfish-Ikegami, Cost 231, Wiart, etc.) and the support of a coverage prediction tool to obtain true C/I maps. However, we are only interested in making an initial assessment to determine whether or not the MCA approach is feasible.

As described in the previous section, the underlying principle of the MCA scheme is to degrade the C/I performance for the I_v -type calls in favor of O_d -type calls. This may be achieved by allowing the O_d -type calls to use channels at a higher power. For example 6dB higher than other types of calls and locking these channels in all co-channel cells for all types of calls except for I_v -type calls. The

worst case may be that a channel is used by an I_v -type call in cell *i* and by O_d -type calls in the co-channel cells of cell *i*. We see that this worst case situation degrades the *C*/*I* performance for the I_v -type calls by 6dB when $R=\sqrt{R_1}$ (see Figure l), where R_1 is the radius of the inner cell. Even with this degradation I_v -type calls *C*/*I* still experiences a *C*/*I* level equal to the O_v -type calls. The interested reader is referred to [5] for a detailed analysis of *C*/*I* in the MCA system.

The fraction of area in which the data users will use higher transmission power is denoted as *p*. For the homogeneous case *p* is given by $1-\left(R_1/R\right)^2$.

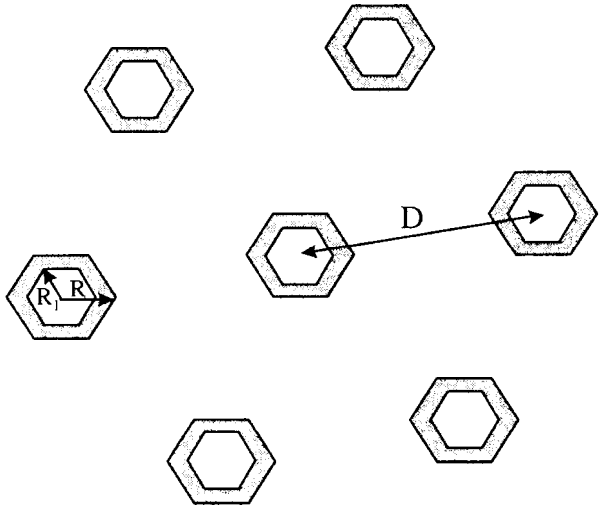

Figure 1. Co-channel interference with six interferers in hexagonal cell layout

2.2 Traffic Performance of the Proposed Scheme

2.2.1 Assumptions

This section is devoted to MCA performance analysis.

1. For simplicity we limit our analysis to the homogeneous case but the analysis can easily be extended to non-homogeneous cases. In the homogeneous case, calls originate uniformly throughout the service area.
2. The speech call origination rate in the inner and outer ring is a Poisson process with parameters λ_{vi} and λ_{vo} respectively.
3. The speech call duration is an exponentially distributed random variable with parameter $1/\mu_v$.
4. The data message arrival process is also a Poisson process with parameters λ_{di} and λ_{do}.

5. The data message length is an exponentially distributed random variable with parameter $1/\mu_d$.

2.2.2 Single-cell Performance

To obtain a simplified solution we suppose that $1/\mu_v = 1/\mu_d = 1/\mu$ and let $\rho = \frac{\lambda_v + \lambda_d}{\mu}$ and $\rho_2 = \frac{\lambda_d}{\mu}$ where $\lambda_v = \lambda_{vi} + \lambda_{vo}$ and $\lambda_d = \lambda_{di} + \lambda_{do}$. Then the behavior of a single cell in MCA can be modeled by a mixed loss and delay queuing system [6] as shown in Figure 2.

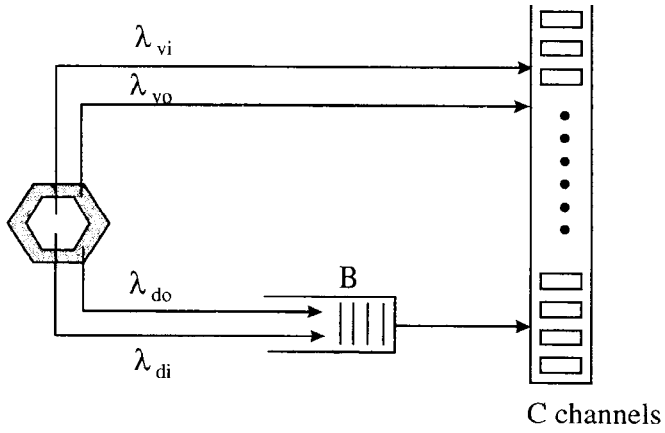

Figure 2. Queuing system diagram of the proposed scheme for a single cell

The blocking probability for voice users in such a system can be written as:

$$P_B = \frac{LE_L(\rho)}{L - \rho_2 + \rho_2 E_L(\rho)} \tag{2}$$

where $E_L(\rho)$ is the *B* formula given as:

$$E_L(a) = \frac{\frac{a^L}{L!}}{\sum_{i=1}^{L} \frac{a^i}{i!}} \tag{3}$$

The mean access delay for data users

$$D = \frac{P_B}{\mu(L - \rho_2)} \tag{4}$$

2.2.3 Multi-cell Performance

For simplicity of analysis, the multi-cell performance analysis is conducted under a pure loss system assumption. New voice and data calls that do not find any free

channels on arrival are blocked and cleared from the system. Using this assumption, a cell in MCA can be in one of a finite number of states. A state is identified by a vector

$$N = (n_{i1}, n_{i2}, n_{i3}, \ldots, n_{iL}) \quad (5)$$

Where n_{ij} *(i = 1,2, . . . , K and j= 1,2, . . . , L)* denotes the *j th* channel in the *i th* cell. We identify the status of the channels in the *i th* cell as:

- nij = 0, channel j is free
- nij = 1, channel j is occupied by an Iv-type call
- nij = 2, channel j is occupied by an Id-type call
- nij = 3, channel j is occupied by an Ov-type call
- nij = 4, channel j is occupied by an Od-type call

Then the total number of channels occupied in cell *i* is given by:

$$L_i = \sum_{j=0}^{L} S_{ij} \quad (6)$$

where

$$\begin{cases} S_{ij} = 1 \; if \; n_{ij} = 1,2,3 \; or \; 4 \\ S_{ij} = 1 \quad otherwise \end{cases} \quad (7)$$

and (i=1, 2, 3, ... , K).

In MCA, the states of the cells using the same channel set are coupled and this renders the problem difficult if not impossible to solve analytically. The results in this paper are given by forming a comprehensive simulation model of the system under study which is described in the next section.

Let $n_{k,j}$ denote the status of the *j th* channel in the co-channel cells of cell *i* and let $C_{i,j} = max\{n_{k,j}\}$. Then the principle of the proposed channel allocation technique for the case of blocked calls cleared can be described as follows:

1. A new I_v -type call originating in cell *i* will be accommodated in the system if $L\text{-}L_i > 0$ otherwise the call is rejected.
2. A new O_v-type call originating in cell *i* will be allocated a free channel *j* if $n_{i,j} = 0$ and $C_{ij} \neq 4$ otherwise the call is denied access to the resource.
3. A new Id-type call originating in cell *i* will be allocated a free channel *j* if $n_{i,j} = 0$ and $C_{ij} \neq 4$ otherwise the call is denied access to the resource.
4. A new O_d-type call originating in cell *i* will be allocated a free channel *j* if $n_{i,j} = 0$ and $C_{ij} = 0$ or *1* otherwise the call is denied access to the resource.

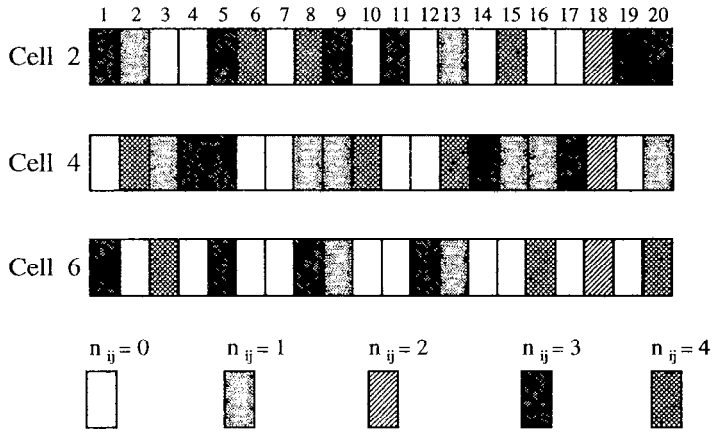

Figure 3. A snapshot of channel status in MCA

A snapshot of channel status in MCA for a one dimensional cell array (Figure 3) with a reuse factor of 2 is shown in Figure 4. For example channel 1 in cell 4 can not be allocated to an O_d -type call since it is being used by O_v -type calls in cell 2 and cell 6 (co-channel cells of cell 4).

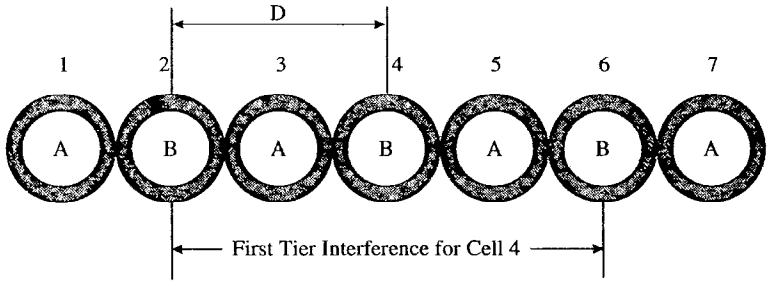

Figure 4. One-dimensional cellular system model

In MCA, channel rearrangements can be used in favor of users that have no access to all free channels. With channel rearrangements, a call for which there is no accessible channel but there are free channels may still be accepted. This is achieved by rearranging the currently occupied channels in the given cell and thus making a channel accessible to the arriving call. We shall call this type of rearrangement as local rearrangement and the corresponding MCA as MCA/LR. In MCA, if a channel cannot be made available to an arriving call by local rearrangement, it is possible to do this by rearranging the occupied channels in the co-channel cells. This type of channel rearrangement will be called Global Rearrangement (GR) and the MCA using this rearrangement in addition to LR as MCA/GR. This idea is similar to Maximum Packing (MP). MP accepts a new call if there is any possible reassignment of channels to calls in progress which results in freeing a channel

within the interference region of the new call's cell. This policy requires complete knowledge of all existing channel assignments in the entire system, and may potentially reassign all existing calls.

3. Simulation Model

In this section we consider a one dimensional cell array as used in major roads and highways. In the 1-D cellular system the main source of interference is from the two co-channel cells comprising the first tier around a target cell as shown in Figure 4. Hence, interference from outside the first tier is neglected. Moreover, since simulations are carried out for a TDMA system we suppose that the base stations and users share global synchronization i.e. cell-to-cell time slot synchronization.

System Variables		
Variable	*Symbol*	*Value*
Number of Channels per cell	L	72
Mean voice call duration	T_c	100 secs
Mean data message duration	T_d	100 secs
Outer voice guard channels	C_0	0
Number of cells	M	20

Simulations are carried out for a UMTS microcellular environment with 72 TDMA slots within a 5 msec frame, implemented through a carrier bit rate of 1.8 Mbits/sec. One carrier is allocated per cell which provides 72 channels per cell.

4. Results and Discussion

Figure 5 shows the behavior of call blocking as the normalized offered load is varied from 0.6 to 0.9. The curve marked MP gives the single cell performance and serves as the reference case for multi-cell performance. Since in the MP case there is no constraint on the outer voice users, the blocking for these users is the same as for inner voice users. The two curves marked MCA/LR give the results for the multi-cell case when local rearrangement of channels is permitted. The blocking probability for I_v-type calls is lower than O_v-type calls because the first type of calls have more channels available than second type calls. It can be observed that the difference between the inner and outer call blocking probability is more pronounced when no channel rearrangement is used. This unfairness can be eliminated by using cutoff priority in favor of outer voice users.

Figure 6 presents the mean access delay for data users as a function of the normalized offered load. As expected, the data users mean access delay is minimum for the MP case. It can also be noted that mean access delay for data users

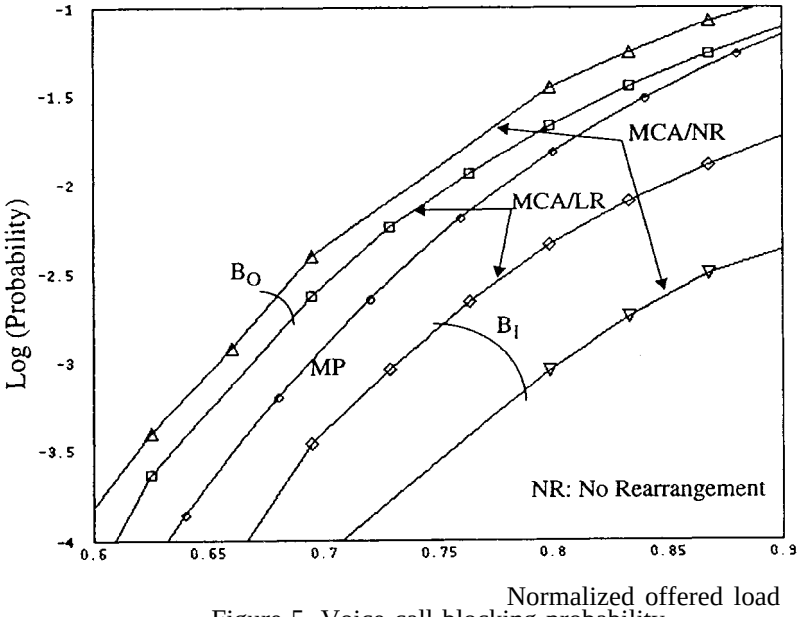

Figure 5. Voice call blocking probability

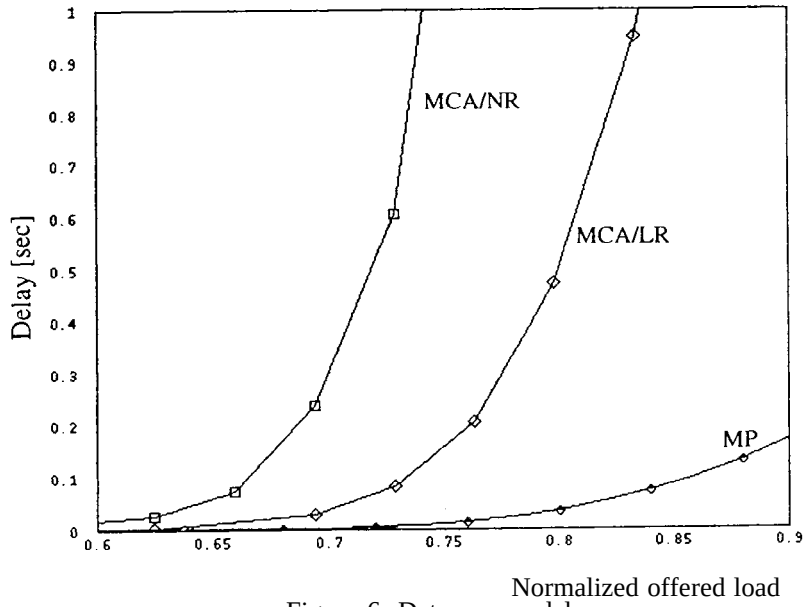

Figure 6. Data access delay

with local rearrangement is much lower than the case with no rearrangement. This improvement for MCA/LR is achieved at the expense of some additional signaling load for rearranging channels.

The comparison of carried traffic for MP, MCA/LR and MCA/NR is depicted in Figure 7. We note that up to 0.75 of normalized offered load, the carried traffic for the three schemes is approximately the same. As offered load is increased beyond 0.75 there is a small difference in the carried traffic for the three schemes. This small difference in the carried traffic for MCA/LR and MCA/NR with respect to MP is explained by higher data access delays for these two schemes.

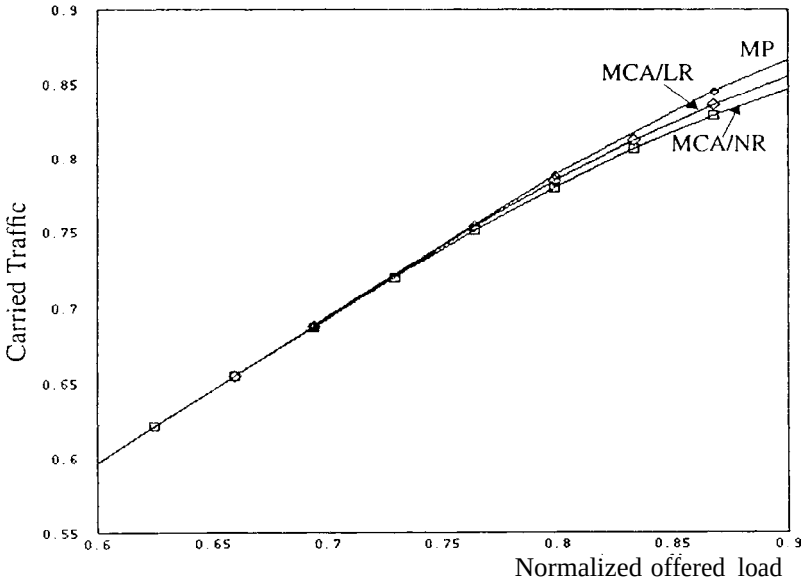

Figure 7. Carried traffic

5. Concluding Remarks

We proposed a channel allocation scheme for wireless networks supporting multiple services. The proposed MCA scheme can be used to guarantee different BER for different services. We observed that the constraints induced on channel allocation results in higher data delays without significant reduction in the system capacity. We expect that the data delays in MCA can be reduced by using efficient queuing disciplines for data users.

References

[1] F. Khan and D. Zeghlache, "Priority-Based Multiple Access for Statistical Multiplexing of Multiple Services in Wireless PCS" *IEEE ICUPC'96*, pp. 17-21.

[2] S. W. Halpern, “Reuse Partitioning in Cellular Systems” *IEEE VTC'83*, pp. 322-327.

[3] W. C. Y. Lee, *Mobile Communications Design Fundamentals*, Wiley series in telecommunications, 1993.

[4] Long-Rong Hu and S.S. Rappaport, "CBWL: A New Channel Assignment and Sharing method for Cellular Communication Systems", IEEE trans. on vehicular technology}, vol. 43, no. 2, pp. 313-322, May 1994.

[5] F. Khan, "Multiple Access and Resource Allocation in Wireless Multimedia Networks" *Ph. D dissertation*, University of Versailles Saint-Quentin en Yvelines, France, February, 1997.

[6] H.Akimaru and K. Kawashima, *Teletraffic: Theory and Applications*, Springer-Verlag, 1993.

This page intentionally left blank.

ESTIMATING THE CELL RADIUS FROM SIGNAL STRENGTH MEASUREMENTS

Pete Bernardin and Meng Yee
NORTEL Wireless Engineering Services
Richardson, TX, 75083-3805

Thomas Ellis
Wireless Software Design & Consulting
13447 Forestway Drive, Dallas, TX, 75240

ABSTRACT : *A robust method for determining the boundaries of cells and the associated reliability of the RF coverage within these boundaries is presented. The procedure accurately determines the effective cell radius using a linear regression of RF signal strength samples. The accuracy of this estimate is quantified both as a radius uncertainty (e.g., ± 100 meters) and as a cell coverage reliability (i.e., area/edge) through 1) simulation, 2) analysis of real data, and 3) theoretical analysis. It is shown that if the estimate of the cell radius meets the desired accuracy, then the corresponding estimates of coverage reliability (both area and edge) are more than sufficiently accurate. It is discovered that estimating the cell radius is a much more critical step in determining the quality of RF coverage than the more common practice of simply estimating the area reliability. In addition, a formula for estimating area reliabiliy is given and shown to be more accurate than the estimate of the cell radius. The validation method presented here is particularly useful in wireless planning since it accurately determines the effective geographic extent of reliable RF coverage. It is recommended that radio survey analyses incorporate this cell radius estimate as part of the coverage validation process.*

I INTRODUCTION

The two most commonly used measures of the reliability of RF coverage are 1) cell edge reliability and 2) cell area reliability. Cell edge reliability refers to the probability that the RF signal strength measured on a circular contour at the cell edge will meet or exceed a desired quality threshold (e.g., -90 dBm). Whereas, cell area reliability is the probability that RF signal will meet or exceed the quality threshold after integrating the contour probability over the entire area of the cell (i.e., across all of the contours of the cell, including the cell edge). D.O. Reudink showed that, for a given propagation environment, cell edge reliability and cell area reliability are deterministically related [2] (see also section IV of this paper). Because of this relationship, estimating the distance to the cell edge can be shown to be theoretically equivalent to determining the reliability of the signal strength within the cell (e.g., see equation (5) and equation (6)). In this study we describe a new measure of RF reliability that has previously not been reported in other wireless investigations. We call this coverage criterion "cell radius inaccuracy," ΔR. We have found this criterion to be very useful in answering the following two (related) questions:

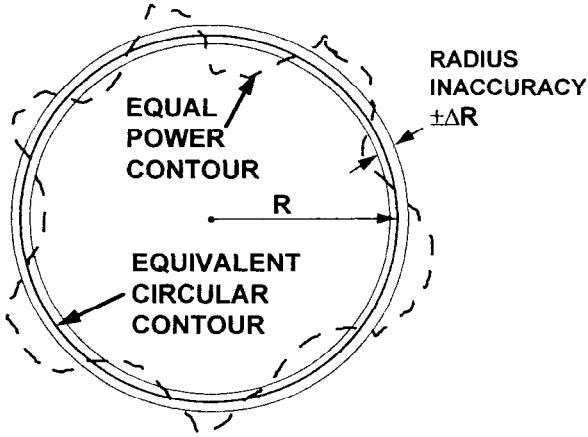

Figure 1. The measurement approach computes the best circular approximation to the equal power contour. The effective radius, *R*, of the cell is measured and the accuracy quantified in terms of a radius inaccuracy ring, $\pm\Delta R$. The average signal strength on the circular contour is equal to the signal strength of the equal power contour.

1) How many signal strength measurements are needed to accurately estimate the spatial extent of reliable coverage?

2) How do we best estimate the coverage reliability of single cells with a finite number of signal strength measurements?

In answering the first question, the equivalent circular contour (i.e., the effective radius, *R*) of the cell is estimated as shown in Figure 1. The relationship between the inaccuracy (ΔR) of this radius estimate and the amount of lognormal fading, σ, in each cell is empirically derived as a function of the number of signal strength measurements, *N* (see equation (11)).

Regarding the second question, perhaps the most important finding of this study is that it is the accuracy of the cell radius estimate (i.e., ΔR), not the accuracy of the area reliability estimate that is the limiting factor in determining the quality of RF coverage. The relationship between cell radius inaccuracy and area reliability is also discussed.

Typically, cell radius estimation and area reliability analyses are not considered together in propagation optimization. It is found that these two problems cannot be considered independently, and the consequence of doing so can lead to inaccurate estimates of RF coverage.

Many vendors of RF tools already use regression to determine the best linear approximation to the median path loss for the purpose of tuning RF prediction models. In this paper we investigate the advantages of also measuring the lognormal fading within each cell to more precisely determine the radius of reliable cellular coverage. These measurements are used to compute a fade margin for each cell which is then incorporated in the estimation of the cell's radius. Thus, the cell radius is defined explicitly in terms of the desired quality of coverage.

It is recommended that, in addition to area reliability, future wireless validations also consider cell radius inaccuracy.

II OVERALL APPROACH

The proposed method estimates the best circular boundary that matches the cell edge at the desired area reliability, as illustrated in Figure 1. It should be emphasized that this approach does not in any way require that the true cell edge be circular. Rather, even the most irregular cell edge can be fitted with a circle such that the average power along the circumference is equal to the power of the true cell edge. This circle encloses the area over which the RF signal meets or exceeds the desired area reliability (e.g., over 90% of the area, the signal power is above -90 dBm). It is the radius of this fitted circle that is estimated. Thus, this radius can be considered the "effective radius" of the cell and is well defined for any cell, circular or otherwise.

The accuracy of the cell radius estimate is quantified in terms of a radius inaccuracy ring, $\pm\Delta R$, also shown in Figure 1, where ΔR is measured in the same units of distance as the cell radius, R (e.g., km). The width of this ring depends mostly on the number of signal strength samples in the regression, and also upon the amount of lognormal fading in the cell.

The method for estimating area reliability involves determining the propagation parameters of individual cells and using this information in conjunction with Reudink's analysis (see equation (6)) to estimate the area reliability. The propagation path is approximated by a two-parameter model similar to Hata [1]. A fade margin based on the actual signal variation within each cell is calculated to ensure the desired cell edge reliability. It is shown that this area reliability estimation technique requires much fewer signal strength measurements than the cell radius estimation approach.

The proposed method can be used to quickly determine the validity of drive test data. Given enough measurements, simulations show that this technique can be made almost arbitrarily exact. It is recommended that this method be included as part of the pre-build validation procedure for any wireless technology (TDMA, AMPS, CDMA, etc.).

III APPROACH FOR ESTIMATING THE CELL RADIUS

The proposed approach for estimating the cell radius is graphically summarized in Figure 2. The measurement method is based on a two parameter propagation model similar to the prediction formulas of Hata

$$P_r = P_t - P_L = P_t \quad - \quad A - B\log_{10} r \tag{1}$$

where P_r is the received power (dBm), P_t is the transmitted power (EIRP) of the base station plus the receiver gain (e.g., P_t = 50 dBm EIRP+G_r), P_L is the path loss (dB), r is the range (km) from the base station, and A and B are the unknown constants to be estimated from the RF data via linear regression [1]. A fade margin based on the actual signal variation within each cell is calculated to ensure the desired cell edge reliability.

Because of the similarity to Hata's model, it is important to clarify that the method does not incorporate Hata's coefficients. Instead, the salient propagation parameters are estimated from the data since the major goal in this study is RF validation, not RF prediction.

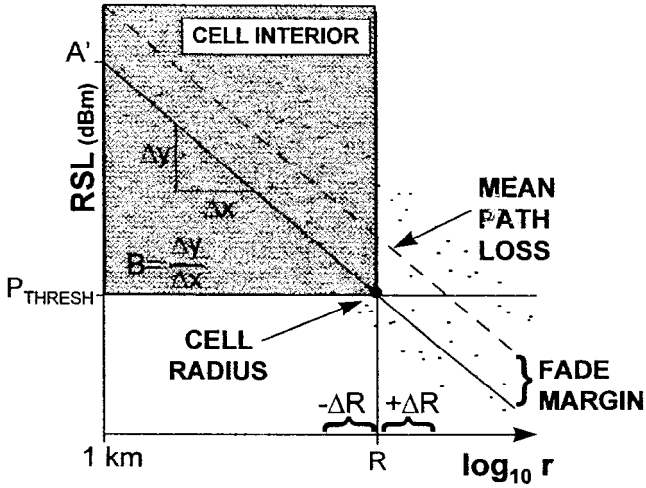

Figure 2. The graphical approach to estimating the cell radius to within ±Δ*R*. The received signal strength level (RSL) is plotted versus the range from the base station to each measurement The mean path loss is computed via linear regression and offset by the fade margin. The cell radius is defined in terms of the desired coverage reliability as the point where the faded line crosses the reliability threshold, P_{THRESH}.

The interior of each cell is divided into approximately 5000 bins which are uniformly sampled both in range and azimuth. The signal strength measurements in each bin are averaged to produce a single (average power) value per bin [10]. The range is then computed from the base station to the center of all of the bins that contain measurements. Each bin represents an average power measurement at a certain range from the base station. The range axis is then mapped to a logarithmic (common log) scale, the transmit power is combined with the parameter, *A*, and the two parameters of the following equivalent model are estimated via linear regression

$$P_r = A' - Br_L \tag{2}$$

where $r_L = \log_{10} r$ and $A' = P_t - A$.

Once the constants *A'* and *B* have been estimated, the mean trend of the propagation data is subtracted from the signal strength measurements and the standard deviation, σ, of the remaining zero-mean process is estimated. The value of σ represents the composite variation due to two primary factors: lognormal fading and measurement error. Both of these factors tend to introduce uncorrelated errors since the regression is computed for range measurements across all azimuth angles which greatly reduces most spatial correlation effects.

A fade margin, FM_σ, that ensures the desired service reliability, *F(z)*, can then be approximated

$$FM_\sigma = z\sigma \tag{3}$$

where $$F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}}\, dt \qquad (4)$$

For example, cell edge reliabilities of 75% and 90% correspond to fade margins of about 0.675σ and 1.282σ, respectively. Slightly more precise expressions can be obtained by limiting the range of the fading to $-4\sigma<z<2\sigma$ [3,4]. However, this provides only a minor improvement in the fade margin estimate since the area under the tails of the Normal density function is quite small.

It is now straightforward to derive the distance to the cell edge, *R*, at any desired signal strength threshold, P_{THRESH}, and service reliability, *F(z)*. From equations (1), (2), (3), and (4)

$$R = 10^{-(P_{THRESH} + FM_{\sigma} - A')/B} \qquad (5)$$

Any additional static (nonfading) margin, such as building penetration losses, can also be easily incorporated into the P_{THRESH} term. Thus, *A'*, *B* and, σ are all that is needed to determine the range from the base station to the cell edge.

Example:
Compute the range to the cell edge assuming the Hata (Cost-231) urban model constants for 1900 MHz and a base station antenna height of 30 meters:

$A = 140$
$B = 35.2$

Also assume

$\sigma = 8$ dB
$P_{THRESH} = -95$ dBm
$P_t = 50$ dBm (EIRP)
F(z) = 75% (i.e., $FM_{\sigma} = 0.675\,\sigma$)

From equation (5), for 75% cell edge reliability the estimated radius is

$$R = 10^{-(-95+5.4-50+140)/35.2} = 0.974\ km$$

Similarly, the radius for 90% cell edge reliability is given by

$$R = 10^{-(-95+10.3-50+140)/35.2} = 0.707\ km$$

With no measurement error, exact knowledge of the propagation parameters *A'*, *B*, and σ is equivalent to the exact knowledge of *R*. The remainder of this paper deals with the details of how to estimate the parameters *A'*, *B*, and σ from actual drive test data and the precision that results from doing so.

σ	6	6	8	8	10	10
B	35	30	4 0	35	30	35
Edge Reliability	75%	90%	75%	90%	75%	90%
Area Reliability	91.65%	96.87%	90.72%	96.57%	87.4%	96.03%

Table 1. The relationship between area and edge reliability for various propagation parameters σ and *B.*

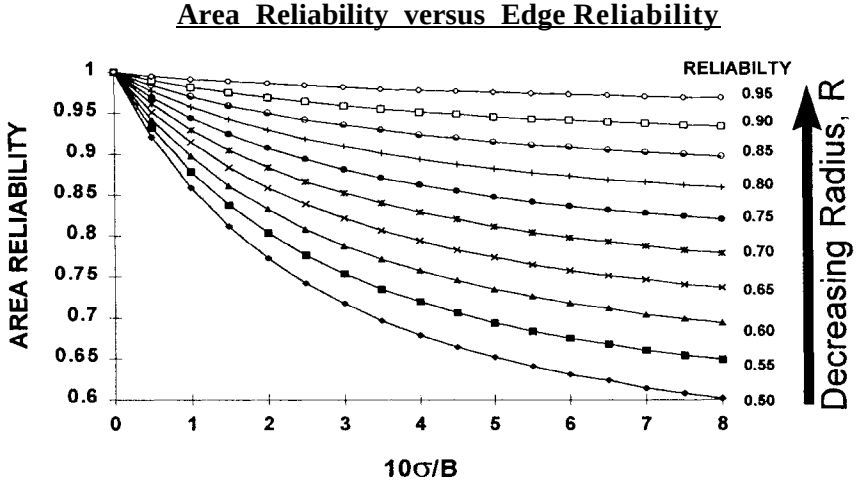

Figure 3. Area reliability (ordinate) and cell edge reliability (see parameter associated with each curve) versus 10σ/*B*, where σ is the standard deviation of the RF signal and *B /10* is the propagation path loss exponent. For a given value of 10σ/*B*, knowledge of the cell edge reliability directly determines the area reliability. (The figure is redrawn from Chapter 2 of reference[2]). Note, increasing cell edge reliability is equivalent to decreasing the radius of coverage.

IV AREA RELIABILITY ESTIMATION APPROACH

The relationship between the reliability of coverage over a circular area and the reliability of coverage on the perimeter of the circle was first established by D.O. Reudink (circa 1974) [2]. The main finding of this study was that cell area reliability and cell edge reliability obey the simple relationship illustrated in Figure 3. As long as the propagation path follows a power law, this relationship is completely determined by the ratio 10σ/*B*, where σ is the standard deviation of the lognormal fading within the cell and *B*/10 is the path loss exponent (e.g., typical values are σ=8 dB and *B*/10=3.52). As shown in Table 1, given exact knowledge of σ and *B*, the cell area reliability (and cell edge reliability) can be exactly computed (see also equation (6)). Note that although 75% cell edge reliability approximately corresponds to 90% cell area reliability, and 90% cell edge reliability approximately corresponds to 97% cell area reliability, their exact values depend on the propagation parameters of each cell (i.e., σ and *B).*

	Example 1	Example 2
Edge Reliability	75%	75%
σ	8	10
B	40	30
Area Reliability at 75% edge reliability	90.72%	87.4%
Radius at 75% edge reliability	1 km	0.9005 km
Radius at 90% area reliability	1.021 km	0.803 km
Edge reliability at 90% area reliability	73.5%	79.5%

Table 2. The relationship between cell area reliability, cell edge reliability and cell radius for different propagation parameters σ and *B* and the same transmit power.

The relationship in Figure 3 is apparently independent of the absolute cell radius, as well as being independent of the transmit and receive power which only serve to scale the radius. This seems to uncouple the problem of determining the coverage reliability from the problem estimating the size of a cell. However, the relationship in Figure 3 does not mean that the cell radius has no effect on coverage reliability. On the contrary, for the same two-way gain and transmit power, making the cell radius larger reduces the coverage reliability and decreasing the cell radius increases the coverage reliability. The dependency on cell radius is implicit through the desired edge reliability and equation (5). Table 2 demonstrates some of the relationships between cell radius, cell area reliability, and cell edge reliability for two cells designed with the same transmit power. In this table, the cell radius is computed from equation (5) and the area reliability is computed from equation (6). For example 1, the cell radius at 75% edge reliability (90.72% area) is 1 km, the radius at 73.5% edge reliability (90%, area) is 1 .021 km. The results are similar for example 2. Observe that changing the cell radius can have a pronounced effect on the reliability of RF coverage.

In Reudink's original derivation of area reliability, the explicit dependence of coverage on the size of the cell radius was purposely eliminated. Since the cell radius is one of the estimated quantities of interest in this paper, it is reintroduced into Reudink's expression in equation (6). The edge reliability on a contour of range, *r,* is

$$1 - P_{out}(r) = P(A' - B\log_{10} r \quad + X \quad > P_{THRESH})$$

$$= Q\left(\frac{P_{THRESH} - A' + B\log_{10} r}{\sigma}\right) = Q(a + b\ln r)$$

where $a = \dfrac{P_{THRESH} - A'}{\sigma}$ and $b = \dfrac{B\log_{10} e}{\sigma}$

The fraction of usable area, F_u, (i.e., area reliability) within the cell can be found by integrating the contour reliability across range [5,7]:

$$F_u = \frac{1}{\pi R^2}\int_0^R \left[1 - P_{out}(r)\right]\ 2\pi r dr = \frac{2}{R^2}\int_0^R Q(a + b\ln r)\quad r dr$$

$$F_u = Q(a + b\ln R) + \frac{e^{\frac{-2a}{b} + \frac{2}{b^2}}}{R^2}\left[1 - Q\left(a + b\ln R - \frac{2}{b}\right)\right] \qquad (6)$$

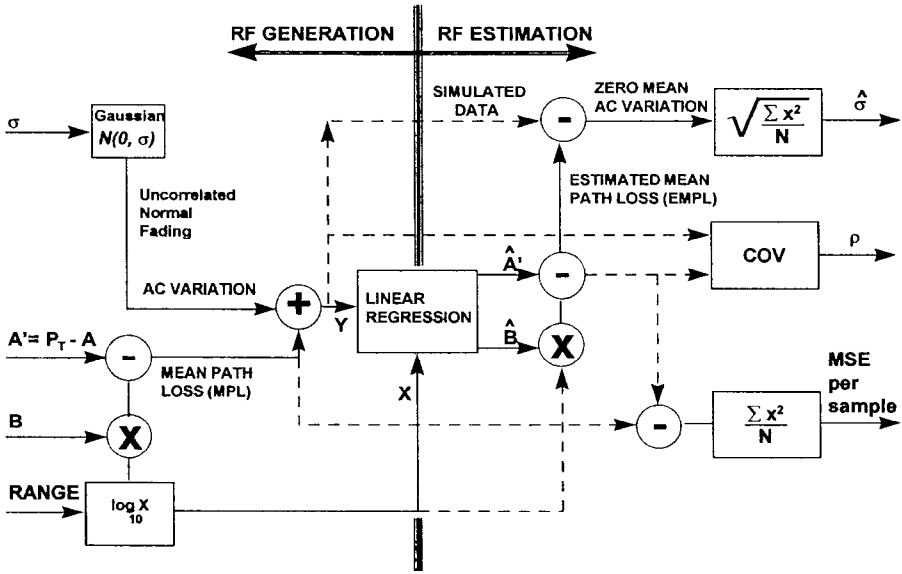

Figure 4. Block diagram of RF propagation simulation. The inputs to the simulation are *A', B,* and σ. The corresponding outputs are estimated as shown above. The mean-square-error per sample between the best fitting line and the true mean path loss is used to measure the performance of the estimation process. The correlation coefficient, ρ, of the best fitting line with the data is also computed.

This is the formula (i.e., $\hat{F}_u$) that is used throughout this paper to estimate the reliability of RF coverage over a circular area. Thus, the approach for measuring area reliability in this study is as follows:

1) measure the propagation parameters $\hat{A}'$, $\hat{B}$, and $\hat{\sigma}$ for each cell via linear regression

2) use these parameters to estimate the cell radius $\hat{R}$ from equation (5)

3) use the radius and the propagation parameters to estimate the reliability of coverage, $\hat{F}_u$, over the cell area using equation (6).

In the following section, a simulation is developed to compare the radius estimation and the area reliability estimation approaches for determining coverage reliability.

V RF PROPAGATION SIMULATION RESULTS

To test the validity of the radius estimation and area reliability estimation approaches, an RF propagation simulation was written, as shown in Figure 4. To reduce the computation, we use a single radial component that is uniformly sampled along its length. It is also assumed that the simulated measurements result from a uniform azimuthal sampling of the cell. Hence, the samples along this single radial represent the composite path loss fluctuations of radials in all azimuth directions. The single radial of this model is thus considered to be a linear superposition of

multiple radials that are uniformly spaced in azimuth. The simulation could explicitly calculate this superposition, but this is computationally inefficient and would have no effect on the results.

For a fixed azimuth angle, the fading is correlated along the radial from the base station. However, the fading between radial components at equally spaced azimuth angles is nearly uncorrelated. Since the regression is actually evaluated across all azimuth angles simultaneously, an uncorrelated Gaussian fading model is chosen. The standard deviation, σ, (typically 5-10 dB) is input into a Gaussian random number generator which produces uncorrelated normal random variables with zero mean and variance σ^2.

The mean path loss is computed for each range value as the product of the logarithm of range and the path loss coefficient, *B*, to which is added the intercept value *A'*. The variation due to fading is then added to the mean path loss (MPL) also shown in Figure 4. This concludes the RF signal generation portion of the simulation. The remainder of the simulation is concerned with estimating *A'*, *B*, and σ. Both $\hat{A}'$ and $\hat{B}$ are computed via linear regression. The estimated mean path loss is then subtracted from the simulated signal strength values and an estimate of the standard deviation, $\hat{\sigma}$, is made from the resulting zero-mean process. Two major criteria are used to evaluate the performance of the estimation procedure in the simulation:

1) the correlation coefficient, ρ, between the simulated data and the best fit line (EMPL). The closer ρ is to unity, the more linear the data. This measure is also used to characterize the reliability of the field data.

2) the mean square error (MSE) per sample between the best fit line (EMPL) and the true path loss in the simulation (MPL). This measure cannot be used in the field, since the true path loss is unknown.

Typical results from the simulation are shown in Figure 5. The following parameters were the inputs used to generate the 515 data points in this figure:

$A = 140$
$P_t = 50$ dBm (EIRP)

hence

$A' = P_t - A = -90$dBm
$B = 35.2$
$\sigma = 10$ dB

The corresponding outputs were

$\hat{A}' = -88.35$
$\hat{B} = 36.35$
$\hat{\sigma} = 10.63$ dB
$\rho = 0.826$ (correlation coefficient, where $\rho = 1$ for a line)
MSE per sample = 0.239 dB

The value of $\rho = 0.826$ is typical of that found in actual drive test data. The simulation estimated the input parameters very well since

$A' - \hat{A}' = -1.65$
$B - \hat{B} = -1.15$

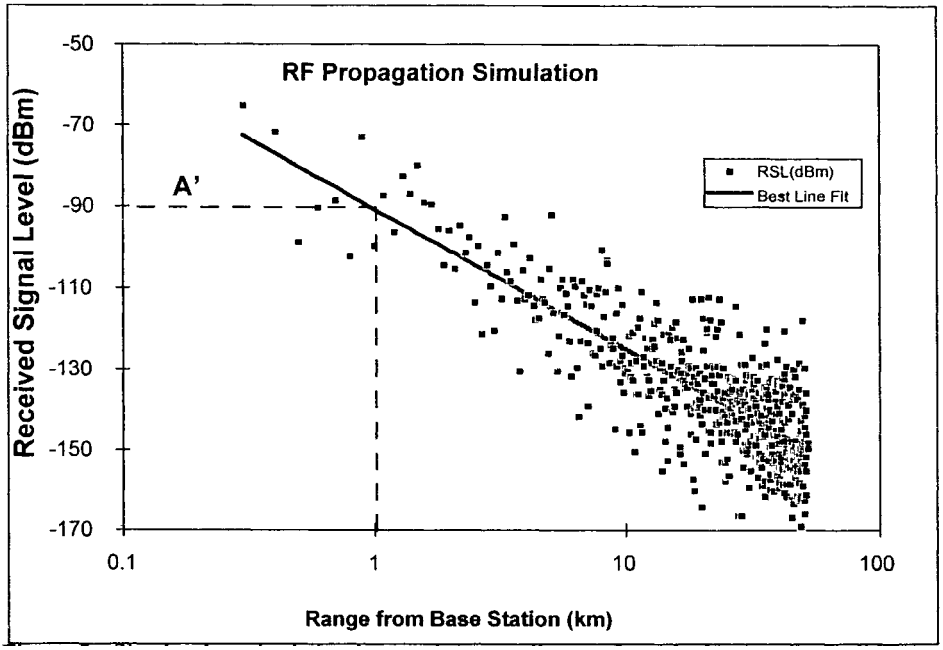

Figure 5. Simulated received signal strength versus distance from the base station and the best fitting linear approximation.

$\sigma - \hat{\sigma} = -0.63$
MSE $= 0.239$

The above four values can be made as close to zero as desired by increasing the number of simulated data points, *N*. Since these errors depend on the number of data samples used to compute the regression, a natural question is "How many data samples are necessary to achieve a given precision?" The accuracy of the measurement approach is examined in more detail in the next section.

VI MEASUREMENT ACCURACY VERSUS THE NUMBER OF SAMPLES

This section deals with determining the measurement error of the overall estimation process. The simulation is used to determine the probability densities of the following two random variables:

$$e_{Fu} = \frac{F_u - \hat{F}_u}{F_u} \quad \text{and} \quad e_R = \frac{R - \hat{R}}{R} \tag{7}$$

where

e_{Fu} is the relative error of the area availability estimate computed from equation (7)

e_R is the relative error of the cell radius estimate computed from equation (7)

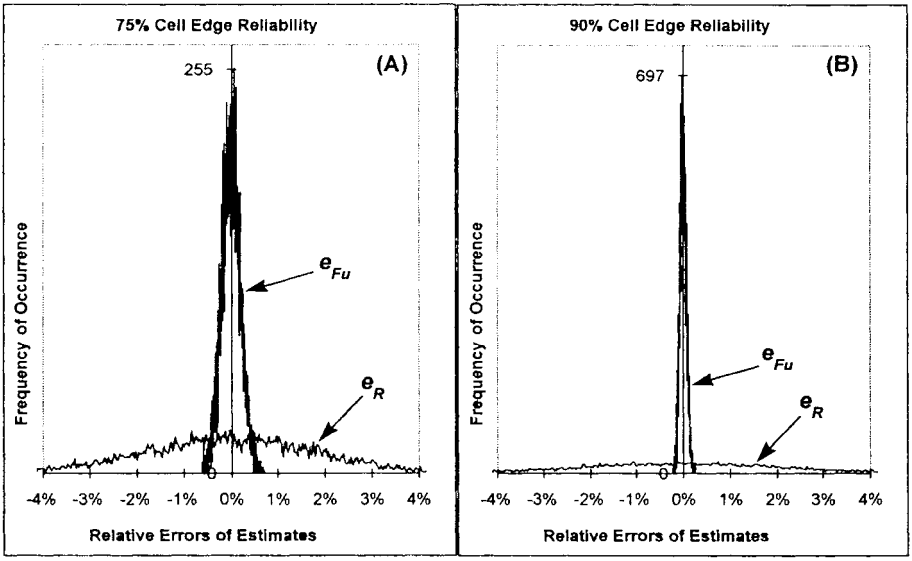

Figure 6. Histograms showing the simulated probability densities of the relative error e_{Fu} of the area availability estimate $\hat{F}_u$ and the relative error e_R of the cell radius estimate $\hat{R}$ for **(A)** a 75% cell edge reliability design **(B)** a 90% cell edge reliability design. The number of samples in the regression is 1000 and the standard deviation of the lognormal fading is $\sigma = 8$ dB.

F_u and $\hat{F}_u$ are the true and the estimated area reliability computed from equation (6)

R and $\hat{R}$ are the true and the estimated cell radius computed from equation (5)

The transformations specified by equation (7) allow a direct comparison of the cell radius estimate with the area reliability estimate, which would otherwise be difficult due to the differences in the dimensions of these two estimators (i.e., R is in kilometers and F_u is a percentage).

Typical probability densities for e_{Fu} and e_R are shown in Figure 6. Observe that the error of the cell radius estimate, e_R, is comparable in Figure 6A and 6l3 (Please note the change of scale on the ordinate). However, e_{Fu} is almost a factor of two smaller for the 90% cell edge reliability design. The evidence in Figure 6 and all of the histograms processed in this study demonstrate that both e_R and e_{Fu} are well modeled as zero-mean Normal random variables, and thus, only their respective variances are needed to characterize the precision of the estimates $\hat{R}$ and $\hat{F}_u$. These are determined empirically via Monte Carlo simulation.

We are interested in determining the inaccuracy, ΔR, of the estimate of the cell radius at a 95% confidence level. The inaccuracy is measured from empirical histograms by simulating e_R and determining ΔR such that

$$P\left(R - \Delta R \le \hat{R} \le R + \Delta R\right) = 95\%$$

The inaccuracy of the radius estimate, ΔR, is determined by the following two-sided test

$$c = F(z_c) = \frac{1}{\sqrt{2\pi}} \int_{-z_c}^{z_c} e^{-\frac{t^2}{2}}\, dt \tag{8}$$

where the z_c variable in equation (8) is chosen to yield the desired confidence level, c. For example, if c=95%, then z_c= 1.96. Since e_R has a mean of zero, the corresponding two-sided normalized radius inaccuracy is

$$\pm\delta_R = \pm\frac{\Delta R}{R} = \pm 1.96\sqrt{VAR(e_R)} \tag{9}$$

where δ_R is a dimensionless percentage of the cell radius, R.

Likewise, the inaccuracy of the area reliability estimate, $\hat{F}_u$, is estimated from histograms of e_{Fu} and determining ΔF_u such that

$$P\left(\hat{F}_u \le F_u + \Delta F_u\right) = 95\%$$

Since e_{Fu} also has a mean of zero, the inaccuracy of the area reliability estimate, ΔF_u, (one sided 95% confidence interval) is

$$\Delta F_u = 1.645\sqrt{VAR(e_{Fu})} \tag{10}$$

Each point in the plots in Figure 7 represents the precision (at 95% confidence) that is obtained after simulating and processing five million signal strength values.

Close inspection of Figure 7 reveals that for a given number of signal strength samples, N, the area reliability is much more precise (by one to two orders of magnitude) than the estimate of the cell radius. For example, 1000 samples in the regression are needed for about a ± 3% inaccuracy in the cell radius estimate. However, even with 500 samples in the regression, the area availability estimate is very precise. The inaccuracy of the area availability estimate is less than 0.5% for cells designed with 75% cell edge reliability and within 0.2% for cells designed with 90% cell edge reliability.

The inaccuracies of both of the estimates $\hat{R}$ and $\hat{F}_u$ can be approximated by the following expressions which were determined empirically from the data in Figure 7

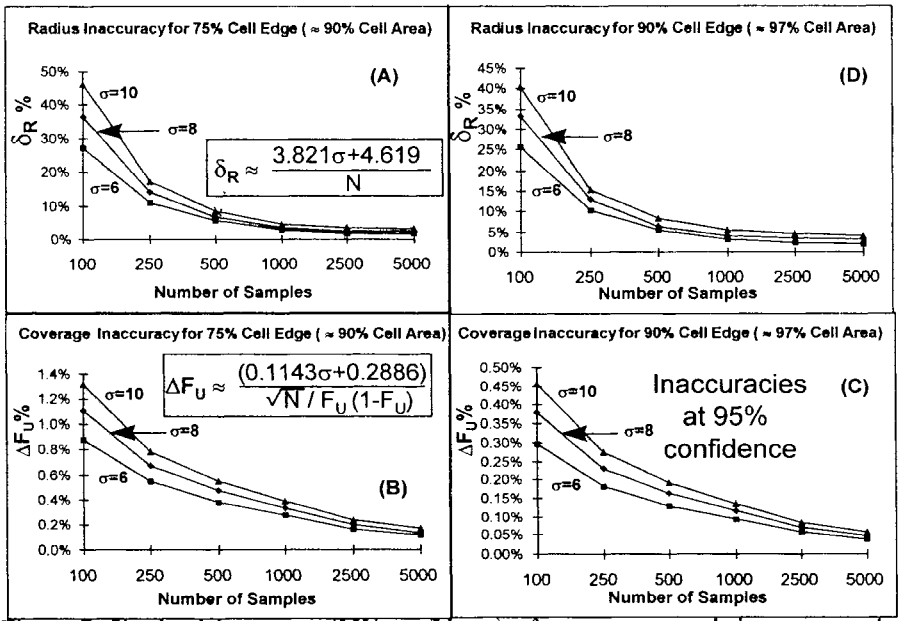

Figure 7. Simulated inaccuracy (95% confidence) of measurement techniques versus the number of samples in the regression; **(A)** cell radius estimate $\hat{R}$ of a 75% cell edge reliability design, **(B)** area reliability estimate $\hat{F}_u$ of a 75% cell edge reliability design **(C)** area reliability estimate $\hat{F}_u$ of a 90% cell edge reliability design, **(D)** cell radius estimate $\hat{R}$ of a 90% cell edge reliability design.

$$\delta_R = \frac{\Delta R}{R} \approx \frac{3.821\hat{\sigma} + 4.619}{N} \tag{11}$$

$$\Delta F_u \approx \frac{(0.1143\hat{\sigma} + 0.2886)\hat{F}_u(1-\hat{F}_u)}{\sqrt{N}} \tag{12}$$

where

N is the number of independent samples in the regression

$\hat{\sigma}$ is the estimated standard deviation of the lognormal fading in the cell

$\hat{F}_u$ is the estimated area reliability computed from equation (6)

Observe that the area reliability inaccuracy in equation (12), ΔF_u, is inversely proportional to $\sqrt{N}$. It is interesting to compare the magnitudes of the inaccuracies of the above area reliability measurements with the cell radius inaccuracies. Note that the area availability estimate, $\hat{F}_u$, shown in Figure 7C is always more than ten

times the precision of the cell radius estimate. Figures 7-B and 7-C also show that Reudink's area reliability estimate is insensitive to small (±2 dB) changes in σ.

The normalized radius inaccuracy in equation (11), δ_R, is inversely proportional to the number of samples in the regression, *N*, and directly proportional to the amount of lognormal fading, σ, in the cell. Interestingly, radio survey engineers have long recognized the negative effects that widely varying terrain and clutter environments have on RF coverage tests. They usually compensate for these effects by taking many more measurements in these areas. Equation (11) is simply the mathematical expression of this practice, specifying the relationship between the desired coverage inaccuracy, δ_R, the number of independent signal strength measurements, *N,* and the terrain fading factor within the cell, σ.

The most important finding of this analysis is that it is the precision of the estimate of the cell radius (i.e., equation (11)) that is the limiting factor in determining the quality of RF coverage, not the precision of the area reliability estimate.

VII DISCUSSION

An accurate method of determining the radius ($\hat{R}$) of individual cells was presented. This led to an even more precise technique for estimating the reliability of coverage over the area of the cell ($\hat{F}_u$).

It was shown that it is possible to obtain an excellent estimate of the area reliability even if the number of samples is insufficient for estimating the cell radius. This raises an interesting question concerning the determination of service reliability: "What is the best metric to use in classifying the quality of RF coverage? " This study indicates that area reliability alone is insufficient.

The results of this paper indicate that estimating the effective radius of a cell is the limiting factor in determining the RF coverage reliability. For a given propagation environment, computing the distance to the cell edge is deterministic (i.e., apply equation (5)). For real drive test data, the true cell radius is unknown and must be statistically estimated. It was shown that as long as the radius estimate is sufficiently precise, so is the area reliability estimate ($\hat{F}_u$).

If cell edge reliability is the desired coverage criterion, an accurate estimate of the cell radius is all that is needed since the cell edge reliability is ensured by the fade margin used to measure the radius.

However, if area reliability is the desired coverage criterion, then a minor adjustment to the cell edge reliability (and cell radius) must be made in each cell to compensate for the variation in the specific values of the propagation constants *A'*, *B*, and σ. This is easily done by first computing the propagation constants via linear regression and then computing the area reliability from equation (6) at fine range increments (e.g., steps of $\Delta R/2$). The desired radius can then be found by inverse interpolation.

The proposed technique is best suited for macrocells. However, it can be modified to work equally well in microcells by eliminating measurements that have line-of-sight to the base station. In many cells, the path loss is better described by a composite of two line segments that intersect at some breakpoint distance near the base station. For these cases, this distance is approximated and measurements before this point are eliminated since virtually no outages occur over this region. The best linear fit to

the path loss in the outer regions of the cell is extrapolated all the way to the base station. For most propagation scenarios, the error in the area reliability estimate due to this approximation is less than 0.5% and less than 1.5% for the radius estimate. Thus, the proposed method needs no additional parameters or modifications to accommodate dual-power law propagation environments. However, if these errors are not tolerable, they can easily be eliminated by incorporating breakpoint distance into Reudink's expression (equation (6)).

Although the precision in this study was determined via simulation, we have processed signal strength measurements from hundreds of cell sites and have found that the results are completely consistent with those of our simulation.

We have also found cell radius inaccuracy to be very useful in determining the sampling requirements of cellular drive tests [8,9].

This validation approach is particularly useful to anyone involved with cell planning since this equates the problem of determining the reliability of RF coverage with that of determining the effective size of the cell. The latter concept is clearly more useful to the cellular network planner.

VIII CONCLUSIONS

Cell radius inaccuracy has been proposed as a new method for measuring RF cellular coverage. The technique measures the distance from the base station to the cell edge (equation (5)) and quantifies the precision by also specifying the uncertainty of the radius estimate. In addition, the approach provides an estimate of the area reliability (equation (6)) which was shown to be much more accurate than the cell radius estimate. Empirical formulas are given that approximate the precision of both of these estimates (equations (11) and (12)).

The recommended technique uses linear regression to estimate the minimum mean square path loss within each cell, and is thus very tolerant to estimation errors due to terrain fluctuations (e.g., lognormal fading). The approach provides the best circular approximation to any equal power contour, at any desired reliability. Thus, this method is ideal for cell site planning with omni antennas for any wireless technology. For example, using standard CW drive test measurements, this technique can help verify that the RF design meets the proper amount of overlap in coverage needed to support the soft handoff regions of CDMA. The method can easily be modified to provide equally valid coverage measurements for sectored cells [9].

The conclusion of this study is that cell radius estimation and area reliability estimation should not be treated separately, and that cell radius inaccuracy is the more critical validation measure.

ACKNOWLEDGMENTS

This paper is dedicated to the memory of Professor Charles Wilhelm Bernardin.

The authors would like to thank the management of NORTEL Wireless Engineering Services for providing the funding and the environment necessary for this research. We would like to acknowledge Dr. Richard Tang for originally suggesting the regression approach. Also, we would like to thank Professor Venugopal Veeravalli for his invaluable suggestions concerning the error performance of the area reliability estimation approach included here. We are especially grateful to Professor Veeravalli for performing the rigorous analysis necessary to derive equation (6). We

would also like to thank Dr. Sudheer Grandhi, Pulin Patel, Reid Chang and Mark Prasse for taking the time to critically evaluate this manuscript.

REFERENCES

[1] Hata, M., "Empirical Formula for Propagation Loss in Land Mobile Radio Services," IEEE Transactions on Vehicular Technology, vol. VT-29, No. 3, August 1980, pp-317-325.

[2] Reudink, D.O, Microwave Mobile Communications, edited by Jakes, W.C., IEEE Press, reprinted 1993, ISBN 0-7803-1069-1, Chapter 2, pp. 126-128.

[3] Steele, R., Mobile Radio Communications, Pentech Press, 1992.

[4] Lee, C.C. and Steele, R., "Signal-to-Interference calculations for Modern TDMA Cellular Systems," submitted to IEE Proceedings-I, Communications, Speech and Vision.

[5] Bernardin, P., Yee, M., and Ellis, T., "Estimating the Cell Radius from Signal Strength Measurements," 6th WINLAB Workshop, March, 20-21, 1997.

[6] Bernardin, P., Yee, M., and Ellis, T., "Estimating the Range to the Cell Edge from Signal Strength Measurements," 47th IEEE Vehicular Technology Conference, May 5-7, 1997.

[7] Bernardin, P., Yee, M., and Ellis, T., "Cell Radius Inaccuracy: A New Measure of Coverage Reliability," submitted 8/27/96 to IEEE Transactions on Vehicular Technology.

[8] Bernardin, P., Yee, M., and Ellis, T., "On the Sampling Requirements of an Omni Cellular Drive Test," submitted 2/18/97 to IEEE Transactions on Antennas and Propagation (Special Wireless Issue).

[9] Bernardin, P. and Faruque, S., "Sampling Requirements in Drive Testing Sectored Cells," submitted 5/28/97 to IEEE Transactions on Communications.

[10] Lee, W.C.Y., Mobile Communications Engineering, McGraw-Hill Book Co., 1982, p. 104.

[11] Gudmundson, B., "Correlation Model for Shadow Fading in Mobile Radio Systems," Electronics Letters, Vol. 27, November, 7, 1991, pp. 2145-2146.

Statistical Model of Spatially Correlated Shadow-Fading Patterns in Wireless Systems

Krishnan Kumaran, Sem Borst
Bell Laboratories, Lucent Technologies
600 Mountain Avenue, Murray Hill, NJ 07974
{kumaran,sem}@research.bell-labs.com

Abstract

We discuss a statistical model to generate correlated shadow-fading patterns for wireless systems in the absence of detailed propagation and landscape information. The current model of shadow-fading postulates a log-normal marginal distribution for the fading values, and does not address correlations. Subsequent introduction of correlations via autocorrelation models for individual mobiles results in anomalous effects that depend on traffic density and mobility. Our approach involves generating a pre-computed set of fading values with the right marginal distributions and spatial correlations. Motivated from statistical physics, the correlations are introduced in terms of "interaction" parameters, which can be computed from local measurements. The model is efficiently implemented using standard linear-algebra methods and is amenable to a statistical-mechanics treatment. Numerical results show that the patterns produced are sufficiently clustered and appear reasonable to visual inspection.

1 Introduction

Propagation is a crucial factor in the design and performance of wireless systems. There are several statistical methods to estimate propagation under different conditions, and most of these methods are computationally intensive. 2D and 3D ray tracing methods, for example, are popular techniques with much effort devoted to enhance their computational viability (see [l], [7] and references therein). Inaccuracies in modeling specific aspects of propagation, such as shadow-fading, could lead to distorted capacity and performance estimates in simulations and approximate analytical calculations. Shadow-fading is the slowly varying component of propagation that arises, as the name suggests,

primarily from obstacles. Raleigh fading, which is a propagation variation on a smaller scale, is also an important factor for stationary or slow moving users of the wireless system, but we do not deal with it here.

For some applications, it is useful to have a statistical model of propagation with an empirical validity for the conditions of interest. It is in this spirit that the log-normal distribution, which is most widely used at the present, was proposed. In this model, the received power at each location is assumed to be independent of other locations and log-normally distributed about a deterministic large-scale decay, i.e.,

$$P(r) = P_0(r)\, e^{s} \tag{1}$$

where s is normally distributed with mean 0 and $\log P_0(r)$ is the average value of $\log P(r)$ received at distance r from the transmitter after large-scale fading. Usual forms for this function are $P_0(r) = A + B \log(r)$ (the Hata Model [5]) or simply the power law decay $P_0(r) = \frac{K}{r^d}$ where A, B, K and d are specific constants depending on the environment. While these models are simple and account for the observed marginal distributions, they fail to capture the spatial correlations inherent in shadow-fading, which cause its slowly-varying behavior. As a result, capacity estimates from simulations of wireless systems, especially sensitive ones such as CDMA, could be highly distorted.

Our interest in generating realistic shadow-fading patterns arose in the context of a simulation study of a novel Interference-Based Dynamic Channel Assignment algorithm. The proposed algorithm relies on periodic interference measurements on the inactive frequencies so as to identify appropriate candidate channels To obtain accurate estimates of the impact on system capacity and voice quality, we were confronted with the problem of generating shadow-fading values consistent with specified marginal distributions and correlation parameters.

The issue of capturing correlations in shadow-fading has been addressed previously by Gudmundson [3, 4] by means of an auto-regressive model that is formulated differently from ours. The correlations considered there are between the shadow-fading values seen by a mobile as it moves in the environment. However, our approach is to precompute a static, spatially correlated shadow-fading pattern with the desired statistics. Thus, we do not encounter anomalies such as the assignment of completely different shadow-fading values for different mobiles at nearby locations or at different times at the same location. In general, the approach is best suited for simulations that are aimed at obtaining perfor-

mance characteristics not directly related to specific propagation aspects, but would benefit from using realistic fading models for generic environments.

2 Description of the Model

Assume that we are given a 2D lattice representation of the region in which we are interested, with the positions of the lattice points being represented by the pair of indices (i, j) which take integer values in the range $[0, N-1]$. While we are using a rectangular lattice in this description, it would be equally valid to use a polar grid with the appropriate local transformation of the coefficients*. We define the random field ϕ_{ij}, which describes the shadow-fading at position (i, j), from which the received power is calculated as before.

$$P_{ij} = P_{0_{ij}}\, e^{\phi_{ij}} \tag{2}$$

We then propose the following joint normal distribution for the $\{\phi_{ij}\}$

$$Prob(\{\phi_{ij}\}) = \frac{e^{-\beta E(\{\phi_{ij}\})}}{\mathcal{Z}} \tag{3}$$

where the "energy" functional $E(\{\phi_{ij}\})$ is given by

$$E(\{\phi_{ij}\}) = \sum_{ij} [\lambda_{ij}\phi_{ij}^2 + \mu_{ij}^x(\phi_{ij} - \phi_{i+1,j})^2 + \mu_{ij}^y(\phi_{ij} - \phi_{i,j+1})^2] \tag{4}$$

The parameters λ_{ij} are related to the inverse of the conditional variances of ϕ_{ij}, while the parameters μ_{ij}^x and μ_{ij}^y introduce correlation between values at neighboring locations, and thereby potentially a global correlation. We restrict our treatment to the case where all the coefficients are positive. We will postpone to a later section the discussion of the relationship between these parameters and the measured variance and correlation parameters. $\mathcal{Z}$ is a normalization constant known as the "partition function" , and is given by $\mathcal{Z} = \sum_{\{\phi_{ij}\}} e^{-\beta E(\{\phi_{ij}\})}$

This terminology is owed to statistical physics, where the μ_{ij} and λ_{ij} are considered "interaction parameters" of the ϕ_{ij}, which are termed "order parameters". The "inverse-temperature" β, which determines the "fluctuations" from the mean configuration, is not essential to our purpose and is set to unity in the rest of this analysis.

* Polar coordinates would, in fact, be better suited to capture the radial nature of the propagation from a point source, by setting different correlation strengths along radial and tangential directions. The following analysis is however applicable to any coordinate system.

We now proceed to examine the properties of equation (4). We can rewrite the energy functional in matrix notation as

$$E(\phi) = \phi^{\mathbf{T}}\, \mathbf{A}\, \phi \tag{5}$$

where ϕ is a column vector containing all the ϕ_{ij}, i.e., $\phi_k = \phi_{i+N,j} = \phi_{ij}$ or $\phi_k = \phi_{int(k/N),mod(k,N),}$. and A is a symmetric, block-tridiagonal matrix of the form

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_1 & \mathbf{D}_1^T & \mathbf{0} & \cdots \\ \mathbf{D}_1 & \mathbf{A}_2 & \mathbf{D}_2^T & \ddots \\ \mathbf{0} & \mathbf{D}_2 & \mathbf{A}_3 & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{pmatrix}$$

The matrices $\mathbf{D}_i$ and $\mathbf{A}_i$ are given by

$$\mathbf{D}_i = \begin{pmatrix} -\mu^x_{i-1,1} & 0 & \cdots \\ 0 & -\mu^x_{i-1,2} & 0 \\ \vdots & 0 & \ddots \end{pmatrix}, \quad \mathbf{A}_i = \begin{pmatrix} \chi_{i1} & -\mu^y_{i1} & 0 \\ -\mu^y_{i1} & \chi_{i2} & \ddots \\ 0 & \ddots & \ddots \end{pmatrix}$$

where $\chi_{ij} = \lambda_{ij} + \mu^y_{i,j-1} + \mu^y_{ij} + \mu^x_{i-1,j} + \mu^x_{ij}$. We have imposed the boundary conditions $\mu_{ij} = 0$ for $i, j \leq 0$ or $i, j \geq N - 1$. Several interesting mathematical properties of the specific distribution (3) are discussed in [6]. In particular, **A** is diagonally dominant and hence positive-definite, and its inverse consists of all positive entries.

3 Algorithm

We now discuss an algorithm to generate numbers obeying the joint normal distribution (3). The matrix **A**, being symmetric and positive-definite, can be decomposed as

$$\mathbf{A} = \mathbf{L}\mathbf{L}^T \tag{6}$$

which is also referred to as taking the "square root" of the matrix **A** or as the Cholesky decomposition of **A**. In our case, the block tridiagonal structure of **A** offers considerable computational advantages by breaking down equation (6) as

$$\mathbf{A} = \begin{pmatrix} \mathbf{L}_1 & \mathbf{0} & \cdots & \cdots \\ \mathbf{B}_1 & \mathbf{L}_2 & \ddots & \ddots \\ \mathbf{0} & \mathbf{B}_2 & \mathbf{L}_3 & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{pmatrix} \begin{pmatrix} \mathbf{L}_1^T & \mathbf{B}_1^T & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{L}_2^T & \mathbf{B}_2^T & \ddots \\ \vdots & \ddots & \mathbf{L}_3^T & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{pmatrix} \tag{7}$$

and the smaller individual sub-matrices are obtained using standard Cholesky decomposition as

$$\begin{aligned} \mathbf{L}_i \mathbf{L}_i^T &= \mathbf{A}_i - \mathbf{B}_{i-1}\mathbf{B}_{i-1}^T \\ \mathbf{B}_i &= \mathbf{D}_i \mathbf{L}_i^{-T} \end{aligned} \tag{8}$$

†This decomposition allows the change of variables $\psi = \mathbf{L}^T \phi$ and it is readily verified that the variables ψ_{ij} are uncorrelated univariate Gaussians with mean 0. Thus the ϕ_{ij} can be obtained by inversion as $\phi = \mathbf{L}^{-T}\psi$. Again, the inversion can be performed more efficiently than in the general case using the special structure of the matrix as $\mathbf{L}_i^T \phi_i = \psi_i - \mathbf{B}_i^T \phi_{i+1}$. Given the matrix $\mathbf{A}$, the algorithm to generate the distribution is hence summarized as follows.

- Perform the decomposition (8) and store the matrix $\mathbf{L}^T$.
- Generate the independent Gaussian variables ψ_{ij} with mean 0 and variance 1. This may be performed by several methods using the standard uniform random-number generators in the range [0, 1]. A simple and efficient option is the cosine method which involves generating pairs of uncorrelated Gaussian variables ψ_1 and ψ_2 using the uniform variables u_1 and u_2 as

$$\begin{aligned} \psi_1 &= \sqrt{-2 \ln u_1} \cos 2\pi u_2 \\ \psi_2 &= \sqrt{-2 \ln u_1} \sin 2\pi u_2 \end{aligned}$$

- Apply $\mathbf{L}^{-T}$ to the vector ψ to obtain ϕ.

Note that successive samples ϕ can be generated by simply applying the inverse of the stored upper triangular matrix $\mathbf{L}^T$ to a different ψ, which can be performed with much greater computational efficiency than the first sample.

4 Measurements and Parameter Estimation

For shadow-fading, measurements available pertain to marginal distributions and correlation lengths. In our case, it is difficult to obtain the marginal distributions exactly in terms of our parameters, since this requires prior knowledge of the inverse of the matrix $\mathbf{A}$. Suitable approximations may be made using perturbation expansions and the properties of the matrix $\mathbf{A}$. For identically distributed ϕ_{ij}, one can derive approximate relationships between the parameters of the model and the variance and correlation of the generated distribution.

†The superscript $-T$ stands for the inverse of the transpose.

For large grid sizes, one may use the infinite lattice approximation to obtain the following estimates of the decay rates and variances. The λ_{ij} are obtained by solving

$$\frac{\log \lambda_{ij}}{\lambda_{ij}} = -4\pi d_x d_y \sigma^2 \tag{9}$$

from which the μ's are obtained as

$$\begin{aligned} \mu_{ij}^x &= \lambda_{ij}\, d_x^2 \\ \mu_{ij}^y &= \lambda_{ij}\, d_y^2 \end{aligned} \tag{10}$$

These results are also applicable when the parameters change with location, either due to changing variances/decay lengths or due to changing grid resolutions. However, we require these changes to be slow, i.e. they occur over spatial scales larger than the correlation lengths. With appropriate local rotation and rescaling of coordinates, the results can hence be applied to polar grid representations provided the region of interest is not too close to the origin.

In the limit, as the grid representation becomes increasingly fine, the exponential decay of correlations can be rigorously proved [2]. Proofs and further discussion of the above results follow in a more detailed version.

5 Conclusion

Numerical results, as seen in figures 1 and 2, show that "realistic" looking patterns can be produced by this method. The computational efficiency and the off-line nature of this computation help speed up simulations. The values generated have consistent statistical properties, and are not affected by mobility, as autocorrelation models are. However, additional memory is required to store the patterns generated, which limits the admissible grid size/resolution.

References

[1] S.J. Fortune, D.M. Gay, B.W. Kernighan, O. Landron, R.A. Valenzuela, and M.H. Wright. WISE Design of Indoor Wireless Systems: Practical Computation and Optimization. *IEEE Computational Science & Engineering*, 2(1):58–68, Spring 1995.

[2] S.E. Golowich. Personal Communication.

[3] M. Gudmundson. Analysis of Hand-over Algorithm. In *41st IEEE Transactions on Veh. Tech. Conf.*, pages 537–542, 1991.

[4] M. Gudmundson. Correlation Model for Shadow Fading in Mobile Radio Systems. *Electronics Letters*, 27(23):2145–2146, November 1991.

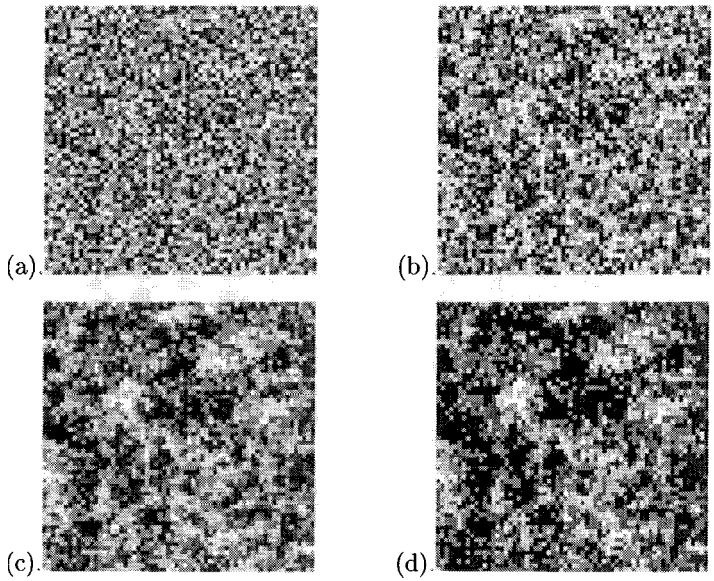

Figure 1: (a) Uncorrelated Gaussian pattern. (b) through (d) - Increasing correlation strengths produce more clustered patterns from (a).

[5] M. Hata. Empirical Formula for Propagation Loss in Land Mobile Radio Services. *IEEE Transactions on Vehicular Technology*, 29:317–325, 1980.

[6] K. Kumaran, D. Geiger, and L. Gurvits. Illusory surfaces and visual organization. *Network: Computation in Neural Systems*, 7(1):33–60, February 1996.

[7] A. Rajkumar, B.F. Naylor, F. Feisullin, and L. Rogers. Predicting RF coverage in Large Environments using Ray-Beam Tracing and Partitioning Tree Represented Geometry. *Wireless Networks*, 2(2):143–154, June 1996.

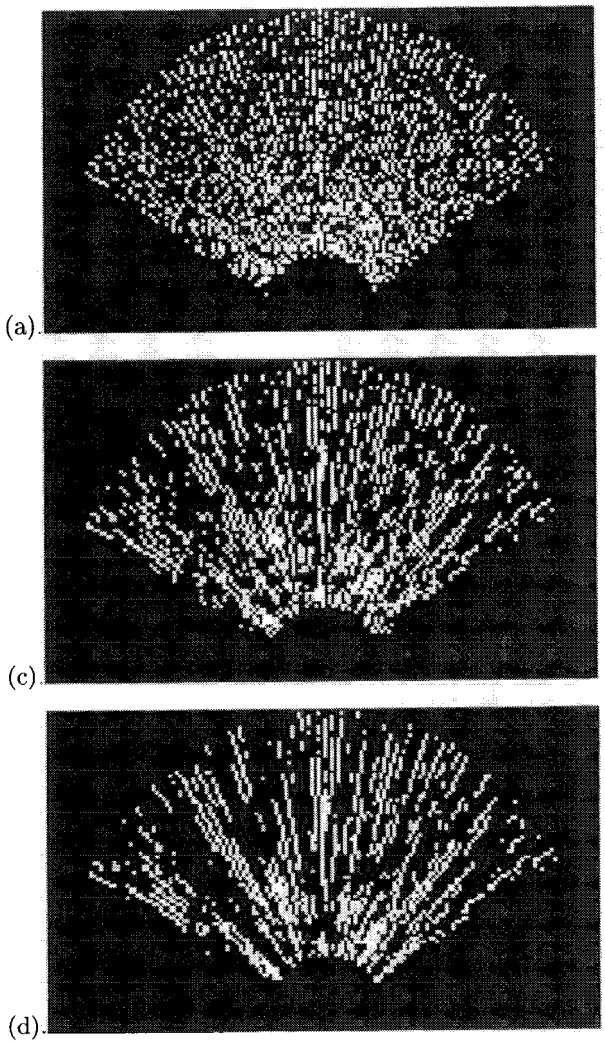

Figure 2: (a) Uncorrelated Radial pattern. (b),(c) - Increasing radial correlation strengths shadow out certain sections along each radial line, while the remaining regions receive a stronger signal.

A MODEL FOR WWW AND RPC TRAFFIC IN A WIRELESS ACCESS NETWORK

Erik Anderlind and Jens Zander

Radio Communications Systems Laboratory
Royal Institute of Technology (KTH), Stockholm, SWEDEN.

Abstract: We propose a simple model for future data traffic in wireless access networks. It is based on an analysis of the TCP/IP protocols for data communication. Model parameters are selected so as to resemble traffic from WWW access and from distributed file systems. By changing a single parameter, the model can be switched between resembling up- or down-link traffic. It is intended for design and performance analysis of radio resource allocation algorithms in future wireless systems.

1.1 INTRODUCTION

Performance analysis of wireless access networks has mainly been done with very simple models consisting of either Poisson packet arrival processes or continuous rate streams with (long) exponentially distributed durations. Traffic from popular multimedia applications does not seem to fit well with these analytically convenient models. Measurements of computer LAN traffic [2, 3, 5, 7, 8, 10] also manifest a large discrepancy from the widespread Poisson traffic assumption.

In this work we develop a network layer data traffic model that can be used for design and performance analysis of algorithms for dynamic channel allocation, power control and access control in wireless cellular networks. We focus on two important data traffic types that we believe future networks must be capable of efficiently servicing: WWW traffic, and Remote Procedure Call (RPC) traffic due to distributed file systems or data base enquiries. This type of traffic could be common to *nomadic* computing terminals, capable of autonomous operation, but benefiting from transparent access to data when this is possible.

A fundamental difference between a cellular radio network and e.g. a wired LAN, such as the ethernet, is that a significantly lower proportion of the system bandwidth can be allotted to a specific user. Target rates for wide area coverage are in the order of 32 to 144 kbps, compared to the 10 Mbps of a standard ethernet. Therefor packet transmission times are significantly longer while the efficiency loss of a 1-10 ms inter-packet time (e.g. due to protocol processing), can more easily be tolerated. Another difference is that cellular networks usually separate up- (from terminal to network) and down-link traffic, either in frequency (FDD) or time (TDD).

In the short term, data traffic can be highly asymmetric. In one direction data is transferred, while in the other there is mainly acknowledgment traffic. The average amount of data transferred during a session is not the same for each direction either. Studies of the distributed file systems AFS [11] and NFS [5] indicate a ratio of $\alpha = 3.2$ and 2.7 data read operations for every write. The cited ratios are for client stations with large disks, allowing caching of previously read data, and utilizing a write-back cache coherency policy. In a study of WWW traffic by Nieminen [8], 90% of the transmitted bytes were in the down-link direction.

A typical session for a packet data user could be the one shown in Fig. 1.. The session consists of several bursts of packets, inter-spaced by idle periods. It normally starts with registration and authentication towards the radio network. Thereafter there may be an additional registration and authentication with a terminal's home agent (e.g. using mobileIP [6,9]). Following these are data traffic bursts due to e.g. WWW browsing or use of the home network's distributed file system. Finally there may be a deregistration sequence.

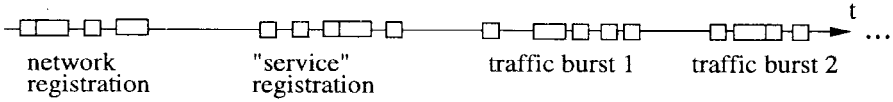

Figure 1. Example of a packet data user session.

The objective of the model presented below is to catch the non-continuous nature of packet transmissions. We believe that as long as the access protocol doesn't utilize specific knowledge about the model's parameters, their exact values are of less importance. In addition to the traffic profile described in the model, there may be numerous

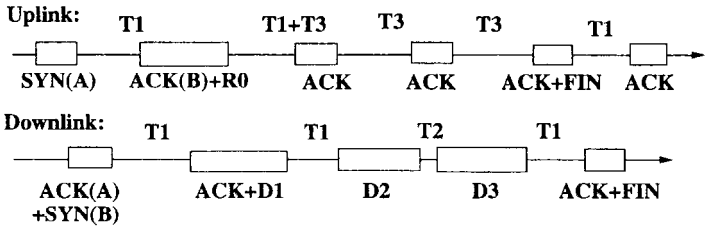

Figure 2. Sequence of packets (on the IP level) for a TCP connection with one request packet R0, and retrieved data in three data packets Dx.

radio resource management control messages e.g. for power control, that might need to be sent over the physical channel. These are outside the scope of this network layer traffic description.

1.2 THE TCP AND RPC PROTOCOLS

As a basis for our model we have studied the TCP [12] and Remote Procedure Call (RPC) protocols. Both protocols use the IP [4, 13] network layer protocol. As IPv6 is set to replace IPv4, we calculate protocol overhead using the later version. RPC implementations use either TCP or the simpler, non-acknowledged UDP transport protocol. We also assume that mobility is handled using care-of addresses (COAs) inserted as IPv6 extension headers. First we study the TCP protocol.

Assume a WWW user requests a WWW . html document using the HTTP protocol [15]. A TCP connection (Fig. 2.) must then be established for the duration of the transfer. First the sender and receiver must synchronize their flow control windows (the SYN ( ) and ACK ( ) packets). The time T_1 corresponds to one round trip delay. Thereafter the file is requested (data packet R0) and down-loaded (data packets Dx). Provided the flow control window is sufficiently large, the time T_2 between data packets depends mostly on the processing speed of the sending terminal and therefore it is short. However, when starting up a TCP connection, the flow control window is normally reduced. On the receiving side, each received IP packet is acknowledged (packets ACKx), where the interpacket time T_3 also includes transmission time for the data packets.

Once the last packet has been acknowledged, the link is closed using a similar double acknowledgment scheme as for the link establishment. The length of SYN and ACK packets is primarily determined by the protocol overhead: The minimum length of a TCP header encapsulated in an IPv6 packet with a single IP extension header for the COA is 20 + 20 + 40 = 80 bytes. To this must be added any additional IP or TCP protocol options.

In HTML v1.0 each embedded image is retrieved using a separate TCP connection. Modern browsers typically open up to four simultaneous TCP sessions when there are

several embedded objects. For future systems we can assume the use of HTML v1.1 improvements allowing concatenation of all requests into one TCP session.

The packet sequence (not shown) for an RPC over UDP is somewhat similar. The requesting client sends an initial packet specifying the targeted file or operation. The request is acknowledged and thereafter the file (or a part of it) is transferred in a set of data packets. Prior to retrieving a file, the client may additionally perform one or several file path-name or directory lookups which generates short or medium size packets. The minimum length of an encapsulated UDP packet header is 68 bytes.

Measurements by Guesela [5] report a bimodal distribution for NFS [14] packet lengths. The short packets, around 144 bytes, are for requests and responses containing NFS and RPC headers, while the long data packets, mostly 1500 bytes, are due to IPv4 packet fragmentation of the transmitted data. From the same study, the protocol processing for a SUN-3 workstation resulted in a fragment interpacket time of $T_2 = 1.9$ ms.

1.3 PACKET DATA MODEL

As we saw in the previous section, data exchange typically consists of a number of short packets with round-trip interpacket delays, needed to set up the request, and then a set of longer data packets with short interpacket delays, transporting the requested data. In order to keep the model simple, we need to find a compromise between WWW and RPC traffic. We first define the model in "pseudo-code" and then discuss the setting of parameter values (See also Fig. 3.). The sequences of bursts alternate between data and acknowledgment traffic according to the previously defined read-write probability α.

Packet Generation Algorithm

1. If "uplink" then $B = \frac{1}{1+\alpha}$ else $B = 1 - \frac{1}{1+\alpha}$.
2. Output packet $L1$. When sent, wait time T_1.
3. Generate uniform RV $U \sim [0,1]$. If $U > B$ goto 7.

* Data *

4. Output packet $L2$. When sent, wait time T_1.
5. Output packet $L3$. When sent, wait time T_1.
6. Output packet $L1$. Goto 9.

* Request + Acknowledgments *

7. Output two packet of length $L1$
8. For i=1 to N
   Wait random time T_3. Output packet L_1.
   end
9. Wait exp. distrib. time with avg T_4. Goto 2.

For simplicity the model has only three packet types. The short packet, $L1 = 80$ bytes, corresponds to SYN, ACK, path-name lookup or specification packets of approximately minimum packet length. The packet $L2 = 1.5$ kbytes symbolizes the TCP

slow start. Assuming a residual link layer BER of 10^{-6}, the standard ethernet packet length results in an approximate TCP retransmission probability of 1.2%. Finally the variable length packet $L3$ contains the rest of the transmitted data.

The time T_1 corresponds to one round trip delay, which we assume is to a remote host connected over a WAN, and therefore set it to $T_1 = 300$ ms. It is measured from the time the previous packet has been successfully transmitted until the next packet is generated. Based on the previously cited studies, but adjusting for WWW traffic, we set the read-write ratio to $\alpha = 4$.

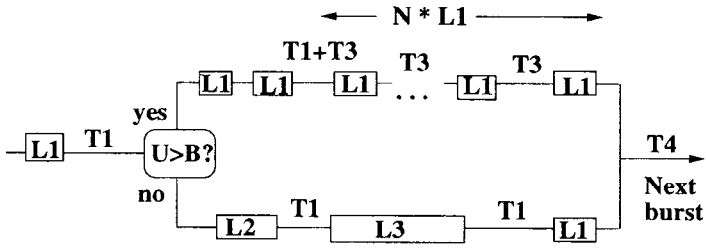

Figure 3. Proposed traffic model: The upper branch is for a request and acknowledgment sequence, while the lower branch represents the transfered data.

Beginning with the data transfer, we observed in the previous section that for congestion control reasons, TCP utilizes a slow start mechanism, which introduces a round-trip delay after the first packet. Once the flow control window is open, the interpacket delays tend to be short and therefore we have set the time T_2 from Fig. 2. to zero and collected all additional data into one large packet of length $L3$. This simplification would however not be suitable if the link data rate is very high.

Continuing with the request and acknowledgment sequence, the data request is modeled as two back-to-back packet of length L_1 . The acknowledgment interpacket time T_3 in a steady state depends primarily on the time it takes to send one IP packet. Based on a 1.5 kbyte packet transmitted at 128 kbps and adding some variability, we model T_3 as uniformly distributed within [50, 150] ms. The time T_3 between the request and the first acknowledgment approximates the effect of initial file server latencies.

The time to the next burst arrival (Fig. 1.) is dependent on user interactions and is modeled as an exponentially distributed variable with average $T_4 = 10$ s. Although actual arrival times surely don't correspond to this simplification and often are significantly longer (see e.g. [7, 8]), the chosen value is sufficiently long to necessitate a release of unused resources and short enough to be practical for simulation studies. Nevertheless, one will probably need to increase this parameter if the traffic model is combined with a mobility model where data users move over more than a small proportion of a typical cell's area. Due to lack of empirical data we have not modeled the effect of a finite session length. In addition, the time will most certainly depend on how the service usage is priced.

What remains to be specified is the amount of transmitted data in each burst, which will determine $L3$ and the number of acknowledgment packets N. Measurements of a very large number of WWW traces [3], report that the average `.html` file was 6.4 kbytes, and the average image file size 13.9 kbytes. As `.html` files often contain embedded links to image files, the study's observed data type ratio of 7 times more image files than `.html` files could help determine the average received document size. However all images are not referenced from within documents and it is probable that the separately retrieved images are significantly larger, which inflates the average size. The WWW traffic measurements in [8] report a median burst size of 6 kbytes and a 90-percentile of 50 kbytes.

For distributed file systems, the transferred amount depends on the file length and the file system data chunk size. In NFS [14] the default data chunk size is 8 kbytes, while in the more modern distributed file system AFS-3 it is increased to 64 kbytes. Measurements by Spasojevic and Satyanarayanan [11] on AFS file transfers (excluding RPC overhead) show a trimodal data chunk distribution with 20% of data chunks 0 – 128 bytes, 34% with 1 – 8 kbytes, and 29% with 32 – 128 kbytes. (The daily per client average amounts of fetched and stored data were 33 and 7 Mbytes.)

Using these observations as base figures, we choose a bimodal data chunk length distribution for $L3$, with equal probabilities for a 6.5 kbytes and a 62.5 kbytes data chunk. With the above IP packet size of 1.5 kbytes, this would result in $N \in \{7,44\}$ acknowledgments (includes one additional TCP tear-down packet). We observe that at a link speed of 128 kbps, transferring the larger amount of data takes 4.3s, which seems within an acceptable time span for interactive sessions.

1.4 DISCUSSION

Although the model is based on the TCP/IP protocol suite, we believe that the mechanisms in other data protocols are fairly similar. It can be used for efficiency studies of algorithms for power control, admission and dynamic channel allocation. It can also be a basis for evaluation of algorithms for link speed renegotiation.

The proposed model is fairly simple and surely does not capture all aspects of future packet data services. It is important to remember the rather narrow service description we imposed in the beginning. In an electronic publishing trial reported in [1], the file length of a single scientific article in the PDF format was in the range 150 to 500 kbytes. These figures were obtained after extensive processing and compression of the original (much larger) postscript files. It is clear that the greater data content is not consistent with our model. Due to access control requiring manual entry of passwords, the traffic profile could also be markedly different. Further work could perhaps extend the modeling methodology in order to devise models for electronic payments, electronic publishing and remote terminal or graphical interface emulation.

An obvious model simplification is the modeling of the constant round trip delay. Network congestion, file server load and differences in physical distances to servers, all contribute to large variability in this parameter. The most difficult parameter to

set was the amount of transmitted data. As the model is for wireless services that do not yet exist, verification against measurement data is unfortunately not possible. A reason for not setting the read-write ratio α too high is that new WWW based services could potentially require content transmission also in the uplink.

The suggested model is suitable for modeling light to moderately loaded systems where performance quality can be maintained. If transfer delays increase, it is probable that the user will alter his or her behavior, either by aborting the session, or tending to choose shorter files (less images). If high loads persist, the retrieved content may also be adapted to the poor network performance by increased compression and/or shorter data and image files. This interaction between provided service and observed traffic is a great difficulty when specifying traffic models. However, the objective of system design/optimization must be to provide adequate quality for popular services. Otherwise the customer may choose an alternative service provider.

1.5 REFERENCES

[1] J. Brassil, A. Choudhury, D. Kristol, A. Lapone, S. Low, N. Maxemchuk, L. O'Gorman, "SEPTEMBER - Secure Electronic Publishing Trial", IEEE Communications Mag., Vol. 34, No. 5, May 1996.

[2] R. Caceres, P. Danzig, S. Jamin, D. Mitzel, "Characteristics of Wide-Area TCP/IP Conversations", Proc. of ACM/Sigcomm, 1991.

[3] C. Cunha, A. Bestavros, M. Crovella, "Characteristics of WWW Client-based Traces", Tech. Report BU-CS-95-010, Computer Science Dept., Boston Univ., July 18, 1995.

[4] S. Deering, R. Minden, "Internet Protocol Version 6 (IPv6) Specification", RFC 1883, Xerox PARC, Dec. 95.

[5] R. Gusella, "A Measurement Study of Diskless Workstation Traffic on an Ethernet", Trans. on Com., Vol. 38, No. 9, Sept. 1990.

[6] D. Johnson, C. Perkins, "Mobility Support in IPv6", `draft - ietf - mobileip - ipv6 - 01 . txt` (work in progress), 13 June, 1996.

[7] R. Jain, "Packet Trains-Measurements and a New Model for Computer Network Traffic", IEEE Jour. on Sel. Areas in Comm., Vol. 4, No, 6, Sept. 1986.

[8] Tapani Nieminen,"Report on WWW-traffic measurements", Report T39, Communications Laboratory, Helsinki Univ. of Technology, Oct. 1996.

[9] C. Perkins, Editor, "IP Mobility Support", IETF RFC 2002, Oct 1996.

[10] Vern Paxson, Sally Floyd, "Wide Area Traffic: The Failure of Poisson Modeling", IEEE/ACM Trans. on Netw., Vol. 3, No. 3. June 1995.

[11] M. Spasojevic, M. Satyanarayanan, “An Empirical Study of a Wide-Area Distributed File System”, Technical Report, Transarc Corporation, Nov. 1994. `/afs/cs.cmu.edu/project/coda/Web/docdir/wadfs–nov94.ps.Z.`

[12] J. Postel, “Transmission Control Protocol”, IETF std 0007, also RFC 0793, September 1981.

[13] J. Postel, “Internet Protocol”, IETF std 0005, September 1981.

[14] “NFS: Network File System”, Version 3 Protocol Specification, Sun Microsystems, Mountain View, CA, Feb. 1994.

[15] T. Berners-Lee, R. Fielding, M. Frystyk, “Hypertext Transfer Protocol – HTTP/1.0”, IETF RFC 1945 (informational document), May 1996.

OPTIMAL PAGING OVER IMPERFECT WIRELESS LINKS

Markku Verkama

Nokia Telecommunications
P.O. Box 300, 00045 Nokia Group
Finland

markku.verkama@ntc.nokia.com

Abstract: This paper addresses the mobile phone location problem when paging is prone to errors because of imperfect wireless links. Such a possibility can make it meaningful to repage some cells after a failed paging attempt. Previous studies on paging and location management have only dealt with cases where paging failures do not occur at all or are equally likely in all cells. This study considers the more realistic situation where the likelihood of failure varies within the network service area. Optimal paging strategies are derived using techniques from search theory. Contrary to previous work based on similar approach? the optimal strategies are implicitly bounded in delay through a design parameter, which reflects the cost of a lost call. In practice this bounds the number of repaging attempts to be made in each cell. The paper proposes how to use and implement these results in systems where location management is based on location areas. The estimation of the search model parameters is considered as well.

1 INTRODUCTION

The objective of using radio resources efficiently has motivated numerous studies on location management in wireless personal communication systems, especially on the optimization of the signaling traffic caused by location update and paging messages over the wireless links, see e.g. [1]. The focus of this paper

is on the paging traffic when the wireless link is imperfect and the quality of these links is different in different cells.

The majority of the work on optimal paging has assumed the wireless link to be perfect so that no errors occur in the paging procedure; that is, the mobile units always receive the paging messages correctly and their responses in turn are received correctly by the system. In practice, this assumption does not hold. For example, mobile units may fail to receive the paging message if they happen to experience excessive interference or are momentarily out of coverage. Yet some of these calls could be carried if paging succeeded. Events where the mobile unit does not respond to paging will be called paging failures in this paper. In live networks as many as 10–20 per cent of first paging attempts may fail (this figure includes, however, also call attempts to mobile stations that are unreachable).

Existing systems that use the location area approach for location management, such as GSM [7], tackle paging failures simply by repeating the paging message in the whole location area. Repaging can be done automatically or only if the mobile unit does not respond within a given time period. The choice between these two modes is not obvious and affects for example which network element should be responsible for repaging. The number of repaging attempts is typically limited to two or three.

The problem addressed in this paper is how to page a mobile unit optimally within a given location area and take into account the possibility of paging failures. The approach taken is to weigh the risk of not finding the mobile unit against the cost of paging. This is done via an optimization model, which is solved using techniques from search theory. The optimal strategy turns out to be sequential and of the same form as in [2], [8] and [l0] but it includes additionally a stopping time, which indicates when to stop paging and deem the mobile unit to be unreachable. In addition to the optimal strategy the paper proposes how to use the results in connection with the traditional paging of location areas. This boils down to a scheme where the whole location area is not repaged automatically but only the cells where repaging is deemed to be most useful. Such a scheme could be used with both fixed and dynamic location areas [11].

This study makes use of probability information concerning the location of the mobile unit. In [2], [5], [6], [8] and [l0] a similar approach has been taken. In [6] the benefits of repaging were acknowledged, but the likelihood of a mobile unit not responding to a paging message was the same in all cells. In [2] and [10] this likelihood was allowed to vary between different cells. However, [10] did not consider delay constraints on locating the mobile unit, while [2] proposed to stop the search procedure after a prefixed amount of time has elapsed. The latter approach can increase the risk of not finding the mobile

unit to an unacceptable level, especially if the location information needed in the optimization is inaccurate. In [5] and [8] paging failures were not considered. The contribution of this paper is thus to take into account the tradeoff between paging costs and the risk of losing a call while recognizing that the quality of the wireless link to vary within the location area. Another contribution are the proposed repaging scheme and ways of estimating the search model parameters.

The rest of the paper is organized as follows. The next section discusses reasons for paging failure and how to model them. The optimization model is presented then, as well as ways of determining the model parameters and using the results in existing mobile systems. A numerical example is provided in section 4 to illustrate the methods. Finally, the implications of the work to radio resource management are discussed in section 5.

2 PAGING FAILURES

When the network needs to establish a connection to a mobile unit, the identity of the mobile unit is broadcast on the paging channels of relevant cells. Typically these cells are the ones that form the location area where the mobile unit is registered. When the mobile unit receives the paging message, it contacts the network via the random access channel of the cell where the mobile unit camps on.

Let us assume that the mobile unit is within the theoretical geographic coverage area of some cell when the network starts paging. Paging may fail, i.e., no connection is established between the mobile unit and the relevant base station because of at least the following reasons:

- The mobile unit does not receive the paging message correctly because of interference or noise
- The mobile unit is momentarily out of the coverage area, e.g., in a tunnel
- The mobile unit is switched off without the network knowing it
- The mobile unit does not have enough transmit power to contact the base station
- The base station does not receive the random access attempt correctly because of temporary interference
- Channel assignment fails

The last two cases are not interesting because the mobile unit can simply repeat the access attempt and establish a connection. The first and the second situations are the most important ones because the problem may be solved

by repeating the paging message. In the remaining two situations, where the mobile unit has no possibility to establish a connection, repeating the paging message does not help. Upon a paging failure, the network should ideally make the repaging decision based on the reason for the failure. The trouble is, of course, that the network has no way of knowing why paging failed.

We shall use the following simple model to describe paging failures. Given that a connection *can* be established with the mobile unit in a cell, paging failures in that cell are independent random events and have a constant probability of occurring. The conditional part implies that only the failures where the mobile unit experiences interference or is momentarily out of coverage are modeled by this probability. Situations where it is really impossible to establish a connection, such as when the mobile unit is switched off or does not have enough transmit power, are excluded.

It seems reasonable to assume that the results of successive paging attempts are independent as long as interference or noise are the reasons for possible failure. The likelihood of a failure is not necessarily constant but could depend for example on network traffic. Failures due to the coverage reason, on the other hand, are more problematic. Suppose that the mobile unit is in a vehicle driving in a tunnel when it is paged for the first time. If repaging is attempted immediately, the mobile unit is still likely to be in the tunnel. Thus the result of the second paging attempt would not be independent of the first one. As more time elapses between two paging attempts, they are likely to become independent . In practice one wants to keep the time between the successive attempts as short as possible while still allowing sufficient time for the mobile unit to make a successful random access and the response to be relayed to the network element in charge of paging control. It is possible that the model does not capture all reasons of paging failures correctly. Nevertheless, we shall be satisfied with the simple model and recognize that a more thorough analysis and validation with real measurements could be in order.

3 OPTIMAL REPAGING

3.1 The Problem of Optimal Stopping

The situation we consider is the following. There are N cells denoted $i = 1,. . . , N$ where the mobile unit may be located. With each cell i there is associated a probability of paging failure α_i, conditional on that the mobile unit can in fact respond from that cell. The problem is to try to find the mobile unit while simultaneously minimizing the number of paging messages sent (i.e., saving radio resources) and maintaining a good quality of service (i.e., carrying calls whenever possible).

The quality of service and radio resource usage objectives are contradictory. If the network operator does not want to lose any calls, the mobile unit should be paged until it is found, no matter how long this takes. If, on the other hand, one is minimizing the use of radio resources only, one should not page at all. Clearly both extremes are unacceptable and one has to find a compromise. One can view this as an optimal stopping problem: For how long should one page to make the risk of losing a call, which could actually be carried, small enough?

We shall approach this question by modeling it as an optimal search and stop problem [3], [4]. First, we assume that the mobile unit is located in cell i with probability p_i, $i = 1, \ldots, N$. Furthermore, we assume that when paging commences there is an *a priori* probability $p_0 > 0$ that the mobile unit can not be reached. Hence $\sum_{i=0}^{N} p_i = 1$. As the objective is to minimize the number of paging messages sent, we define that sending a paging message in any cell costs one unit. The tradeoff between the quality of service and the usage of radio resources is modeled with a penalty cost C that is charged if the mobile unit is not found. Hence C is a way of measuring the cost of a lost call in terms of radio resources used in the paging.

A search strategy is denoted by $\delta = (\delta_1, \ldots, \delta_s)$ where δ_i is the ith cell to be paged given that the mobile unit has not responded by then and $s \in \{0, 1, \ldots, \infty\}$ is the *stopping time.* The stopping time means that if the mobile unit is not found with s messages, one stops and deems the mobile unit to be unreachable and the penalty cost is charged. The problem is to find the optimal search strategy δ^* and the optimal stopping time s^* that minimize the expected cost of locating the mobile unit.

It turns out that the optimal strategy δ^* is simply the myopic strategy that always pages the cell with the maximum probability of finding the mobile unit, and the whole problem reduces to that of choosing the optimal stopping time [3]. In other words, the optimal strategy δ^* sends the nth paging message in cell i where i yields the maximum of $p'_j(1 - \alpha_j)$, $j = 1, \ldots, N$. Here $p' = (p'_0, p'_1, \ldots, p'_N)$ is the location distribution prior to sending the nth paging message and conditioned on the previous paging messages.

The optimal stopping time can be determined by considering the equivalent problem of determining the set of posterior location probabilities for which the optimal expected cost of the search equals the penalty cost [3]. Let $f(p, \delta)$ denote the expected paging cost when p is the vector of prior location probabilities and δ is used as the search strategy, and let $f(p) = \inf_\delta f(p, \delta)$. The optimal stopping region is then defined by $S = \{p : f(p) = C\}$.

The exact optimal cost function is difficult to solve but a computationally feasible approximation has been presented in [4], [9]. It is possible to approxi-

mate $f(p)$ through a sequence of functions $f_n(p)$ defined as

$$f_0(p) = C, \quad f_1(p) = \min\{C, \min_i[1 + (1 - (1 - \alpha_i)p_i)C]\},$$
$$f_n(p) = \min\{C, \min_i[1 + (1 - (1 - \alpha_i)p_i)f_{n-1}(T_i p)]\} \quad \text{for } n > 1, \tag{1}$$

where $T_i p = [(T_i p)_0, (T_i p)_1, \ldots, (T_i p)_N]$ is the vector of posterior location probabilities after an unsuccessful paging attempt in cell i, given by

$$(T_i p)_j = \begin{cases} p_j/(1 - (1 - \alpha_i)p_i) & \text{for } j \neq i, \\ \alpha_i p_i/(1 - (1 - \alpha_i)p_i) & \text{for } j = i. \end{cases} \tag{2}$$

The interpretation is that $f_n(p)$ gives the minimum expected cost when at most n steps are allowed. One can use this to approximate the optimal stopping time by letting

$$s_n = \min\{k : f_n(p(\delta^*, k)) = C\}, \tag{3}$$

where $p(\delta*, k)$ is the vector of posterior location probabilities after k unsuccessful searches with the strategy $\delta*$.

The numerical example in section 4 demonstrates how to apply this technique in practice. For practical purposes it seems sufficient to use the approximation with $n = 1$ or $n = 2$.

3.2 Estimation of the Parameters

To be able to use the model one must of course determine the location probabilities and failure probabilities. Let us consider how to determine them when measurements can be made from one fixed location area.

The overall *a priori* failure probability can be estimated simply as the ratio of the number of failed mobile-terminated call attempts to the number of all mobile-terminated call attempts in a given location area, i.e.,

$$\hat{p}_0 = \frac{\#\text{ of failed call attempts}}{\#\text{ of all call attempts}}. \tag{4}$$

Here failed call attempts refer to situations where a mobile unit, which is supposed to be reachable in the location area, does not respond to paging. Data should be gathered from each location area individually.

Once $\hat{p}_0$ is estimated the other location probabilities could simply be taken as

$$\hat{p}_i = (1 - \hat{p}_0)/N, \quad i = 1, \ldots, N. \tag{5}$$

This distribution reflects complete uncertainty as to the whereabouts of the mobile unit within the location area. This level of accuracy is sufficient for the

repaging scheme that will be presented in section 3.3. More accurate location information would be useful especially if separate paging groups and sequential paging were used [5], [6], [8]. An alternative to (5) would then be to use the location accuracy matrix method proposed in [6]. A third alternative is to determine how often mobile-terminated calls have been answered from each cell and use the respective frequencies as location probabilities. These figures would have to be scaled to accommodate the *a priori* failure probability $\hat{p}_0$. The more refined location probability estimates are likely to depend on the time of day, and therefore one would ultimately have different paging strategies for different times of day.

Let us then consider the cell-dependent paging failure probabilities α_i. Suppose that M mobile-terminated call attempts arrive to mobile units in the location area over a given period of time. Of these, theoretically M_i are to mobile units that camp on cell i and can be paged successfully. Given the model of paging failures used, on the average in $(1 - \alpha_i) M_i$ cases the first paging message will be answered, and in $\alpha_i(1 - \alpha_i)M_i$ cases the mobile unit answers after the second paging message. Let us denote the actual numbers, which can be collected from switch statistics, with m_{i1} and m_{i2}; that is, the first paging message has been answered m_{i1} times from cell i and the second one m_{i2} times. Asymptotically these will approach their respective expected values, and hence one can estimate

$$\hat{\alpha}_i = m_{2i}/m_{1i}. \tag{6}$$

If the failure probabilities are not constant but depend on the traffic in the network, say, the estimation would have to be made over periods of similar traffic and one would obtain different estimates for, say, busy hours and quiet hours. Note that if $m_{2i} = 0$, so that there are no recorded events of the second paging attempt having resulted in success in cell i, then $\hat{\alpha}_i = 0$. Hence, to enable the estimation of the failure probabilities one must initially operate the network so that at least one repaging attempt is made in every cell after the first attempt has failed. The measurement period has to be sufficiently long to collect enough data from all cells.

3.3 Implementation Issues and a Repaging Scheme

Suppose that the *a priori* location probabilities $p_0, p_1, \ldots, p_N$ and the failure probabilities $\alpha_1, \ldots, \alpha_N$ have been estimated for the location area, and one chooses some penalty cost C. One can then compute the optimal search strategy and an approximation of the optimal stopping time, s_n, for sòme n. If the penalty cost is small, e.g., if $C = 0$, then it is clearly optimal to stop immediately and not to page at all. When the penalty cost increases, it becomes worthwhile to search for a longer and longer period of time. Thus, one can

tune the paging behavior by changing C. Once the value of C has been fixed for some location area, the same value can be used in the whole network to guarantee a consistent overall quality of service.

The real optimal paging strategy is sequential. It may thus lead to an unacceptably large delay in finding the mobile unit even if the penalty cost bounds the delay of the whole strategy implicitly. Therefore, we propose an alternative strategy based on modifying the traditional blanket paging of the whole location area. This strategy has two different implementations depending on the capabilities of the network element in charge of repaging.

Let $\delta* = (\delta_1^*, \ldots, \delta_{s_n}^*)$ be the optimal strategy. From this one can calculate the quantities $r(1)$, …, $r(N)$ where $r(i)$ is the number of occurences of cell i in $\delta*$. The interpretation is that $r(i)$ gives the maximum number of paging attempts in cell i. In practice $\delta*$ will be such that $r(i) \geq 1$ for all cells and in particular, $r(i) = 1$ for cells with $\alpha_i = 0$. In other words, all cells will be paged at least once, and those cells where repaging is not useful will not be paged more than once. If the network element that handles repaging is capable of relating paging reposponses to paging requests, the modified blanket paging proceeds as follows. First the whole location area is paged. If the mobile unit does not answer within a predefined time period, one pages next in cells where $r(i) \geq 2$. If even this fails, one continues paging in all cells with $r(i) \geq 3$ and so on. A more simple scheme must be used if binding of paging responses to requests is not possible in the entity that handles repaging. An example is the base station in the GSM system. Then repaging takes place "blindly" but the number of messages sent in each cell varies according to the quantities $r(i)$, which would be stored in respective network elements. On base station level the blind repetitions could be omitted if the paging channel is congested.

The benefit of this scheme is that it can be easily implemented in existing systems that use location areas. Other alternatives still compatible with location areas exist too. For example, [6], [8], and [5] have studied schemes where location areas are partitioned into smaller paging groups that are paged sequentially. Such schemes can result in considerable reduction of paging traffic. One option would be to use these schemes as such as the first step. If the mobile unit fails to respond, the schemes would be repeated but with the cells with $r(i) = 1$ removed and so on. Alternatively blanket paging over the cells with $r(i) \geq 2$ could be used.

In practical implementations one can calculate the quantities $r(i)$ in advance for each location area if the *a priori* location distribution is static, but also on-line calculation is possible. The values can be stored in the network entity responsible for paging control. For example, in GSM this would be the mobile switching center (MSC) if the more intelligent repaging scheme were used.

To implement the simple repaging scheme proposed here, MSC can run the following algorithm for every mobile-terminated call attempt:

```
Counter := 1
init:
PageList := nil
for all cells in the location area
    if cell.MaxPageCount ≥ Counter then
  add cell to PageList
if PageList ≠ nil then
    Page all cells in PageList
    Wait for response
    if the mobile unit responded then
  Continue call setup; abort paging
    else
  Counter := Counter+1
  goto init
    endif
else
    The mobile unit is not reachable; abort paging
endif
```

In GSM, MSC would initiate paging in the relevant cells by sending a message to their respective base station controllers using the BSSMAP protocol [7]. The cell identities in the PageList would be given as parameters of the BSSMAP PAGING message.

To estimate paging failure probabilities using equation (6) MSC would have to store for each successful mobile-terminated call the identity of the cell from which the mobile unit responds and the current value of the counter from the above algorithm. As noted earlier, MaxPageCount should be large enough for all cells during the estimation. In GSM the cell identity is carried to MSC in the BSSMAP COMPLETE LAYER 3 INFORMATION message so no changes in the GSM system specifications would be needed for the implementation.

4 EXAMPLE

Let us consider a fictitious numerical example to illustrate the method. The location area in question comprises five cells. Suppose that the *a priori* probability of the mobile unit being unreachable is 5 percent and let us use the uniform location probability distribution. Thus $p_0 = .05$ and $p_i = (1 - .05)/5 = .19$ for $i = 1, \ldots, 5$. Furthermore, we shall assume that $\alpha_1 = \alpha_2 = \alpha_5 = 0$, $\alpha_3 = .01$, and $\alpha_4 = .05$. In other words, repaging has not been successful in cells 1, 2, and 5 during the period when the paging failure probabilities have been estimated.

Table 1 The paging strategy for the penalty cost $C = 1000$.

k	p_0	p_1	p_2	p_3	p_4	p_5	δ_k^*	$f_2(p)$	Action
1	.05	.19	.19	.19	.19	.19	1	622	Page in 1
2	.062	0	.235	.235	.235	.235	2	533	Page in 2
3	.081	0	0	.306	.306	.306	5	392	Page in 5
4	.116	0	0	.442	.442	0	3	144	Page in 3
5	.207	0	0	7.85E-3	.785	0	4	218	Page in 4
6	.814	0	0	3.09E-2	.155	0	4	824	Page in 4
7	.955	0	0	3.63E-2	9.07E-3	0	3	957	Page in 3
8	.990	0	0	3.76E-4	9.41E-3	0	4	992	Page in 4
9	.999	0	0	3.8E-4	4.75E-4	0	4	1000	Stop

From these figures it is possible to calculate the optimal search order. Note that the order is independent of the penalty cost, which only affects the stopping time. For the first step the quantities $(1 - \alpha_i)p_i$ are calculated and compared. The maximum is given by $i = 1$, 2, and 5, and the optimal strategy can thus begin by paging in any one of them. We choose cell 1 first. If the mobile unit does not respond to the paging message, the location probabilities are updated using equation (2). The result is shown in Table 1, which shows the course of the optimal paging strategy δ^*. Note that since the conditional failure probability is zero in cell 1, the posterior probability $p'_1 = 0$ after an unsuccessful paging attempt. The same holds for cells 2 and 5 too. Figure 1 illustrates the behavior of the location probabilities graphically during a sequential search.

The stopping time is obtained simultaneously when the optimal search strategy is calculated. Let us use the approximation $f_2(p)$ and take $C = 1000$. At each stage of the search one can calculate $f_2(p)$ from equation (1) for the current vector of location probabilities. If $f_2(p) < C$ the cell indicated by the optimal strategy is paged; otherwise, one stops and concludes that the mobile unit is unreachable. The values of $f_2(p)$ and the consequent actions are shown in Table 1 as well. The stopping time is $s_2 = 8$. For comparison, if the penalty cost were $C = 100$ the stopping time would be $s_2 = 7$; the paging order and the location probabilities would be the same as in Table 1.

To implement the scheme proposed in section 3.3 one needs to calculate the quantities $r(i)$. For the penalty cost $C = 1000$ we get $r(1) = r(2) = r(5) = 1$, $r(3) = 2$, and $r(4) = 3$ from Table 1. Hence one would first page the whole location area. Upon failure paging would be repeated in cells 3 and 4, and if necessary, final repaging would take place in cell 4.

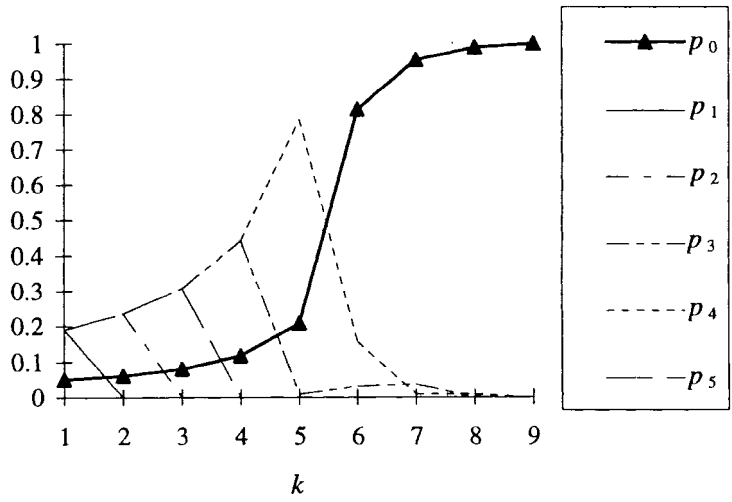

Figure 1 The behavior of the location probabilities during the sequential search.

5 RADIO RESOURCE MANAGEMENT

This section discusses the benefits of the presented results from the point of view of saving radio resources. The key of the paging scheme proposed in section 3.3 is to take into account the differing quality of the wireless control links. This quality is reflected in the failure probabilities α_i.

The application of the results is expected to yield concrete benefits in situations where cells are different with respect to the paging failure probability. Then a varying number of messages may be sent in different cells, see Figure 2. This means that for every mobile-terminated call attempt, the expected number of paging messages varies between the cells. Given the anticipated mobile-terminated traffic, the expected messages turn into requirements concerning the paging channels, which must have enough capacity to carry the paging traffic. In other words, the results can be used as input when planning the capacities of the paging channels. In FDMA or TDMA systems this would typically mean dedicating a different portion of the available radio resources for paging in each cell. In CDMA systems, where such explicit allocations are not necessarily made, the benefits would be in reducing the interference level slightly in the network.

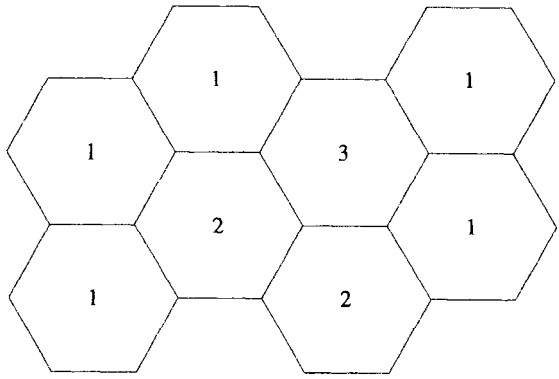

Figure 2 The maximum number of paging attempts depends on the cell.

Sometimes another dimensioning approach can be feasible in FDMA or TDMA systems. By identifying the cells with the smallest failure probabilities and by dimensioning all paging channels according to the requirements of these cells, one can isolate paging delays due to channel congestion to the cells with worse link quality. For example, in the situation of Figure 2 one could dimension all paging channels as if only one message per call attempt were needed. Then congestion and delays would be limited to cells where actually more than one message may be needed.

The concrete benefits and their extent depend of course on the parameters of the network. If all cells in a location area are equal in terms of paging failure probability and if the location probability distribution is uniform, the paging scheme proposed here is no different from the traditional one. Then the benefit of the analysis is only to give some degree of understanding of the tradeoff between the paging cost and the risk of losing calls unnecessarily.

6 CONCLUSION

This paper has dealt with the mobile phone location problem when the mobile phone may fail to respond to paging because of interference or other reasons. Optimal paging strategies have been derived using search theory. The optimal strategy is sequential and pages cells in decreasing order of probability of finding the mobile unit. Moreover, the optimal strategy includes a stopping time when to stop paging and conclude that the mobile unit is unreachable. This is the main difference to previous work on sequential strategies [2], [8], [10].

Methods to estimate the model parameters have been presented as well. The work concerns of course not only paging due to mobile-terminated speech call attempts, but paging due to any mobile-terminated connection establishment needs, such as delivery of mobile-terminated short messages.

The paper has also proposed a simple paging scheme where the optimal number of repaging attempts is taken into account in the traditional blanket paging of the whole location area. Implementation issues concerning this strategy have been discussed. The concrete benefits of optimizing the number of repaging attempts would have to be evaluated on the basis of real network data and would probably vary from network to network. Such an evaluation has been out of the scope of the present paper.

References

[1] Akyildiz, I. F. and Ho, J. S. M. "On location management for personal communications networks," IEEE Communications Magazine, vol. 34, no. 9, pp. 138–145, September 1996.

[2] Awduche, D. O., Ganz, A. and Gaylord, A. "An optimal search strategy for mobile stations in wireless networks," in Proc. 5th IEEE International Conference on Universal Personal Communications, Cambridge, MA, September 29 - October 2, 1996, Vol. 2, pp. 946–950.

[3] Chew, M. C., Jr. "A sequential search procedure," Annals of Mathematical Statistics, vol. 38, pp. 494–502, 1967.

[4] Chew, M. C., Jr. "Optimal stopping in a discrete search problem," Operations Research, vol. 21, pp. 741–747, 1973.

[5] Goodman, D., Krishnan, P. and Sugla, B. "Minimizing queueing delays and number of messages in mobile phone location," Mobile Networks and Applications, vol. 1, pp. 39–48, 1996.

[6] Madhavapeddy, S., Basu, K. and Roberts, A. "Adaptive paging algorithms for cellular systems," in Proc. Fifth WINLAB Workshop on Third Generation Wireless Information Networks, 1995, pp. 347–361.

[7] Mouly, M. and Pautet, M.-B. *The GSM system for mobile communications.* Mouly and Pautet, 1992.

[8] Rose, C. and Yates, R. "Minimizing the average cost of paging under delay constraints," Wireless Networks, vol. 1, pp. 211–219, 1995.

[9] Ross, S. M. “A problem in optimal search and stop,” Operations Research, vol. 17, pp. 984–992, 1969.

[10] Verkama, M. “Optimal paging — A search-theory approach,” in Proc. 5th IEEE International Conference on Universal Personal Communications, Cambridge, MA, September 29 - October 2, 1996, Vol. 2, pp. 956–960.

[11] Xie, H., Tabbane, S. and Goodman, D. “Dynamic location area management and performance analysis,” in Proc. IEEE Vehicular Technology Conference, Secaucus, NJ, 1993, pp. 536–539.

LOCATING MOBILE STATIONS WITH NON-LINE-OF-SIGHT MEASUREMENTS

Marilynn Wylie-Green[1] & Jack Holtzman
Wireless Information Network Laboratory
Rutgers University

1 Problem Introduction

Abstract

We consider the problem of locating mobile stations using range measurements from several base stations when the direct path from transmitter to receiver may be intermittently or, quite possibly, continuously blocked. This non-line-of-sight phenomenon is known to be a major source of error in position location because it systematically causes the mobile to appear farther away from the base station than it actually is. In this paper, we present a first order model for the non-line-of-sight error and show that, under certain conditions, it is possible to (1) detect the non-line-of-sight error using a hypothesis testing procedure and to (2) reconstruct the line-of-sight by exploiting a priori knowledge of the statistics of the standard measurement noise. Simulation examples are presented in order to demonstrate performance of this algorithm.

There are a number of possible commercial applications for mobile position estimation, including fleet management and vehicle navigation [1]. A critical application of position location in cellular systems is in its use with Enhanced-911 (E-911), which is designed to locate the caller during an emergency. Although this service is widely available to wireline subscribers, it is not currently an integral part of the 911 services offered to wireless callers. For this

[1] Marilynn Wylie-Green is currently a Research Engineer with Nokia Research Center, and a Visiting Scholar at WINLAB.

reason, the FCC has required that by the end of 2001, that wireless service providers have the capability to locate the caller in 2-D with a horizontal precision of 125 meters 67% of the time [2]. This mandate, coupled with the demand for location-dependent services in wireless networks, has motivated industry-wide interest into the problem of mobile location.

One method for locating a mobile terminal requires measuring the times of arrival of the radio signal sent by the mobile at a minimum of three base stations. The travel time of the signal at any particular base station can be modeled as:

$$R = c\,T \tag{1}$$

where R is the range, c is the speed of light and T is the one-way travel time of the radio signal. Each distance measurement generates a circle which is centered at the measuring base station and which has a radius equal to the range. The location of the mobile can be estimated by using, for example, the least squares estimate of the intersection of three or more of these circles. However, in a dense urban environment, due to reflection and diffraction, the first detectable wave may actually travel *excess* path lengths on the order of hundreds of meters [3]. This non-line-of-sight (NLOS) problem has been recognized by others as a critical issue, possibly a "killer issue" for mobile location [4]. Thus, accurate location will require algorithms that are robust to the NLOS error.

Silventoinen and Rantalainen [4] conducted simulations and found that the mean absolute location error increased linearly with the increase in the mean of the NLOS errors. Caffery and Stuber [5] developed a constrained location algorithm to mitigate the effects of NLOS range measurements. Mizusawa and Woerner [6] demonstrated the adverse effect of using uncorrected NLOS measurements for mobile location. A survey of positioning techniques for wireless networks can be found in [7].

This paper examines three aspects of location estimation in a NLOS environment. In the first part of our work, we discuss modeling the NLOS error as an AR(1) process. Second, we consider the problem of mode estimation. Each base station measures its own mobile-to-base distance. By operating on segments of the range data to calculate a sample test statistic and exploiting our knowledge of the standard deviation of the standard measurement noise, we can often estimate whether a group of measurements are predominantly LOS or NLOS. The third part of this paper concentrates on LOS reconstruction. If there is strong evidence that the sequence of measurements are NLOS, then they must be corrected prior to location estimation. We show that it is possible to correct the NLOS error when there is some a priori knowledge of the (approximate) support of the standard measurement noise over the real axis.

2 The NLOS Error Model

When a mobile loses LOS with the base station, the travel time of the first detected ray at the base station will be greater than the travel time of a LOS path. This phenomenon is a randomly occurring event whose spatial correlation varies as a function of the mobile's speed and the clutter in the surrounding environment. The NLOS error is defined as:

$$u(k) = \bar{u}(k) + \mu_v \tag{2}$$

where

$$\begin{aligned} \bar{u}(k+1) &= \alpha \bar{u}(k) + (1-\alpha)v(k) \\ v(k) &\sim N(0, \sigma_v^2). \end{aligned} \tag{3}$$

μ_v is a positive bias, characteristic of range measurements in an outdoor environment. Typically, $\mu_v \sim 500m$ [4]. $\alpha = \varepsilon^{(v \bullet T_s / D)}$ denotes the spatial correlation between adjacent samples; where v is the mobile's speed, T_s is the sampling period, D is the decorrelation distance and $\varepsilon < 1$. One can easily imagine that as the mobile moves through the environment at a "slow" speed, its path length error will be highly correlated over small sampling intervals. Conversely, if the mobile station moves at "faster" speeds through the same environment with the same sampling rate, one would expect for the path error to decorrelate sooner. The autoregressive model captures this behavior.

In an urban environment, the mobile may only have LOS with the base station intermittently. The state of either receiving a LOS or a NLOS measurement is modeled as a two state Markov switching model. Switching between states (LOS versus NLOS) is a process that occurs at random intervals and is governed by a set of state transition probabilities which quantify the probability of having LOS or NLOS with the base station during the current sample *given* its state at the last measurement time. (In [8], the NLOS problem is considered for the situation in which the measurements are either continually LOS or continually NLOS). We represent the NLOS error at time t_k as

$$w(k) = z(k)u(k). \tag{4}$$

$z(k)$ is a two state Markov random variable that equals zero/one when the measurement is LOS/NLOS. The state transition probabilities are given as:

$$\begin{aligned} P_L(t_k) &= P[z(k) = 0] \\ P_N(t_k) &= P[z(k) = 1] \\ P_{L|L}(t_k|t_{k-1}) &= P[z(k) = 0|z(k-1) = 0] \\ P_{L|N}(t_k|t_{k-1}) &= P[z(k) = 0|z(k-1) = 1] \\ P_{N|L}(t_k|t_{k-1}) &= P[z(k) = 1|z(k-1) = 0] \\ P_{N|N}(t_k|t_{k-1}) &= P[z(k) = 1|z(k-1) = 1]. \end{aligned} \tag{5}$$

All of the modeling is, of course, subject to verification with real data.

3 Range Measurements

The range measurements at the mth base station (m=1,...,M) at time t_k can be represented as the superposition of three terms:

$$y_m(k) = L_m(k) + n_m(k) + w_m(k). \tag{6}$$

$L_m(k)$ is the mobile-to-base range for the mth base station in 2-D:

$$L_m(k) = \sqrt{(x(k) - b_{mx})^2 + (y(k) - b_{my})^2}, \tag{7}$$

where $(x(k),y(k))$ and (b_{mx}, b_{my}) are the Cartesian coordinates of the mobile and the mth base station (respectively) at t_k. $n_m(k) \sim N(0, \sigma^2_m)$ is the standard measurement inaccuracy. $w_m(k)$ in Equation (4) models the NLOS error at t_k.

4 Mobile Location

Locating a mobile in two coordinates requires the use of range measurements from three or more base stations. One method of solving for the mobile's coordinates is to find the values of $(\hat{x}(k), \hat{y}(k))$ that minimize:

$$S(k) = \frac{1}{M}\sum_{m=1}^{M}\left(y_m(k) - \sqrt{(\hat{x}(k) - b_{mx})^2 + (\hat{y}(k) - b_{my})^2} \right)^2. \tag{8}$$

It is not known a priori which range measurements (if any) contain NLOS errors. One of the major effects of the NLOS error is to positively bias the range measurements, so that it appears that the mobile station is farther away from the base station than it actually is. In the sections following, we discuss the proposed technique for detecting the Markov process state (NLOS versus LOS), then correcting these measurements deemed to be from the NLOS state before location estimation.

4.1 Mode Estimation

Accurate mobile location requires the ability to correctly identify the type of measurements (LOS versus NLOS) that have been received so that (if

necessary), we can correct the error before doing least squares estimation. Naturally, estimating the mode cannot be effectively performed unless we have some a priori knowledge of the behavior of the measurements under LOS conditions so that we can detect statistically significant deviations.

When the mobile is moving, a sequence of LOS measurements is a nonstationary Gaussian random process that can be expressed as

$$y_m(k) = L_m(k) + n_m(k) \tag{9}$$

(m=1,...,M). Its mean value is the unknown LOS distance, $L_m(k)$, and its standard deviation, σ_m, is the standard deviation of the noise, $n_m(k)$. However, when the NLOS error is also present,

$$y_m(k) = L_m(k) + n_m(k) + w_m(k) \tag{10}$$

and the measurements are temporally correlated as well as nonstationary. Assuming that $n_m(k)$ and $w_m(k)$ are uncorrelated with each other, it follows that the mean and variance of $y_m(k)$ will be

$$\begin{aligned} \mu_{y_m(k)} &= L_m(k) + \mu_{w_m(k)} > L_m(k) \\ \sigma^2_{y_m(k)} &= \sigma^2_m + \sigma^2_{w_m(k)} > \sigma^2_m, \end{aligned} \tag{11}$$

where

$$\begin{aligned} \mu_{y_m(k)} &= E\{y_m(k) \mid L_m(k)\} \\ \mu_{w_m(k)} &= E\{w_m(k)\} \\ \sigma^2_{w_m(k)} &= E\{(w_m(k) - \mu_{w_m(k)})^2\}. \end{aligned} \tag{12}$$

Hence, the two major effects of the NLOS error are to *increase* the mean and the variance of range measurements. Abrupt changes in the measurements that can be easily detected by inspection really reduce to detecting when $\mu_{y_m(k)}$ either increases/decreases significantly. However, we are also interested in those cases in which the change is not readily apparent or when all of the measurements are either NLOS or LOS and, therefore, no change has occurred. Our strategy in estimating the Markov process is to find a test statistic that behaves statistically similarly to the additive white Gaussian noise, $n_m(k)$, when conditions are LOS so that we can detect deviations from the LOS condition.

The difficulty in developing a test for jump detection in this problem is due to the fact that the range measurements are generally nonstationary. When the

data are stationary, and the means and or variances before and after the jump are *known*, then the problem of jump detection reduces to a hypothesis testing problem using likelihood ratios [10]. However, in this case, since the LOS distance is the unknown mean value and it is changing with time, we cannot directly apply the standard likelihood ratio test.

The proposed technique for detecting a jump in the data is based on using a sliding window to fit a second order polynomial locally to the data. Each window contains P samples, and the set of measurements associated with the ith sliding window can be modeled as:

$$\begin{aligned} y_m(i-\overline{P}) &= \left(\sum_{n=0}^{2} a_{nm}(i)t_0^n\right) + n_m(i-\overline{P}) + w_m(i-\overline{P}) \\ &\vdots \\ y_m(i+\overline{P}) &= \left(\sum_{n=0}^{2} a_{nm}(i)t_{P-1}^n\right) + n_m(i+\overline{P}) + w_m(i+\overline{P}). \end{aligned} \tag{13}$$

The window size is $P = 2\,\overline{P} + 1$. Equation (13) can also be expressed as:

$$\mathbf{y}_m(i) = \mathbf{A}\mathbf{a}_m(i) + \mathbf{n}_m(i) + \mathbf{w}_m(i) \tag{14}$$

where

$$\begin{aligned} \mathbf{y}_m(i) &= [y_m(i-\overline{P})\ldots y_m(i+\overline{P})]^{\mathrm{T}} \\ [\mathbf{A}]_{j,k} &= t_{j-1}^{k-1} \\ \mathbf{a}_m(i) &= [a_{0m}(i)\ \ a_{1m}(i)\ \ a_{2m}(i)]^{\mathrm{T}} \\ \mathbf{n}_m(i) &= [n_m(i-\overline{P})\ldots n_m(i+\overline{P})]^{\mathrm{T}} \\ \mathbf{w}_m(i) &= [w_m(i-\overline{P})\ldots w_m(i+\overline{P})]^{\mathrm{T}} \end{aligned} \tag{15}$$

for base station $m=1, \ldots, M$; $i = \overline{P}, \ldots N - \overline{P} + 1$; $j = 1, \ldots, p$; *and* $k=1,2,3$. (N is the total number of samples).

The least squares estimate of $\mathbf{a}_m(i)$ is given by the pseudo-inverse

$$\hat{\mathbf{a}}_m(i) = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y}_m(i). \tag{16}$$

The residual estimation error, which we denote $\mathbf{r}_m(i)$, is then

$$\mathbf{r}_m(i) = \mathbf{y}_m(i) - \mathbf{A}\hat{\mathbf{a}}_m(i). \tag{17}$$

By substituting Equation (16) into Equation (17), we can also express $\mathbf{r}_m(i)$ as

$$\begin{aligned}\mathbf{r}_m(i) &= \left(\mathbf{I} - \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\right)\mathbf{y}_m(i) \\ &= \tilde{\mathbf{A}}\mathbf{y}_m(i),\end{aligned} \tag{18}$$

where $\tilde{\mathbf{A}} = \mathbf{I} - \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\ \mathbf{A}^T$ projects onto the space that is orthogonal to the column space of **A**.

The more accurately one can estimate $\mathbf{a}_m(i)$, the closer $\mathbf{r}_m(i)$ will "resemble" the input noise process and this implies that under ideal circumstances the average squared deviation of the elements in $\mathbf{r}_m(i)$ should be on the order of σ_m^2. A second interpretation of $\mathbf{r}_m(i)$ is that since it is orthogonal to the data, that it is in some sense an "estimate" of the input noise vector and one would expect for it to behave statistically similar to the noise. Following this logic, the property of $\mathbf{r}_m(i)$ that we will exploit for mode estimation is:

$$\hat{\sigma}_m^2 = \frac{1}{P}\mathbf{r}_m^T(i)\mathbf{r}_m(i) \sim O(\sigma_m^2), \tag{19}$$

which is a reasonable assumption, given that the second order polynomial is a good approximation to the behavior of the range over short intervals.

The two possible modes follow:

$$\begin{aligned}M_1 &: \text{predominantly line - of - sight} \\ M_2 &: \overline{\mathrm{M}}_1\end{aligned} \tag{20}$$

and the test that we will use is

$$\begin{aligned}M_1 &: \text{if } \hat{\sigma}_m^2 < T \\ M_2 &: \text{if } \hat{\sigma}_m^2 > T\end{aligned} \tag{21}$$

where T is the threshold now to be determined. We will chose as our test statistic

$$\lambda(\mathbf{r}_m(i)) = \frac{P}{2}\log\left(\hat{\sigma}_m^2(i)/\sigma_m^2\right) \tag{22}$$

Our test rule will reject M_1 for large values of $\hat{\sigma}_m^2(i)/\sigma_m^2$. Using $P\hat{\sigma}_m^2(i)/\sigma_m^2 \sim \chi_P^2$ [11], it follows that

$$\text{Prob}_{H_o(i)|M_1}[P\hat{\sigma}_m^2(i)/\sigma_m^2 > x_P(1-\gamma)] = \gamma, \tag{23}$$

where $x_P(1-\gamma)$ is the $(1-\gamma)^{\text{th}}$ quantile of the χ_P^2 distribution. The decision rule is to reject M_1 when

$$\hat{\sigma}_m^2(i) > x_P(1-\gamma)\sigma_m^2/P. \tag{24}$$

To estimate the state of the mode, we define

$$\bar{z}_P(i) = \begin{cases} 1 \text{ if } \hat{\sigma}_m^2(i) > x_P(1-\gamma)\sigma_m^2/P \\ 0 \text{ if } \hat{\sigma}_m^2(i) < x_P(1-\gamma)\sigma_m^2/P \end{cases} \tag{25}$$

for $i = \overline{P}, \ldots, N-\overline{P}+1$. Following state estimation, it may be necessary to perform LOS reconstruction. In the next section, we define the conditions under which we can correct the NLOS error.

4.2 LOS Reconstruction

When M_1 is rejected at a base station, we have evidence that the NLOS error is (at least) intermittently present within the P sample window of observation. Under that condition, we employ a technique that exploits our a priori knowledge of the statistics of the standard measurement noise, $\mathbf{n}_m$, in order to reduce the positive bias and to increase the accuracy of our position estimate.

The LOS reconstruction technique is best understood by fast examining the statistical properties of the composite error,

$$e_m(k) = n_m(k) + w_m(k). \tag{26}$$

Because $n_m(k) \sim N(0,\sigma_m^2)$, there is a 95% probability that it will assume values between $(-1.96\sigma_m, 1.96\sigma_m)$. The NLOS error, $w_m(k)$, is modeled as a nonnegative random variable which (approximately) ranges over some interval $(0, \beta)$, where $\beta > 0$. By considering these truncated regions of support, one can gain a great deal of insight into the proposed solution for this problem.

The NLOS error is, by definition, nonnegative since it models excess path length and in Equation (2) we modeled it as a first order process. Since $v(k) \sim N(0,\sigma_v^2)$ and μ_v is constant, it follows that the mean and standard variation of $u(k)$ are given, respectively, by

$$\begin{aligned} \mu_{u(k)} &= \mu_v \\ \sigma_{u(k)} &= \sigma_v \sqrt{(1-\alpha)^2/(1-\alpha^2)}. \end{aligned} \tag{27}$$

Since $u(k)$ is Gaussian, we can show that 95% of the values that it can assume range over the region $(\mu_v - 1.96\sigma_{u(k)}, \mu_v + 1.96\sigma_{u(k)})$. Further, given that the NLOS error is present, $e_m(k)$ is also Gaussian, with an approximate region of support over $\left(-1.96\sigma_m + \mu_v - 1.96\sigma_{u(k)}, 1.96\sigma_m + \mu_v + 1.96\sigma_{u(k)}\right)$.

Suppose the measurements over a P sample window were fit to a second order polynomial to generate the smoothed curve, $s_m(k)$, and that the major effect of the NLOS error was to positively bias this curve. Let $\delta_m(k) = s_m(k) - y_m(k)$. Then, the LOS reconstruction technique is to

- Find the time, $k = K_0$ at which $\delta_m(k)$ is maximum. $y_m(K_0)$ will be the point of maximum deviation of the measured range from the smoothed curve and:

$$y_m(K_0) \approx L_m(K_0) - 1.96\sigma_m + \mu_v - 1.96\sigma_v \sqrt{(1-\alpha)^2/(1-\alpha^2)}. \tag{28}$$

- Calculate $\tilde{s}_m(k) = s_m(k) - \delta_m(K_0)$.
- Displace $\tilde{s}_m(k)$ to generate the final LOS estimate $\hat{L}_m(k) = \tilde{s}_m(k) + 1.96\sigma_m$.

Using this technique, the final bias will be $\mu_v - 1.96\sigma_v\sqrt{(1-\alpha)^2/(1-\alpha^2)}$ which is less than the original bias, μ_v. This technique will reduce the effect of the bias for all K_0. In particular, if $\mu_v \sim 1.96\sigma_v\sqrt{\frac{(1-\alpha)^2}{(1-\alpha^2)}}$, then the bias will have been removed. However, if the degree of spatial correlation between adjacent samples approaches unity i.e., $\alpha \to 1$, then the NLOS error will approach a constant bias which will be unresolvable unless one has access to a series of LOS measurements and can detect a jump when conditions change from LOS to NLOS (or vice versa).

For simulation purposes, we have assigned numerical values to σ_m, σ_v, and β based on the real measurements reported in [4]. In [4], the support of the composite errors is (-400, 1700) meters. From this one could

argue that if the standard measurement noise is symmetrically distributed about the origin, then its support ranges from (-400,400) meters. This implies that the NLOS error ranges in value from (0, 1300) meters and that $\beta = 1300$ meters, $\sigma_v = 330$ meters and $\mu_v = 647$ meters. Based on the relation $400 = 1.96\sigma_m$, we have that $\sigma_m = 204$ meters.

In the following examples, we illustrate the LOS reconstruction algorithm for different decorrelation distances, D, for the case in which the NLOS error is continuously present. The numerical values chosen are given in Table 1.

b_{lx}	b_{ly}	$x(k)$	$y(k)$	σ_v	μ_v	$\sigma_1 = \ldots \sigma_M$
0	0	$100+30k$	$-100+10k$	330	647	150

Table 1: The simulation parameters used in Figures (1-6).

All distances are expressed in meters, time in seconds. The value of D changes in each example from D = 30, 60, 120, respectively, in order to demonstrate the effect of the increased correlation on the performance of the LOS reconstruction algorithm. Results from the first simulation (D = 30), shown in Figures (1-2), indicate that if the correlation is small that it is possible to reconstruct the LOS. Results from the second example (D=60) are reported in Figures (3-4) and show how the increased correlation has begun to adversely affect the reconstruction algorithm. This effect is even more pronounced in the third example (D = 120) (see Figures (5) and (6)).

4.3 2-D Mobile Location

In this section, we discuss 2-D location of the mobile. Following mode estimation at each base station, there are two possible cases. For case (1), the majority of the measurements have been declared LOS, while in case (2), the majority of the measurements have been declared NLOS. For case (1), location of the mobile is straightforward. The measurements can be first smoothed using a polynomial fit in order to reduce the effect of any undetected NLOS measurements and then the smoothed fit can be used along with the (smoothed) range measurements from other base stations to find the least squares position estimate. In case (2), when a clear majority of the measurements have been declared NLOS, then we can use the LOS reconstruction technique in order to reduce the effect of the bias. After error correction, we will then use these measurements, along with the measurements from other base stations, to estimate position.

In each of the following simulations, all of the measurements were treated as LOS/NLOS if over 60% of them were declared to be LOS/NLOS, respectively.

5 Simulations

In all of the following simulations, the mobile -base station geometry is the same as is shown in Figure (7). The mobile's coordinates were given by

$$\begin{aligned} x(k) &= -10 + 10k \\ y(k) &= -200 + 10k \end{aligned} \tag{29}$$

for k=0,...,200 and the base stations were located at

$$\begin{aligned} (b_{1x}, b_{1y}) &= (0,1500) \\ (b_{2x}, b_{2y}) &= (-700,-500) \\ (b_{3x}, b_{3y}) &= (700,-500). \end{aligned} \tag{30}$$

The mean NLOS error was $\mu_v = 500$, $\sigma_m = 150$ and $\sigma_v = 200$ (m=1,...,3). The spatial correlation was $\alpha = 0.34$.

5.1 Simulation I

In this simulation, we demonstrate the performance of the mode estimation algorithm for the case in which the transition probabilities are given by:

$$\begin{aligned} P_L(t_0) &= 0.5 \\ P_{L|L}(t_k|t_{k-1}) &= 0.85 \\ P_{N|N}(t_k|t_{k-1}) &= 0.8. \end{aligned} \tag{31}$$

The true mode is shown in Figure (8a), while the estimated mode is shown in Figure (8b). The window length, P=10.

5.2 Simulation II

In this simulation,

$$\begin{aligned} P_L(t_0) &= 0.001 \\ P_{L|L}(t_k|t_{k-1}) &= 0.001 \\ P_{N|N}(t_k|t_{k-1}) &= 0.95. \end{aligned} \tag{32}$$

The simulated, true, and corrected ranges are shown in Figure (9), and indicate that the NLOS correction technique was able to reduce the bias in the range. The estimated trajectory of the mobile station is compared to the true trajectory in Figure (10). The window length, P=25.

5.3 Simulation III

We repeated the simulation in Example II over 200 independent trials in order to assess the performance of the location algorithm in the presence of the NLOS error with P=25. In Figure (1l), the mean location estimate is compared to the true location and the average error is well below the mean of the NLOS error, μ_v = 500 meters. As was mentioned in the introduction, the FCC has mandated that 67% or the time, the location estimate should be within 125 meters of the true value. In Figure (12), the ordinate is the mark. The NLOS error, if detected, was corrected and then the corrected measurements were used for estimating location. In most cases, the error appears to be less than twice the upper bound specified by the FCC.

6 Conclusion

We have presented a new tracking algorithm that is capable of discriminating between LOS versus NLOS range measurements and correcting the NLOS error by using a priori knowledge of the approximate support of the noise over the real axis. Simulation results indicate that the positive ranging bias, which is caused by the NLOS error, was reduced by several orders of magnitude after performing the LOS reconstruction technique. The results encourage further investigation into variations of the algorithms and such issues as correlated measurements, other scenarios, sensitivities to parameters, etc., and, of course, investigation with real measurements.

References

[1] M.I. Silventoinen and T. Rantalainen, “Anytime, anywhere... Big Brother is watching you”, *Mobile Europe*, September 1995, pp. 43-50.

[2] FCC Regulation Proposal: Notice of Proposed Rule Making, FCC Docket 94-237, Adopted Date: Sept. 19, 1994. Released Date: Oct. 19, 1994.

[3] M.I. Silventoinen and T. Rantalainen, "Mobile Station Emergency Locating in GSM", *IEEE International Conference on Personal Wireless Communications*, India, February 1996.

[4] M.I. Silventoinen and T. Rantalainen, “Mobile Station Locating in GSM”, *IEEE Wireless Communications System Symposium*, Long Island, NY, November 1995.

[5] J.A. Caffery and G.L. Stuber, “Radio Location in Urban CDMA Microcells”, *Proceedings of the Personal, Indoor and Mobile Radio Communications* (PIMRC ‘95), vol. 2, pp. 858-862.

[6] G.A. Mizusawa and B.D. Woerner, “Performance of Hyperbolic Position Location Techniques for Code Division Multiple Aceess”, Technical Report MPRG-TR-96-29, Virginia Polytechnic Institute and State University, Aug., 1996.

[7] T.S. Rappaport, J.H. Reed, and B.D. Woerner, “Position Location Using Wireless Communications on Highways of the Future”, *IEEE Communications Magazine*, October 1996, pp. 33-41.

[8] M.P. Wylie and J. Holtzman, ‘The Non-Line-of-Sight Problem in Mobile Location Estimation”, *Proceedings IEEE 5th International Conference on Universal Personal Communications*, 1996, pp. 827-831.

[9] D. J. Torrieri, “Statistical Theory of Passive Location Systems”, *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-20, No. 2., March 1984, pp. 183-198.

[10] Y. Liu and S.D. Blostein, “Quickest Detection of an Abrupt Change in a Random Sequence with Finite Change-Time”, *IEEE Transactions on Information Theory*, Nov. 1994, pp. 1985-1993.

[11] P.J. Bickel, and K.A. Doksum, Mathematical Statistics: Basic Ideas and Selected Topics, Holden-Day, Inc., chapter 5.

[12] H.L. Van Trees, Detection, Estimation and Modulation Theory Part I, John Wiley & Sons, New York, 1968, chapter 2.

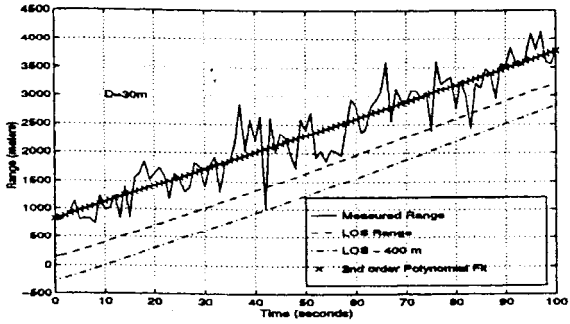

Figure 1: Example I: The measured range is shown, in comparison to the actual LOS distance. The parameter values are D = 30 meters, α = 0.08, σ_v = 330, σ = 204 and μ_v = 647.

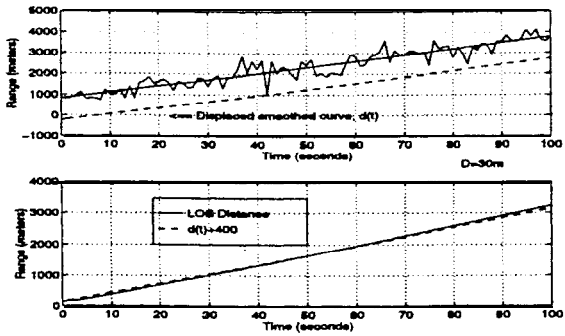

Figure 2: Example I: The LOS is reconstructed by displacing the smoothed curve through the point of maximum deviation and then up by 1.96σ = 400 meters. The parameter values are D = 30 meters, α = 0.08, σ_v = 330, σ = 204 and μ_v = 647.

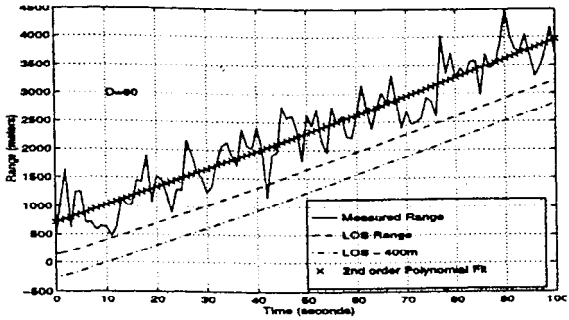

Figure 3: Example II: The measured range is shown, in comparison to the actual LOS distance. The parameter values are D = 60 meters, α = 0.29, σ_v =330, σ = 204 and μ_v = 647.

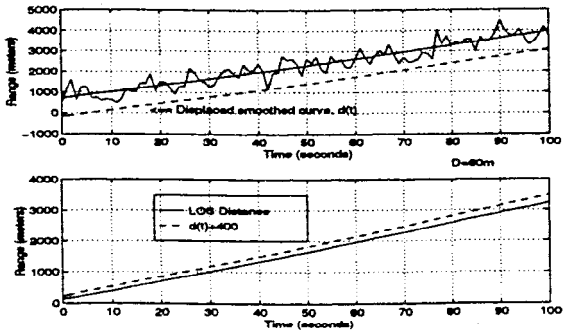

Figure 4: Example II: The line of sight is reconstruction by displacing the smoothed curve through the point of maximum deviation and then up by 1.96σ = 400 meters. The parameter values are D = 60 meters, α = 0.29, σ_v =330, σ = 204 and μ_v = 647.

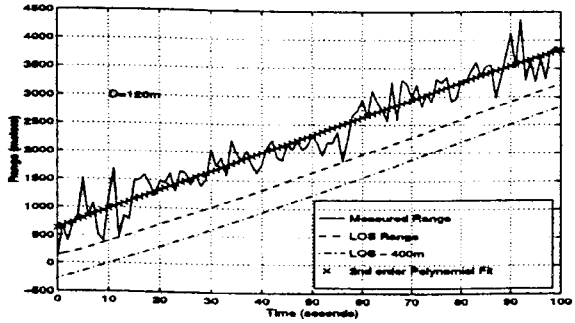

Figure 5: Example III: The measured range is shown, in comparison to the actual LOS distance. The parameter values are D = 120 meters, α = 0.55, σ_v = 330, σ = 204 and μ_v = 647.

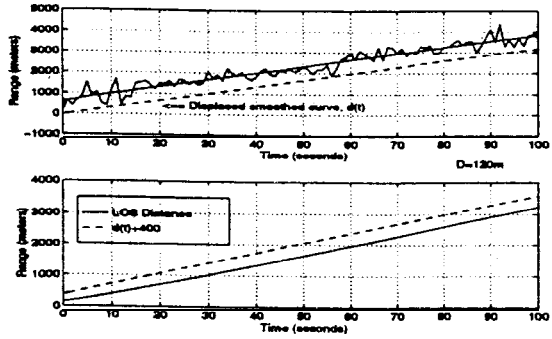

Figure 6: Example III: The LOS is recontructioned by displacing the smoothed curve through the point of maximum deviation and then up by 1.96σ = 400 meters. The parameter values are D = 120 meters, α = 0.55, σ_v = 330, σ = 204 and μ_v = 647.

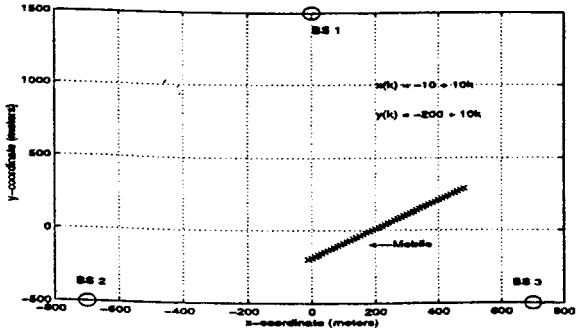

Figure 7: The mobile-base station geometry is shown for Simulations I, II and III. The mobile follows the trajectory $(-10 + 10k, -200 + 10k)$.

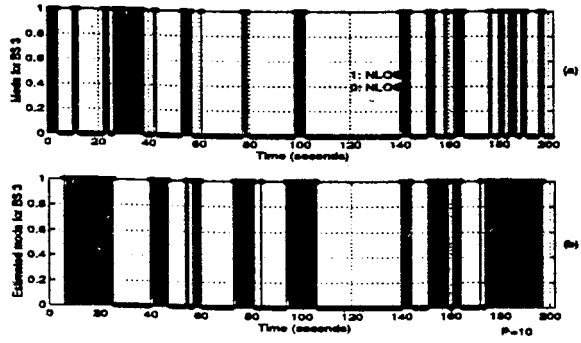

Figure 8: Simulation I: The true mode of the measurements taken by base station 3 is shown in (a) and the estimated mode is shown in (b). $P_L(t_0) = 0.5$, $P_{L|L}(t_k|t_{k-1}) = 0.85$ and $P_{N|N}(t_k|t_{k-1}) = 0.8$, $\mu_v = 500$, $\sigma = 150$, $\sigma_v = 200$, $\alpha = 0.34$ and $P = 10$.

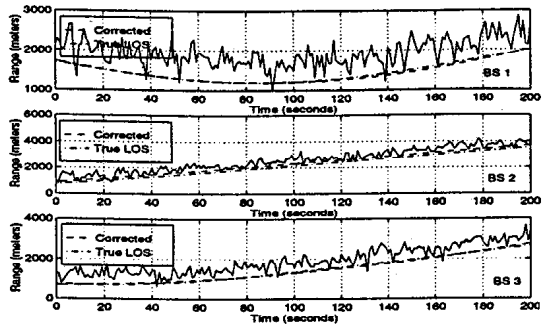

Figure 9: Simulation II: The simulated range, containing the NLOS error, is shown for each base station, along with the true line of sight radial distance and the corrected distance. In each case, the measurements were dominated by the NLOS error. $P_L(t_0)$ = 0.001, $P_{L|L}(t_k|t_{k-1})$ = 0.001, $P_{N|N}(t_k|t_{k-1})$ = 0.95, μ_v = 500, σ = 150, σ_v = 200, α = 0.34 and P = 25.

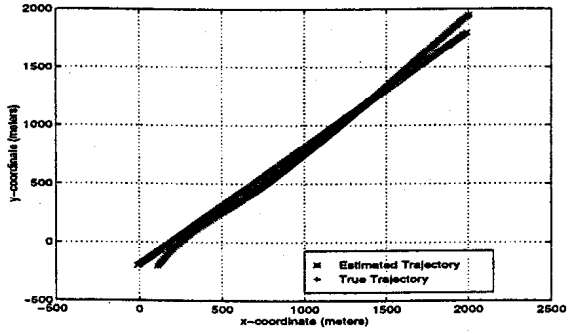

Figure 10: Simulation II: The true and estimated trajectories are shown in the figure above. $P_L(t_0)$ = 0.001, $P_{L|L}(t_k|t_{k-1})$ = 0.001, $P_{N|N}(t_k|t_{k-1})$ = 0-95, μ_v = 500, σ = 150, σ_v = 200, α = 0.34 and P = 25.

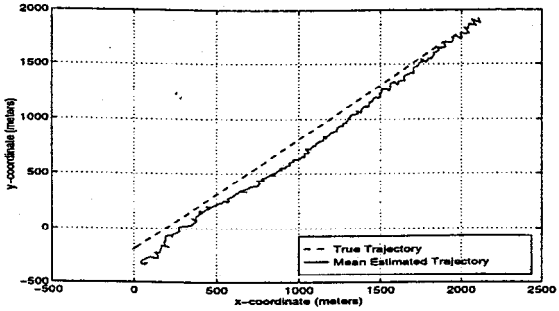

Figure 11: Simulation III: Over 200 independent trials were run and the average trajectory is shown in direct comparison with the true trajectory. $P_L(t_0) = 0.001$, $P_{L|L}(t_k|t_{k-1}) = 0.001$, $P_{N|N}(t_k|t_{k-1}) = 0.95$, $\mu_v = 500$, $\sigma = 150$, $\sigma_v = 200$, $\alpha = 0.34$ and $P = 25$.

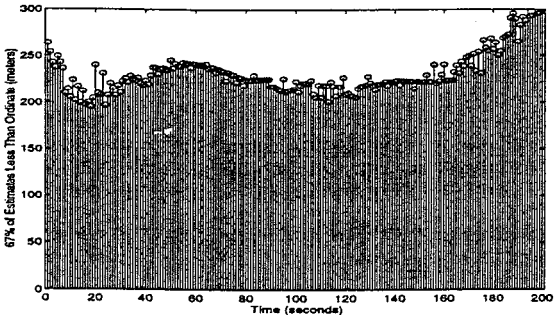

Figure 12: Simulation III: 67% of the location estimates fall below the values indicated on the ordinate as a function of time. The FCC mandate requires that 67% of the location estimates fall within 125 meters of the true radial distance. $P_L(t_0) = 0.001$, $P_{L|L}(t_k|t_{k-1}) = 0.001$, $P_{N|N}(t_k|t_{k-1}) = 0.95$, $\mu_v = 500$, $\sigma = 150$, $\sigma_v = 200$, $\alpha = 0.34$ and $P = 25$.

This page intentionally left blank.

INDEX