Technical University of Budapest
Department of Telecommunications and Telematics

Design and Analysis of Cellular Mobile Data Networks

András Gergely Valkó

High Speed Networks Laboratory
Department of Telecommunications and Telematics
Technical University of Budapest

Ph. D. Dissertation

Advisors

Dr. Tamás Henk
High Speed Networks Laboratory
Department of Telecommunications and Telematics
Technical University of Budapest

Prof. Andrew T. Campbell
Center for Telecommunications Research
Columbia University, New York

Budapest, 1999

Budapesti Műszaki Egyetem

Távközlési és Telematikai Tanszék

Cellás mobil adatátviteli hálózatok tervezése és elemzése

Valkó András Gergely

Nagysebességű Hálózatok Laboratóriuma
Távközlési és Telematikai Tanszék
Budapesti Műszaki Egyetem

Ph. D. disszertáció

Tudományos vezetők

Dr. Henk Tamás
Nagysebességű Hálózatok Laboratóriuma
Távközlési és Telematikai Tanszék
Budapesti Műszaki Egyetem

Prof. Andrew T. Campbell
Center for Telecommunications Research
Columbia University, New York

Budapest, 1999

Contents

1 Introduction 1
1.1 The Cellular Principle 1
1.2 State of the Art 3
1.2.1 The Global System for Mobile Communications 4
1.2.2 Third Generation Cellular Systems 4
1.2.3 Internet Mobility Proposals 5
1.3 Research Objectives 6

2 Hybrid-Hierarchical Simulation Architecture 8
2.1 Problem Statement 8
2.2 Related Work 9
2.3 Simulation Architecture 10
2.3.1 Overview 10
2.3.2 Model 12
2.3.3 Communication and Synchronization 13
2.4 Analysis 14
2.4.1 Implementation Validation 14
2.4.2 Single-link example 14
2.4.3 Network example 16
2.5 Discussion 18

3 An Efficiency Bound of Cellular Mobile Systems 20
3.1 Problem Statement 20
3.2 Related Work 21
3.3 Model 22
3.4 Analysis 23
3.4.1 Efficiency of Systems with Deterministic Advance Reservation 24
3.4.2 Efficiency of Single Application Systems with Statistical Local Admission Control 27
3.5 Discussion 31

4 An Architecture and Protocol for Cellular Wireless Data Networks 35
4.1 Problem Statement 35
4.2 Related Work 37
4.3 Model 39
4.4 Design Principles 41
4.5 Cellular IP 44
4.5.1 System Overview 44
4.5.2 Protocol Details 46

4.5.3 Implementation 50
4.6 Discussion 53

5 Performance Evaluation of Cellular IP Networks 55
5.1 Problem Statement 55
5.2 Related Work 56
5.3 Model 57
5.4 Methodology 58
5.4.1 Simulation Environment 58
5.4.2 Experimental Setting 59
5.5 Analysis 60
5.5.1 Handoff Performance 60
5.5.2 Mobility Management Cost 64
5.5.3 Scalability 70
5.5.4 Service Quality Provisioning 73
5.6 Discussion 77

6 Conclusions 79
6.1 Contribution of Thesis 79
6.2 Future Research Directions 80

Acknowledgments

I would like to express my greatest thanks to my two advisors, Prof. Tamás Henk at the Technical University of Budapest and Prof. Andrew T. Campbell at Columbia University, New York. As head of the High Speed Networks Laboratory in Budapest, Tamás helped my first steps as a PhD student and directed my choice of topic. His encouragement and comments have always meant a lot to me. Andrew's support during my time at the COMET group was also invaluable. Working with him has greatly influenced my thinking, let alone my technical writing style.

A substantial part of my research was financed by Ericsson. I am grateful to Dr. Miklós Boda, head of Ericsson Traffic Analysis and Network Performance Laboratory for having believed that this might be a reasonable investment. Even more important to me was Miklós' help with critical comments and sound judgment throughout the past few years.

Many thanks are due to lots of friends and colleagues at the High Speed Networks Laboratory, at Ericsson Traffic Laboratory and at the COMET group. Special thanks to Gábor Fodor, András Rácz and Zsolt Haraszti for many fruitful discussions and for the fun we had working together. I would like to express my appreciation to Lars Westberg for sharing his innovative and exciting work style with me. Special thanks are also due to Javier Gomez and Sanghyo Kim for their immense contribution to Cellular IP and for the great time we had when nothing seemed to work as planned. And thank you, Emilia.

Összefoglalás

Az egyre kisebb méretű hordozható számítógépek és a vezeték nélküli mobil adatátviteli berendezések elterjedése gyökeresen változtathatja meg a távközlés mai kultúráját. A jövőben az Internet hozzáféréssel rendelkező, zsebben hordható számítógép ugyanolyan mindennapi eszközzé válhat, mint amilyen természetes ma a mobiltelefon. Az Internetet már ma is számtalan olyan célra használjuk, ami nem kötődik a hagyományos értelemben vett számítógépekhez, és amit szívesen végeznénk irodánktól távol, hordozható mikroszámítógép segítségével. Ilyen alkalmazás például az elektronikus levelezés, és a mostanában elterjedő Internet telefon. A hordozható számítógépek és a nagysebességű vezeték nélküli adatátvitel árának csökkenésével ezeknek az alkalmazásoknak a köre várhatóan bővülni fog, és megnő a mai mobiltelefonhálózathoz hasonló cellás mobil Internet-hozzáférést biztosító rendszerek jelentősége is. Ez a változás új kihívásokat jelent a cellás mobilrendszerek tervezése, működtetése és vizsgálata számára. Kutatómunkám során e kihívások némelyikét vizsgáltam meg.

A disszertáció első felében olyan problémákat elemzek, amelyek a ma is működő, vagy a szakirodalomban javasolt cellás mobil adatátviteli hálózatokhoz kapcsolódnak. Az első tézisben új számítógépes szimulációs módszert javasolok, amelynek segítségével nagy bonyolultságú mobil rendszerek vizsgálata a hagyományos szimulációs módszereknél hatékonyabban végezhető el. A javasolt "hibrid-hierarchikus" szimulációs módszer előnyeit szimulációs példákon keresztül mutatom meg. A második tézisben cellás mobil távközlési rendszerek elvi hatékonyságát vizsgálom. Analitikus módszerek segítségével megmutatom, hogy a "szoft handovert" nem alkalmazó, állandó csatornakiosztású rendszerekben az elérhető hatékonyság csökkenő cellamérettel csökken, és kiszámolom a lokális hívásengedélyezési módszerekkel elérhető legmagasabb rendszerhatékonyságot.

A disszertáció második felében megállapítom, hogy a ma is működő, illetve a szakirodalomban javasolt cellás adatátviteli rendszerek nem minden tekintetben felelnek meg az Internet-forgalom és az internetes alkalmazások által támasztott követelményeknek, és új cellás adatátviteli technológiat dolgozok ki. A harmadik tézisben e technológia alapelveit fogalmazom meg, illetve a javasolt rendszer működését meghatározó algoritmusokat írom le. A rendszer működőképességének alátámasztásaképpen ismertetem egy működő kísérleti megvalósítás főbb elemeit. A negyedik tézisben a javasolt új eljárás teljesítményelemzését végzem el, különös tekintettel a szolgáltatásminőségi paraméterekre és a rendszerhatékonyságra. Az elemzéshez analitikus, szimulációs és kísérleti módszerek kombinációját használom. A vizsgálatok alapján összehasonlítom a javasolt megoldást a létező, illetve az irodalomban előforduló hasonló célú megoldásokkal. A disszertációt az eredmények összefoglalása, illetve a további kutatási irányok felvázolása zárja.

List of Figures

1.1 Channel assignment map . . . 2
1.2 General architecture of a GSM network . . . 4
1.3 UMTS architecture . . . 5
2.1 Hierarchical simulation: traditional approaches . . . 10
2.2 Hybrid-hierarchical simulation architecture . . . 11
2.3 Hybrid-hierarchical simulation example (illustration) . . . 13
2.4 Simulation results for the single-link case . . . 15
2.5 Performance penalty of increased accuracy in the single-link case . . . 16
2.6 Network example - configuration . . . 17
2.7 Blocking probabilities with and without load sharing . . . 17
2.8 QoS violation without load sharing [min] . . . 18
2.9 QoS violation with load sharing [min] . . . 19
3.1 Efficiency as a function of call holding time. Single class of calls. . . . 25
3.2 System efficiency vs. diversity of call characteristics . . . 26
3.3 Long-term balance of the single class system . . . 30
3.4 Markov chain of the single class system with $\alpha'_{\text{step}}(n)$ . . . 30
3.5 Upper limit of the mean number of active channels per cell . . . 32
3.6 Maximum achievable efficiency vs. user mobility . . . 32
3.7 Maximum achievable efficiency vs. tolerated call failure probability . . . 33
4.1 Wireless access networks and Mobile IP . . . 37
4.2 Wireless access network design methodology . . . 37
4.3 Routing to/from wireless mobile hosts . . . 40
4.4 Wireless access network reference model . . . 42
4.5 Cellular IP handoff scenario . . . 45
4.6 Downlink routing algorithm . . . 48
4.7 Uplink routing algorithm . . . 49
4.8 Mobile host state machine . . . 50
4.9 Node implementation reference model . . . 52
4.10 Routing module in Cellular IP nodes . . . 52
5.1 Cellular IP testbed . . . 59
5.2 Gateway node architecture in the Cellular IP testbed . . . 60
5.3 Downlink packet loss at handoff (measurement) . . . 62
5.4 TCP sequence numbers at handoff (downlink case, simulation) . . . 63
5.5 TCP sequence numbers at handoff (uplink case, simulation) . . . 64
5.6 Throughput of TCP download (measurement) . . . 65
5.7 Location management cost vs. T_{ru} (dotted lines: measurement) . . . 66
5.8 Paging traffic rate generated by some applications (measurement) . . . 68
5.9 Network configuration for paging trade-off simulation . . . 69
5.10 Node throughput (measurement) . . . 71

5.11 Packet loss at advance binding handoff (simulation) 75
5.12 TCP throughput at hard handoffs and advance binding handoffs (measurement) 75
5.13 Packet loss at hard and semisoft handoff (measurement) 77
5.14 TCP throughput at hard, advance binding and semisoft handoffs (measurement) 78

List of Abbreviations

AAL2	ATM Adaptation Layer 2
AMPS	Advanced Mobile Phone System
ATM	Asynchronous Transfer Mode
B-ISDN	Broadband Integrated Services Digital Network
BS	Base Station
BSC	Base Station Controller
BSS	Base Station Subsystem
BTS	Base Transceiver Station
CAC	Call Admission Control
CDMA	Code Division Multiple Access
CDPD	Cellular Digital Packet Data
D-AMPS	Digital Advanced Mobile Phone System
ETSI	European Telecommunication Standards Institute
FA	Foreign Agent
FDMA	Frequency Division Multiple Access
GGSN	Gateway GPRS Support Node
GPRS	General Packet Radio Service
GSM	Global System for Mobile
GTP	GPRS Tunnelling Protocol
HA	Home Agent
HLR	Home Location Register
IETF	Internet Engineering Task Force
IMT-2000	International Mobile Telecommunications 2000
IP	Internet Protocol
ISDN	Integrated Services Digital Network
ISP	Internet Service Provider
JTACS	Japanese Total Access Cellular System
LAN	Local Area Network
MS	Mobile Station
MSC	Mobile Switching Center
NMT	Nordic Mobile Telephone
NSS	Network and Switching Subsystem
OSS	Operation and Maintenance Subsystem
PC	Personal Computer, Paging Cache
QoS	Quality of Service
RC	Routing Cache
RNC	Radio Network Controller
SGSN	Service GPRS Support Node
SMS	Short Message Service

TACS	Total Access Cellular System
TCP	Transmission Control Protocol
TDMA	Time Division Multiple Access
UDP	User Datagram Protocol
UMTS	Universal Mobile Telecommunication System
VLR	Visitor Location Register
WWW	World Wide Web

Nomenclature

K	Number of applications (call types) in cellular system
N	Number of wireless cells in cellular system
D_i	Capacity of ith wireless cell
λ_j	Arrival rate of calls of type j
μ_j	Inverse of type-j calls' mean holding time
b_j	Bandwidth occupied by type-j calls
B_j	Bandwidth reserved for a type-j call in deterministic reservation system
R	Wireless cell radius
ρ	Density of mobile users in service area
η	Constant specific to mobility pattern
$\mathbf{S}(\mathbf{t})$	Set of active calls at time t
P_h	Handoff blocking probability
P_f	Probability of call failure due to handoff blocking
ϵ	Resource efficiency of wireless system
M_j	Expected number of handoffs during a call of type j
A_j	Offered traffic from type-j calls
$\overline{b}$	Average used bandwidth in the system
$\overline{B}$	Average reserved bandwidth in the deterministic reservation system
γ	Cluster size
T_H	Mean time between handoffs
$\beta_{i,k}$	Rate of handoffs from the kth to the ith cell
$\mathbf{N_i}$	Set of cells neighbouring cell i
$n_k(t)$	Number of active calls in the kth cell at time t
$\alpha'(n)$	Accepted rate of new call attempts into cell with n calls
n^*	Mean number of active calls in cell
θ	$(M/(M+1))^2$
n_{loss}	Number of data packets lost at handoff
T_L	Handoff loop time
w	Rate of data packets sent to mobile host
T_{ru}	Inter arrival time of route-update packets (route-update time)
T_{pu}	Inter arrival time of paging-update packets (paging-update time)
α	Ratio of route-timeout and route-update time
β	Ratio of paging-timeout and paging-update time
r	Bit rate of data sent to mobile host
p	Fraction of time an active mobile host is not transmitting data
λ_P	Arrival rate of paging sessions
R_P	Mean amount of data sent in a paging session
C_i	Mobility cost of idle hosts
C_a	Mobility cost of active hosts

R_{ru}	Size of route-update packets
R_{pu}	Size of paging-update packets
T_a	Advance binding delay

Chapter 1

Introduction

The development of affordable palmtop devices with built in high-speed radio interfaces will have a major impact on the mobile communications industry. Large numbers of mobile users equipped with wireless Internet enabled communicators will require access to web based services anywhere anytime. The ubiquitous availability of wireless Internet access may superceed the popularity of cellular telephony and change the way we communicate. This environment places significant demand on existing and next generation mobility solutions.

The recent years have seen a rapid development of mobile communications technology. The *cellular principle* allows for the efficient use of the scarce radio resources and helps to support large subscriber populations. Advances in microelectronics, on the other hand, have made cellular telephones a commodity. The growing number of cellular phone users suggests that mobility will soon become the norm in communications, rather than the exception. While state of the art cellular mobile systems are still optimized for voice communication, they support an increasing variety of data services [13], [61]. Recent initiatives to augment the Internet with mobility support indicate the increasing interest in mobile data services [15], [14].

Future technologies for the support of wireless Internet access should leverage experiences from both cellular telephone systems and Internet technology. Flexible and scalable solutions are required that can adapt to a wide range of environments. Users must be offered seamless mobility across possibly heterogeneous systems which need to interact and co-operate to provide the best service available. The efficient use of the wireless interface, which continues to be the bottleneck in mobile communications, will become increasingly important with the emergence of mobile multimedia services. In this dissertation we address some of the challenges imposed by the design and analysis of wireless mobile communication systems in this new environment.

1.1 The Cellular Principle

Wireless communication systems face the common problem of spectrum scarcity. Due to the limited availability of radio capacity, the total rate of simultaneously transmitted traffic at any time is limited. Unlike wired communication systems that may increase the supported data rate in exchange for added equipment cost, wireless capacity limits are hard constraints in system design.

The first mobile telephone systems (e.g., Mobile Telephone Service, St.Louis, 1946) used a single base station which covered the entire service area. In these systems the number of simultaneous connections is limited by the number of available radio channels [29]. To make efficient use of the available radio capacity, state of the art wireless mobile communication systems exploit *channel reuse.* Channel reuse is based on the observation that the same wireless communication

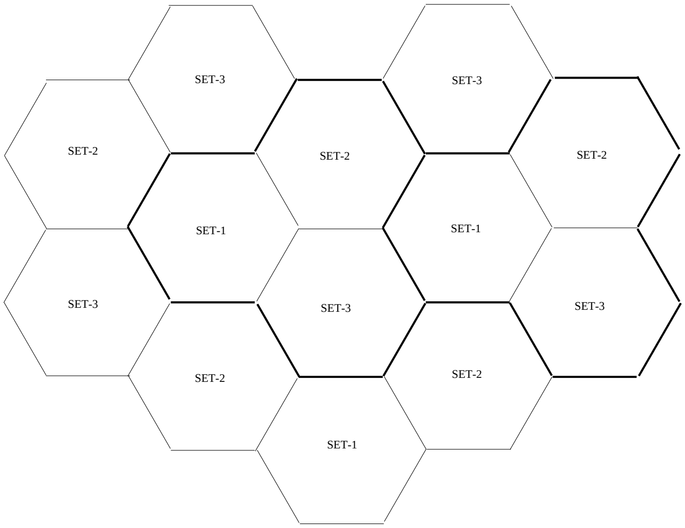

Figure 1.1: Channel assignment map

channel can be used for multiple simultaneous communication sessions in a system if these sessions are sufficiently distant to avoid interference. To allow for channel reuse, the total service area is divided into smaller areas called *cells*. In the case of *Fixed Channel Assignment* each cell is statically assigned a subset of the logical radio channels available for the system. The logical channels are distributed among cells such that adjacent cells use disjoint subsets of logical channels, but distant cells may reuse the same channels. In *Dynamic Channel Assignment* systems the association of channels with cells may change in time in response to changing traffic conditions. (In the discussion that follows a 'channel' denotes a logical resource that may be a frequency domain, a set of time slots or an assigned code, depending on the access technology.)

Figure 1.1 illustrates radio channel assignment in a cellular system where radio channels are partitioned into three sets. Each cell is associated with one out of the three sets of channels. The regular assignment of channels to cells creates repetitious patterns in the channel assignment map. These patterns (i.e., *clusters*) are shown by thick lines in Figure 1.1. The number of cells in a repetitious pattern is commonly called the *reuse factor*. The reuse factor in the cellular system shown in Figure 1.1 is three. For a given cell size the reuse factor determines the distance between cells that use the same radio channels. Increasing the reuse factor decreases the interference between such cells but at the same time decreases the capacity offered in one cell. The reuse factor is chosen based on system specific interference limits and is typically fixed for each type of radio interface standard.

Dividing the service area into cells increases the total rate of possible simultaneous communication sessions in the system. At constant reuse factor, the total available rate depends on the size of wireless cells. Shrinking cell size (while transmission power is reduced accordingly) increases the number of cells where a given logical radio channel can be simultaneously used thus increasing the total available user data rate. This gives rise to micro and pico cellular systems. Splitting the service area into small cells, however, has a number of drawbacks and system capacity can not be infinitely increased this way.

In contrast to early wireless systems where the entire service area shares the same radio resources, cellular systems require a dedicated wireless access point (i.e., base station) in each cell. These base stations represent a cost to the network operator. In addition to the base stations themselves, cellular systems require a *fixed network infrastructure* that interconnects base stations. This infrastructure is also often referred to as a “cellular network”. In this dissertation we primarily focus on networking issues of cellular wireless systems and hence use the terms ‘cellular system’ and ‘cellular network’ interchangeably throughout the thesis.

The cost of base stations and of connecting infrastructure increases with decreasing cell size representing a price for increased system capacity. In addition to this cost, the cellular principle gives rise to mobility related phenomena not present in non-cellular wireless systems. We use the term “migration” to refer to a user’s moving from one wireless cell to another one. A migration while the user is engaged in active communication is called a “handover” or “handoff”. Control functions associated with migrations and handoffs also appear as costs to the network operator. Handoffs must be handled by the network with little or no disturbance to ongoing communication sessions. Decreasing cell size results in increased handoff rate which increases the control messaging and processing associated with handoff. Finally, handoff also represents a cost to the network operator by decreasing system efficiency as discussed in Chapter 3. This cost also increases with increasing handoff frequency.

The mobility of users among cells imposes another problem that did not exist in non-mobile communication systems. In order to be able to quickly establish communication paths toward mobile users, the system must maintain information related to the location of users in its service area. Without such location management information a user would need to be searched for in the entire service area before data can be routed to the user. Maintaining and updating the location management data base, however, represents a cost to the operator which increases with increasing user speed and decreasing cell size. To avoid overloading the system by location update messages, most existing cellular systems separate the location management of users actively engaged in a communication session from location management of other (i.e., ‘idle’) users. While the location of active users must be exactly known to the system the location of idle users is only approximately recorded. Before establishing a data path to an idle user, its exact location is determined in a process called *paging*.

1.2 State of the Art

The first generation of cellular mobile systems uses Frequency Division Multiple Access (FDMA) technology and analogue modulation. The most widely deployed first generation standards are the Advanced Mobile Phone System (AMPS) that nearly ubiquitously covers North America and Nordic Mobile Telephone (NMT) that was developped in Scandinavia and is now widely spread in Europe. In the United Kingdom, a first generation system called Total Access Cellular System (TACS) is used, a modified version of which is deployed in Japan (JTACS).

In contrast to first generation cellular standards, second generation systems use digital voice coding. Examples of second generation cellular systems are IS-136 which uses Time Division Multiple Access (TDMA) radio technology and IS-95 which uses Code Division Multiple Access (CDMA). In Section 1.2.1, we present a brief overview of the Global System for Mobile Communications (GSM), a second generation cellular system that now serves over 110 million subscribers [86].

While second generation cellular standards are still optimized for conversational voice, they also provide data services to the mobile user. Third generation cellular systems will support voice, data and multimedia services in an integrated environment. An introduction to third generation cellular systems is presented in Section 1.2.2. In addition, recent initiatives to augment the Internet with mobility are discussed in Section 1.2.3.

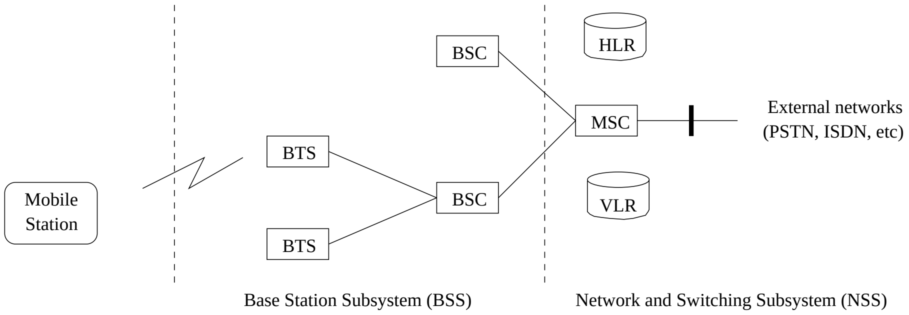

Figure 1.2: General architecture of a GSM network

1.2.1 The Global System for Mobile Communications

The GSM system is composed of the Base Station Subsystem (BSS), Network and Switching Subsystem (NSS) and Operation and Maintenance Subsystem (OSS) [16]. The network architecture is depicted in Figure 1.2. The BSS consists of Base Transceiver Stations (BTS) and Base Station Controllers (BSC) and is in charge of providing and managing transmission paths between the Mobile Stations (MS) and Mobile Switching Centers (MSC) which are the primary building blocks of the NSS. The NSS includes switching and location management functions. In particular, it comprises the Home Location Register (HLR) and the Visitor Location Register (VLR) functions which represent GSM's location management data bases. The NSS is also responsible for interfacing external networks such as the public telephony network. Finally, the OSS provides functions for network management interactions to all the above listed entities.

Though GSM is primarily designed to support conversational voice services, it offers a variety of data services. GSM users can send and receive short alphanumeric messages using the Short Message Service (SMS). In addition, fax and circuit switched data services (up to 9600 bps) are provided. Recently, the importance of providing packet data services to cellular mobile users has grown due to the increasing role of IP based networks. As a response to this demand, GSM will in the near future be extended by the General Packet Radio Service (GPRS) [13]. Similar to the Cellular Digital Packet Data (CDPD) extension to AMPS used in North America, GPRS will reuse the existing radio interface for the transmission of data packets. In the wired network infrastructure, GPRS will to a large extent follow the GSM architecture. The network consists of Gateway GPRS Support Nodes (GGSN) and Service GPRS Support Nodes (SGSN). To tunnel data packets between GGSNs and SGSNs GPRS uses a special protocol called GPRS Tunneling Protocol (GTP). GPRS is primarily intended to support applications which generate bursty traffic such as World Wide Web (WWW), e-mail and other Internet applications [83].

1.2.2 Third Generation Cellular Systems

The increasing importance of data and multimedia services in addition to voice communication calls for a new generation of cellular systems. Third generation cellular systems are expected to provide a variety of services to mobile users anywhere anytime. The concept referred to as International Mobile Telecommunications 2000 (IMT-2000) includes high quality access to the Internet and to future broadband integrated services digital networks (B-ISDN) as key components. Standardization of this new system is carried out mainly by the ITU-T SG11 on an international level and by the European Telecommunication Standards Institute (ETSI) in Europe [61], [82],

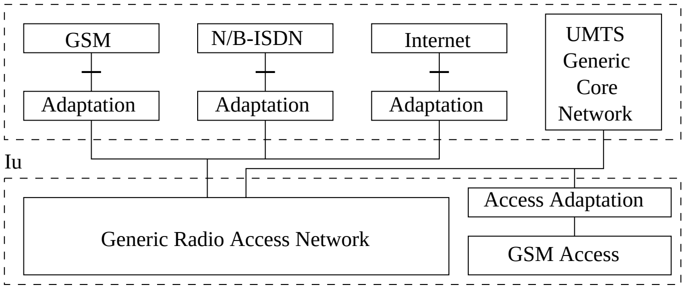

Figure 1.3: UMTS architecture

[65].

In order to gain wide acceptance, the European initiative for IMT-2000, called Universal Mobile Telecommunication System (UMTS) includes smooth evolution from second generation mobile systems, particularly the GSM system. The schematic architecture of UMTS is illustrated in Figure 1.3 [66]. To efficiently support a mix of voice and bursty data traffic, the International Mobile Telecommunications 2000 (IMT-2000) standard suggests using Code Division Multiple Access (CDMA) and Asynchronous Transfer Mode (ATM) in third generation mobile systems [64], [65], [82], [61], [67]. To support the strict delay requirements of low bit-rate encoded conversational voice over an ATM based cellular network infrastructure, ITU-T has recently standardized a new ATM Adaptation Layer, AAL2 [34], [39]. Overviews of CDMA/ATM mobile systems for the support of voice and multimedia and of related standardization activities are provided in [61] and [68]. A performance evaluation framework for voice and IP services in UMTS is presented in [J1].

1.2.3 Internet Mobility Proposals

Independent of the initiatives to add data services to cellular mobile communication networks, recently a multitude of proposals have appeared to augment the Internet with host mobility support. In [33] a host is defined mobile if, as it migrates around the local or wide area network, a user cannot differentiate its operation and performance from that of a fixed host.

A basic difficulty that protocols in support of such host mobility must cope with is that the host address in the Internet Protocol (IP) has dual significance. First, as a unique host identifier it should be kept constant regardless of mobility. Second, in its role as a location pointer it should change as hosts change location [33]. These are competing requirements that mobile host protocols should efficiently resolve. A fundamental problem to solve is therefore the separation of these two roles while an up-to-date mapping of host identifiers to location information is made available. It has been shown in [15] that most of the proposed solutions can be viewed as special cases of a “two tier addressing” architecture where a mobile host is logically associated with two IP addresses; that is, its home address that serves as an unchanged host-identifier and an address that reflects its point of attachment to the Internet. This general architecture comprises three fundamental components. A Location Directory represents a data base that contains the most up-to-date mapping between the two address spaces. The translation of the host identifier to the actual destination address in each packet is performed by Address Translation Agents. The final component of the generalized architecture is the Forwarding Agent that performs the inverse translation in order to ensure that packets arriving at the mobile host have its constant home address in the destination field. An overview of Internet mobility proposals founded on this general concept is provided in [15].

A set of requirements that mobile host protocols should fulfill in addition to solving the address translation problem is provided in [33]. *Operational transparency* means that the user should not need to perform any special actions, such as manual reconfiguration, before or after host migration. Host mobility protocols should provide transparent interworking with correspondent hosts automatically, without restarting or reconfiguring the host at migration. In addition to operational transparency, mobile host protocols should provide *performance transparency*, meaning that the performance of a host is not degraded by mobility. In order to achieve performance transparency, the mobility protocol should aim at optimal routing of packets to and from mobile hosts.

Besides operational and performance issues, mobility protocols should take complexity and implementational cost into account. *Backward compatibility* is important because mobility will be gradually introduced in the Internet and mobile and non-mobile hosts must smoothly interwork. To reduce the cost of mobility, mobile host protocols should require little added infrastructure to what is present in IP networks. Finally, mobile host protocols must incorporate *user authentication* and *security* functions.

In the following, we describe the Mobile IP protocol adopted by the Internet Engineering Task Force (IETF) [14]. In this host mobility solution, each mobile host is assigned an IP address which serves as its unique identifier regardless of its actual location. This IP address is called a *home IP address* and its routing domain is referred to as the *home domain* for the given mobile host. When the host visits a network other then its home domain, then it is assigned a temporary IP address that reflects its current point of attachment to the network. This address is called a *care-of address.*

Hosts willing to send data packets to the mobile host are not supposed to be aware of its actual location and hence send the packets to its home address. In the home domain, a *Home Agent (HA)* is responsible for intercepting these packets and tunnelling them to the care-of-address. For this avail, the mobile host must communicate its care-of-address with the Home Agent each time it moves to a new network. Packets arriving at the care-of-address must be decapsulated (that is, the tunnelling header is removed). This task is performed by the *Foreign Agent (FA)* which can be the mobile host itself or a router in the visited domain.

This scheme fulfills the requirement of operational transparency because neither the correspondant host nor the transport and application layers of the mobile host notice host mobility. It does not, however, fulfill the requirements of performance transparency and optimal routing. The route path from a correspondant host toward the mobile host will always traverse the Home Agent which may impact packet delay and service quality. This property, often referred to as *triangular routing*, reflects a design decision that prioritizes security in exchange for service quality. The triangular route can only be cut through if the correspondant host or other entities were informed about the mobile host's actual location. This possibility would, however, raise significant security concerns and is therefore precluded in the present version of the Mobile IP protocol [14]. Future versions will be extended by secure communication channels between the mobile host and correspondant hosts and will therefore allow for route optimization [26], [30].

1.3 Research Objectives

The growing importance of mobile communications and of the Internet indicate that the demand to access web based services from wireless mobile devices will increase in the near future. The Internet is more and more used for applications that are hardly or not related to computers in the traditional sense. E-mail, World Wide Web, IP telephony and other dominant Internet applications are services that we commonly access through desktop or laptop computers but they could equally be used through Internet enabled palmtop devices, mobile telephones, intelligent pagers or other portable devices. The Internet that has traditionally been used to interconnect computers is becoming a global communications infrastructure that carries voice, data and multimedia services to users world wide.

In this environment, data services provided by today's cellular mobile communications systems will no longer be sufficient to meet future user demands. Flexible and scalable cellular wireless Internet access networks that support a large number of attached subscribers are required. In this dissertation, we adopt the concept of Wireless Overlay Networks [31] and assume the co-existance of a large variety of cellular Internet access network technologies. Each of these technologies will be optimized for a different geographic region and service level. While some technologies can provide high data rate in an indoor environment, others may cover a metropolitan area but offer lower data rate. Such networks can independently be operated and offered to wireless mobile terminals. Mobile terminals will scan available access networks and select one based on service quality, cost and other parameters. We define the future Wireless Internet as a combination of these heterogeneous wireless access technologies together with a mobility enabled wired Internet.

In this dissertation we address some of the challenges that existing and future cellular wireless mobility solutions will meet in an environment of ubiquitous wireless Internet availability. In the first part of the dissertation, we investigate cellular performance issues related to existing approaches. We propose a simulation architecture optimized for complex mobile communications systems (Chapter 2) and study the resource utilization of cellular wireless mobile networks (Chapter 3).

The IMT-2000 concept of third generation cellular systems and the Internet mobility proposals address the issue of wireless mobile data networks from two different perspectives. We argue that both of these approaches have a number of shortcomings. Internet mobility proposals represent simple and scalable global mobility solutions but are not appropriate to support fast and seamless handoffs. In contrast, third generation cellular systems offer smooth mobility support but are built on complex networking infrastructure that lacks the flexibility offered by IP based solutions. The second part of this dissertation is dedicated to the design and analysis of an alternative solution that represents a 'third way' in cellular mobile data networks (Chapters 4 and 5). We conclude the dissertation by comparing our proposal to existing solutions and discussing future research directions.

Chapter 2

Hybrid-Hierarchical Simulation Architecture

Together with measurements and analytical methods, the simulation-based evaluation of cellular systems will be increasingly important as the deployment of new mobile applications imposes new requirements both on the radio interface and on the fixed network infrastructure. Efficient allocation of the network's resources must be based on reliable and flexible performance evaluation techniques. In this chapter we propose a simulation environment optimized for the performance analysis of complex mobile networks. To handle the complexity of the system without losing low-level details due to a high-level abstraction, a hierarchical simulation structure is proposed which also relies on analytical techniques built into the simulator.

2.1 Problem Statement

Future mobile communication networks will support voice, data and multimedia traffic over the same infrastructure. Providing the required service quality for a variety of traffic types in a mobile environment is a challenging task. Mobile users may move from one network access point (base station) to another while engaged in a communication session and expect the network to handle these migrations with little or no disturbance to the application. The network operator must ensure that these handoffs are successful with high probability. At the same time, the precious wireless and network resources must be utilized efficiently.

In this environment network behaviour can not be described at a single time scale. Service quality parameters such as packet loss and delay are determined by events in the packet time scale. The frequency of handoffs depends on user speed and on the size of wireless cells. Calls and data sessions are initiated and terminated at an even higher time scale. Network planning and management must consider events at all these time scales and be aware of interactions between them. The complexity of this task calls for a combination of evaluation techniques. While analytical methods provide the most general view of a system's behaviour, tractable models that capture events at all time scales are not always available. Prototyping and measurements give accurate information about the system under study but are time consuming and expensive. In what follows, we will concentrate on computer simulation techniques. The advantage of simulations is that the level of abstraction can be freely determined as dictated by the studied phenomena and the required accuracy. The simulation of complex systems, however, becomes time consuming and sometimes even impossible unless the simulation model simplifies and hides some low level details. A number of techniques have been proposed in the literature to increase simulation power and reduce simulation time. A brief overview of such techniques is provided in Section 2.2.

In this chapter we propose *hybrid-hierarchical simulations*, a technique relying on a hierarchical decomposition of the simulation task and on the integration of analytical techniques into simulation. The proposed structure allows for the simulation of a large and complex network without hiding the low level details behind a high-level abstraction. The simulation architecture comprises a network simulator, a device level simulator and an assignment queue. The studied system is primarily simulated in the network simulator that uses the device level simulator to zoom in and investigate details where necessary. The output of a simulation study is the combination of simulation results gained in the network and device level simulators.

In Section 2.2, we provide an overview of advanced simulation techniques. A hybrid-hierarchical simulation architecture is outlined in Section 2.3 and analyzed in Section 2.4. Conclusions are presented in Section 2.5.

2.2 Related Work

For large and complex systems a fully detailed simulation of the entire problem is often unrealistic. A byte-level simulation of a single ATM connection is so time-consuming that it is impractical in real investigations. While in simpler systems (PSTN or other constant bit-rate, single application communication systems) a higher level investigation may be appropriate, a more sophisticated system's characteristics such as bit error rate or delay can depend largely on lower level behaviour.

In the simulation of packet switched cellular mobile networks an additional difficulty arises from the fact that events at various levels of abstraction and at various time scales need to be modeled and simulated. For instance, low level changes in the quality of the radio interface may trigger a handoff event at the connection level, which, in turn, may have cell level consequences inside the affected switches. We observe that this basic characteristic has two major general requirements for an efficient and practically useful simulator:

- the description, modeling and simulation of the system must be able to capture relevant events at whatever level of abstraction they happen;
- the description and modeling of the system must support the simulation of events at whatever time scale they happen.

Extending the classification of [49] and [50] the various techniques for enhancing modeling and simulation efficiency of complex systems fall into the following broad categories:

- hybrid models increase the efficiency of the simulation by combining analytical models with simulation, see e.g., [56], [57] and [40]. Our method inherits the basic idea of combining analytical and simulation techniques, as described in Section 2.3.
- variance reduction techniques improve computational efficiency by using statistical methods to obtain more accurate performance measures, as in [60], [44], [43], [36], [47] and [48]. We have found that finding a good probability transform at various abstraction levels and time scales can be difficult. Even though these methods offer a considerable increase of simulation speed without requiring more processing capacity so far their applicability has only been shown for relatively simple examples and their extension for more realistic problems needs further research. For an overview of these and other special simulation techniques including hybrid and hierarchical simulation see [49] and [50].
- extrapolative methods increase computational efficiency of a simulation by employing statistical methods to estimate the tail probability distribution outside the sample range [51], [52], [53], [54].

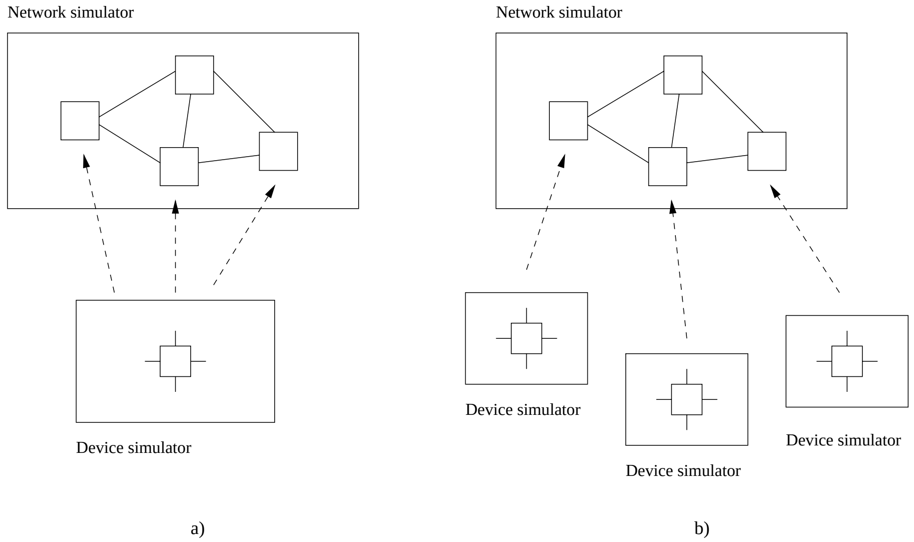

Figure 2.1: Hierarchical simulation: traditional approaches

- parallel and distributed methods attempt to increase the simulation time by employing more computer resources, see e.g., [59] and [58] and the references therein. The performance of even advanced parallel simulation techniques, however, does not seem to justify the additional programming effort which is needed in the decomposition and synchronization tasks inherent in such techniques.
- co-simulation techniques aim at loosely interconnecting two or more independently running simulators of different abstraction levels by allowing them to exchange messages. This approach, though attractive, often suffers from problems caused by timing and causability constraints [41]. The challenge of efficient communication between the various levels in multiple time scale simulations is addressed in e.g., [42], but the solution proposed there is not directly applicable to communication networks. Our approach is in fact a one directional co-simulation technique, also importing ideas from the hybrid approach. The main benefit of these changes is that the higher level simulator never needs to await results from the lower level counterpart. Instead, when needed, the higher level simulator uses predictions.

2.3 Simulation Architecture

2.3.1 Overview

The analysis of cellular mobile networks is often focussed on the trade-off between network utilization and per-connection service quality parameters. Typically, the network's response to various control and routing strategies needs to be evaluated with service quality requirements as optimization constraints. This kind of investigation requires that the entire network be studied while the model is detailed enough to include the internal structure of network elements down to queues and processors. As this is not feasible in one simulator we propose a hierarchical decomposition of the problem.

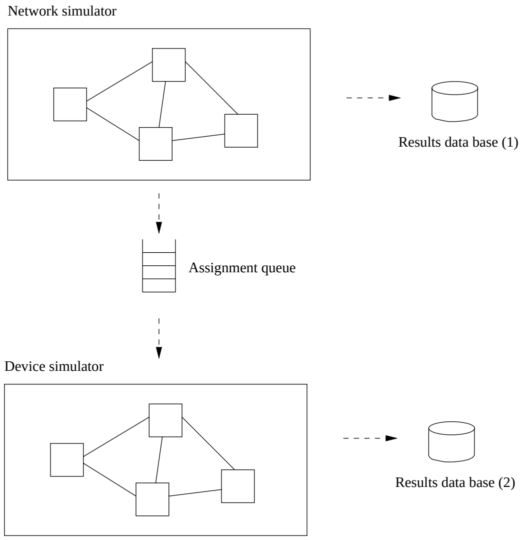

Figure 2.2: Hybrid-hierarchical simulation architecture

As illustrated in Figure 2.1, in a traditional approach to hierarchical simulation the lower level simulator(s) either provide(s) characteristics about a number of identical or similar network entities (Figure 2.1a) or a dedicated lower level simulator must be assigned to each network element of interest (Figure 2.1b). Both approaches have drawbacks, however. The former solution is based on the investigated system's specific inherent feature of having a number of identical network elements working in similar circumstances, which does not necessarily apply for cellular systems. The latter requires the use of a number of simulators in parallel, which might come back to the problem of insufficient processing capacity with the additional problem of requiring a specific simulator for each network element of interest.

The hybrid-hierarchical simulation architecture is motivated by the fact that a fully detailed simulation of all network elements may not only be unrealistic, but often superfluous. Service quality parameters such as packet loss or delay values are only of interest when they are close to or above their specified limits. To save simulation capacity the hybrid-hierarchical approach, illustrated in Figure 2.2, uses estimates on these parameters and relies on device level simulation only when more accurate information is required. The studied system is primarily simulated in a network level simulator that covers device level details behind an abstract system model. The hybrid-hierarchical architecture offers benefits if there exist low level system parameters that can not be determined in the network simulator but are of interest when they leave their respective pre-defined tolerance ranges. The network simulator maintains estimates on these parameters to detect when the parameters are close to their limits. To determine a parameter's exact value, in these cases the network simulator initiates a device level simulation session focussed on the given parameter.

Device level simulation assignments defined by the network simulator are processed sequentially by the device level simulator. Simulation results gained from these sessions are not fed back to the network simulator. Rather, they are stored in a data base and made available to the user after the network simulation is terminated. In contrast to co-simulation techniques, this approach does not allow device level simulation results to affect the network level simulation.

Traffic control decisions performed in the network simulator must be based on data available in the network level abstraction or on the estimated low level parameters. Despite the limitations of this approach, it tends to model real network that are controlled by inaccurate estimates but allow for measurements of “arbitrary” precision. Call admission control decisions, for example, may be based on the effective bandwidth calculation [81]. The quality of service offered to established calls, however, can be measured accurately.

A hybrid-hierarchical simulation session is complete when the network simulator is terminated and the device level simulator has processed all simulation assignments. If the number of device level simulation sessions is high, the total simulation time may become comparable to a pure device level simulation. The advantage of the approach compared to device level simulation is therefore larger if the number of device level simulation sessions is low, that is, the low level parameters rarely exceed their respective tolerance ranges.

2.3.2 Model

Network Level

The network level simulator used in our hybrid-hierarchical simulation implementation is an extended version of the PLASMA simulator, inheriting most of its modeling capabilities. PLASMA is a generic event driven simulation platform for packet based communication networks, particularly ATM and IP. A detailed description of PLASMA structure and functionalities is provided in [38]. In what follows, we outline the most important modelling features of the simulator.

Traffic in the network simulator is modelled at a flow level where users are characterized by calling behaviour and mobility parameters. Users generate data flows of type j ($j = 1 \dots K$) according to a stochastic process ψ_j. Each flow type is assigned a packet level traffic description and a set of service quality parameters. The packet level description can be an arbitrary stochastic process determining packet arrival times and sizes. In addition, each flow type may be assigned an abstract traffic descriptor (e.g., effective bandwidth) to be used by traffic control algorithms in the network level simulator. We note that though this approach suggests that only connection based communication networks can be studied, the notion of flows is purely an abstraction used in the simulator and does not limit the applicability to IP networks.

Typical events in the network simulator are the establishment, rerouting or release of communication flows. Rerouting may occur due to user mobility or in response to a change of network routing state or traffic conditions. Both wired and wireless links are modelled by their bandwidth, delay and error rate. Routing decisions can be arbitrary functions of the traffic source and destination identifiers, the flow type and its abstract traffic descriptor and the load conditions in the network. Optionally, admission control can be performed on a per hop basis using the same input parameters.

Device Level

In the device level simulator traffic is modelled on a per packet basis. Traffic sources (i.e., users) generate data according to the stochastic traffic descriptors assigned to flow types $j = 1 \dots K$. Depending on the simulated system, user data is broken into ATM, AAL2 or IP packets. Packets propagate through a network of links, queues, switches (routers) and multiplexers. Signalling is not simulated but end-to-end flow control (e.g., TCP) is modelled. Radio propagation issues are covered by an abstract radio link model where link capacity and error rate are modulated by a stochastic time function. The establishment and release of flows is not modelled in the device level simulator. The set of active flows is specified as an input parameter to the simulator and is unchanged during the device level simulation session.

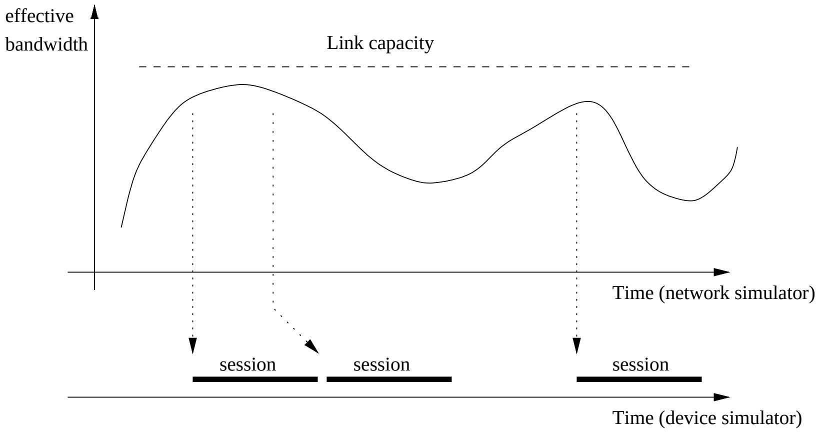

Figure 2.3: Hybrid-hierarchical simulation example (illustration)

2.3.3 Communication and Synchronization

Figure 2.3 schematically illustrates the operation of the hybrid-hierarchical simulation system in an elementary example. In this example connections are being established and released over a physical link. The network level simulator controls this process by applying admission control based on the connections' aggregate effective bandwidth. The operation of the system is therefore determined by this possibly inaccurate estimate. Through the device level simulation sessions, however, we learn about per connection service quality to a depth that is not available in the network simulator.

The upper part of Figure 2.3 shows the aggregate effective bandwidth of active connections calculated by the network simulator throughout the simulation session. In this example, device level simulation sessions will be triggered when this aggregate effective bandwidth is close to the link's capacity. Points in simulated time when such sessions are triggered are indicated by vertical dotted arrows. In the lower part of Figure 2.3 we show the sequence of device level simulation sessions. Each such session provides deep insight about a single point of time in the network simulator's session. If a device level simulation session is not terminated before the next session is triggered, as is the case in the figure, the device level simulator will have a backlog compared to the network simulator. Simulation assignments sent from the network to the device level simulator are then queued in an assignment queue (see Figure 2.2). Based on the device level simulation sessions initiated at critical periods, at the end of the simulation in addition to a network level view we will have exact information on per connection service quality. As a side effect, device level simulation results provide a cross-checking of estimation made in the network simulator and eventually give indications of its errors.

In periods of time when the aggregate bandwidth is low, service quality is likely to be satisfactory and the exact quality parameters are of little importance making device level simulation unnecessary. The advantage of hybrid-hierarchical simulation compared to a full device level simulation comes from omitting detailed simulation in these periods.

This example also illustrates how in the hybrid-hierarchical approach simulation accuracy can be increased in exchange for increased simulation time. In the description above we assumed that device level simulation sessions are triggered when the aggregate effective bandwidth of active connections is "close" to the link's capacity. By adjusting the definition of being "close" to the capacity limit, we can effectively increase of decrease the accuracy of information gained on per connection service quality. A numerical example of this adjusting possibility is provided in the following section.

2.4 Analysis

In this section, following an overview of validation results, we provide two simulation examples to illustrate the hybrid-hierarchical simulator's capabilities. The examples are taken from the analysis of ATM Adaptation Layer Type 2 (AAL2) [39] where standardization activity was based on detailed performance evaluations [34], partly using the hybrid-hierarchical simulation environment. AAL2 supports efficient transport of delay sensitive compressed voice connections over ATM and will hence play important role in ATM based cellular mobile systems.

2.4.1 Implementation Validation

The hierarchical structure of the proposed simulation environment implies that validation must be performed for both the upper and the lower level simulators. In order to validate the network simulator we have considered a number of cases detailed in [46] and [35]. In [35] a model for multirate circuit switched loss networks with non-zero call processing time is developed, which allowed us to compare simulation output to analytical and approximative results in non-trivial cases.

To validate the device level simulator, we have used a series of test cases where comparison with analytical/approximative techniques is feasible. In particular, we have considered single queue – single server systems with batch arrivals as in [45]. The $D^{[x]}/D/1$ queueing system is chosen because it plays an important role in the modelling of systems which carry compressed variable bit rate voice samples over ATM, most notably in the modelling of GSM/UMTS systems with AAL2 transport. A series of simulation results are presented and compared to theoretical results in [J1] showing nice match between simulation and analytical results.

2.4.2 Single-link example

In this example voice and data connections are established and released on a link of capacity $C = 1.5$ Mbps. 50 voice and 20 data sources initiate calls according to Poissonian arrival processes with parameters $\lambda_v = 0.002$ and $\lambda_d = 0.001$, respectively and maintain the connections for exponentially distributed times with parameters $1/\mu_v = 500$ sec and $1/\mu_d = 1000$ sec, respectively. Active voice sources generate packets with a constant inter-arrival time $T = 10$ ms where the packet size is determined by an embedded state machine of four states such that the mean rate is 9 kbps and the peak rate is 20 kbps. The measurement based four-state model is extensively described in [J1]. Active data sources are of on-off behaviour with exponentially distributed "on" and "off" period lengths with parameters $\alpha_{on}/\alpha_{off} = 0.23$ and rate $r = 64$ kbps in the "on" state. Traffic sources are all independent. Both applications tolerate a maximum packet loss probability of 10^{-3}.

Active voice sources are assigned dedicated AAL2 connections each and are all statistically multiplexed in a single ATM VCC. In addition, each active data source is assigned an ATM VCC. The VCC carrying AAL2 voice connections is statistically multiplexed with and is prioritized over the data VCCs. Such scenarios are expected in ATM based cellular networks [34]. Figure 2.4 shows the effective bandwidth estimation maintained by the network-level simulator during a 1000 minute simulation. In this example, device level simulation sessions were triggered when the estimated effective bandwidth exceeded a pre-defined *critical threshold.* In Figure 2.4, three critical thresholds are shown as horizontal solid lines and results from some device level simulation

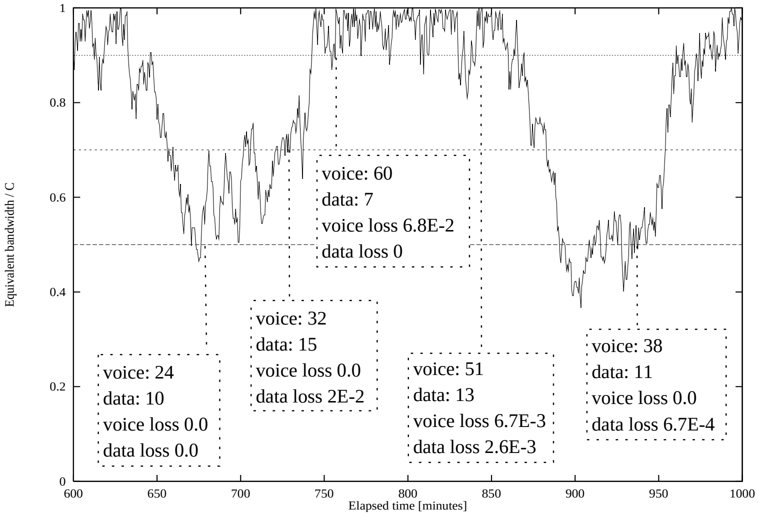

Figure 2.4: Simulation results for the single-link case

sessions are shown in the dotted boxes hanging from the effective bandwidth curve. In each box we specify the number of active voice and data connections and the packet loss probability perceived by voice and data users, as obtained in the device level simulator.

We recall, that this simulation example is for illustrative purpose only. Since the accuracy of the effective bandwidth calculation for the two application types may be different, a thorough system analysis would require that the triggering of device level sessions also depends on the ratio between voice and data connections in the aggregate effective bandwidth. The example shows, however, the added information gained from device level simulation sessions. Setting the critical threshold at 90% of the link capacity, device level results indicate that at the traffic peaks service quality was poorer than required. This shows that the effective bandwidth estimation was too optimistic, but it does not tell us the per-connection service quality perceived by users throughout the simulation. By lowering the critical threshold to 70% we trigger more frequent device level simulation sessions. We observe that this gives more accurate information on QoS parameters. By further lowering the critical threshold, the pure device level simulation can be approached. Results obtained in device level simulations at this lowest threshold show that service quality is satisfactory at all observation points.

This example showed that by determining the conditions that trigger device level simulation sessions, the hybrid-hierarchical simulation environment can be freely tuned in the trade-off between accuracy and simulation time. This phenomenon is illustrated in Figure 2.5. Here we varied the critical threshold from 100% to 50% of the link's capacity. Setting the threshold to 100% corresponds to a pure network level simulation while a threshold of 50% is practically equal to a pure device level simulation. To illustrate the extra information gained from device level sessions, we plotted the number of occurrences when the device level simulation session proved the effective bandwidth estimation to be incorrect. This quantity is shown as solid line in Figure 2.5. In addition, we plotted the total number of performed device level simulations (dashed line) which represents the "cost" of the hybrid-hierarchical approach in terms of processing requirement. In accordance with expectations, by lowering the critical threshold the extra information gained from the device level simulator increases in exchange for increased processing cost. We observe,

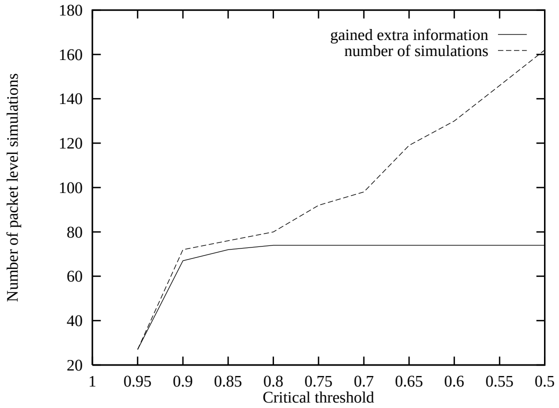

Figure 2.5: Performance penalty of increased accuracy in the single-link case

however, that approaching the pure device level simulation this cost drastically increases while the gain compared to a network level simulation saturates. This phenomenon indicates that the optimal simulation accuracy is between pure network or device level simulations and justifies the hybrid-hierarchical approach. Determining the optimal operation point in more complex cases is subject to further study.

2.4.3 Network example

Our next simulation example illustrates the speed-up we can achieve using the hybrid-hierarchical simulator. The simulated network configuration is illustrated in Figure 2.6. This small network follows a typical cellular architecture consisting of two base station sub-systems. Each sub-system consists of two base stations and one Radio Network Controller (RNC) that are connected in a ring for reliability purposes. The two sub-systems are connected to a Mobile Switching Center (MSC). Mobile users generate voice and data traffic with traffic parameters and QoS requirements as in the previous example.

In this example we investigate the benefits of a direct RNC-RNC connection in a local overload situation and show that the hybrid-hierarchical simulation environment allows for an analysis that would not be feasible using standard simulation techniques. Direct RNC-RNC connections such as the one studied here are rare in today's hierarchically built cellular networks but will become more common in ATM based IMT-2000 systems which allow more flexible establishment and release of inter-node connections.

In the beginning of our simulation analysis, mobile hosts are evenly distributed among the four base stations of Figure 2.6. By forcing mobile hosts to migrate to the first base station sub-system (left side) we cause an overload in this sub-system and on the corresponding RNC-MSC link. The solid line in Figure 2.7 shows that call blocking probability, seen in the network simulator, increases on this sub-system as the total offered load increases. In order to limit this condition and to distribute load evenly in the network, we now take advantage of the direct RNC-RNC link by applying load sharing. Load sharing directs some incoming calls toward the RNC-RNC link instead of the RNC-MSC link and allows the two sub-systems to share the load despite the uneven distribution of mobile hosts in the service area. The dashed line in Figure 2.7 shows call blocking probabilities in the overloaded sub-system when load sharing is applied. We observe that load sharing decreased call blocking probability.

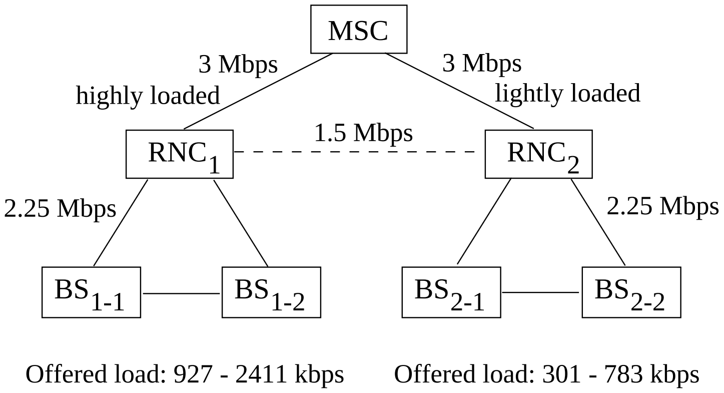

Figure 2.6: Network example - configuration

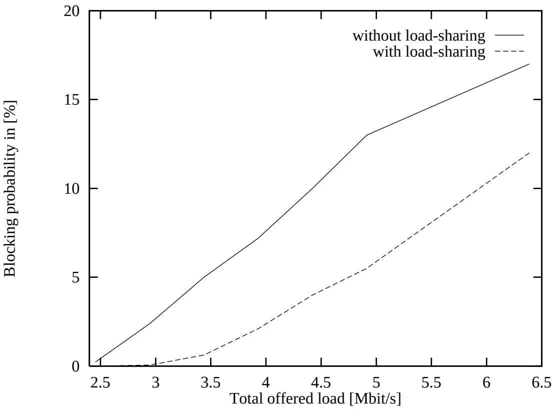

Figure 2.7: Blocking probabilities with and without load sharing

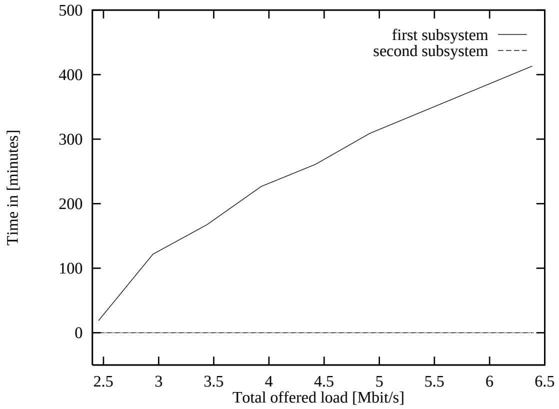

Figure 2.8: QoS violation without load sharing [min]

These results were obtained from the network level simulator and require no device level simulation sessions. However, the overload situation and our management action also affect per-connection service quality as perceived by the mobile users. These parameters are not shown by network simulation results. By exploiting the hierarchical-hybrid simulation we can monitor the packet level QoS without an unacceptable simulation time. By setting again critical thresholds on the aggregate effective bandwidth of connections on each link in the network, we can trigger device level simulation sessions whenever QoS requirements are likely to be violated. In Figure 2.8 and Figure 2.9 results from these simulation sessions are shown in the overload situation without and with load sharing, respectively. For simplicity, we only plotted an aggregate characteristic of per-connection service quality, that is the total time a user perceived unsatisfactory service quality, out of a 500-minute simulation. (The service quality constraints were identical to those in the single-link example.)

In Figure 2.8, we observe that QoS is often violated in the first (overloaded) sub-system, but never in the second sub-system. Figure 2.9 indicates that by applying load sharing, service quality became balanced in the two sub-systems. We note that these results do not reveal unexpected phenomena in the system under study. They show, however, that the hybrid-hierarchical environment allows for the analysis of a system at both the network and packet level, without an unacceptable simulation time. With this setting, our 500-minute simulation took 300 to 700 minutes, depending on the total offered load. A complete device level simulation of the same setting would take approximately 120 minutes per simulated minute resulting in a total simulation time of approximately 42 days, making the analysis infeasible. The speed-up is a result of focusing the device level simulation power to points in time where service quality is likely to be violated and omitting device level simulations when this is not the case.

2.5 Discussion

In this chapter we have introduced hybrid-hierarchical simulations, a simulation technique relying on hierarchical decomposition and on including analytical estimations into the simulator. We have shown that hybrid-hierarchical simulation provides great benefit in cases where a complex system is to be simulated with the primary objective of studying high level parameters but

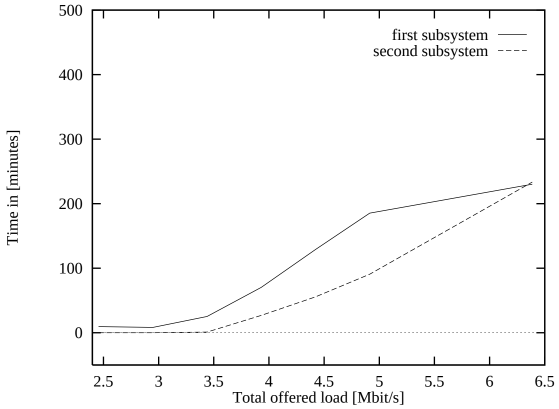

Figure 2.9: QoS violation with load sharing [min]

network operation is constrained by bounds on a set of lower level parameters (e.g., per connection service quality parameters). If periods of “overload” (when these bounds are violated) are isolated during simulation, the majority of simulation is performed at network level and the low level simulator is only used to zoom in on events of interest. In such cases, the hybrid-hierarchical simulator can achieve a significant speed-up that eventually makes it capable of performing analysis tasks that are infeasible using traditional simulation approaches. The speed-up, however, depends largely on the amount of time where device level simulation sessions are necessary. If the system under study is operating in continuous overload, the hybrid-hierarchical simulation comes back to a pure device level simulation and becomes inefficient. We conclude that the hybrid-hierarchical simulation environment can not be considered as a general method to increase simulation speed, but can provide great benefit in cases where the studied system and traffic conditions allow for its exploitation. Further work is needed to determine the appropriate conditions (e.g., critical thresholds) that trigger device level simulation sessions in complex simulation problems. This will become increasingly important in the case of future micro and pico cellular multi-application systems which must be carefully planned and managed to fulfill expected service quality constraints while using system resources efficiently. Resource efficiency of such systems is analyzed and is related to user call and mobility characteristics in the following chapter.

Chapter 3

An Efficiency Bound of Cellular Mobile Systems

In order to limit handoff blocking probability, in cellular systems admission control can be applied to new call requests and resources reserved for future handoff attempts. Such resource reservation strategies have been widely studied in the literature. In this chapter we show that due to this kind of advance reservation the throughput of a cellular system is upper bounded even if no constraint is specified on the blocking probability of new call requests. Reduced throughput decreases the network operator's revenue and creates a trade-off with handoff performance. In what follows, we analyze this trade off. In particular, we study the impact of user mobility on a cellular system's efficiency which we define as the system's throughput normalized by the theoretically optimal throughput at zero user mobility (see Section 3.3). As a reference case, we first determine the efficiency of a hypothetical system with deterministic advance reservation. Next, we consider systems with statistical reservation and calculate the maximum achievable efficiency, under the set of possible local admission control strategies subject to a hard constraint on the probability of call failure due to handoff blocking. Local admission control strategies are investigated because they represent efficient, yet simple strategies that lend themselves to easy implementation. We note that trunk reservation is a special case of local policies. We relate the calculated efficiency bound to call and mobility characteristics and show that the achievable efficiency decreases with decreasing cell size and with increasing proportion of multimedia calls.

3.1 Problem Statement

Two important quality measures of cellular mobile systems are the ratio of blocked new call requests and the ratio of calls blocked at a handoff attempt due to unavailability of radio resources. The goal of the cellular operator is to maximize network efficiency subject to constraints on these quality measures. Recently, a number of control algorithms for cellular mobile systems have been proposed and analyzed in the literature. In [1], three different formulations of this problem are addressed: MINOBJ to minimize a linear objective function of the two blocking probabilities, MINBLOCK to minimize, for a given amount of radio resources, the new call blocking probability subject to a hard constraint on handoff blocking probability and MINC to minimize the amount of radio resources subject to hard constraints on both blocking probabilities.

A general assumption in most of these approaches is that regarding system quality as experienced by the mobile user, handoff blocking is more disturbing than the blocking of new calls. This is not only due to the fact that a blocked call is more annoying than an unsuccessful call attempt, but because a call may have to perform a large number of handoffs during its lifetime (i.e., holding time). A call can be considered successful if it is successfully established and it does not get blocked during any of the handoff attempts it initiates. Therefore in the call success ratio, the handoff blocking probability has increasing importance at increasing handoff frequency.

In [2], for example, a global system performance measure is used which is defined as a weighted sum of the two blocking probabilities with higher weight for handoff blocking. To limit handoff blocking probability, most researchers solicit admission control for new call requests and resource reservation for future handoff attempts. An overview of related studies is presented in Section 3.2. Some of these proposals also show that due to the protection of handoff attempts the efficiency of the cellular system will be lower than what could be achieved if handoff and new call attempts were treated equally. In the shadow-cluster concept proposed in [3], for example, by increasing the size of the shadow clusters efficiency is traded for handoff success probability.

In this chapter we study the relationship between resource efficiency and user mobility. To measure the impact of mobility and handoff on cellular resource efficiency, we consider a system where a hard constraint is specified for the ratio of handoff blocking but no constraint exists on the new call blocking probability. The efficiency limit in this system is clearly a consequence of mobility only and is in this sense an upper bound on achievable cellular efficiency. We compute the maximum achievable resource efficiency of this system under the possible local resource control strategies. We relate this maximum efficiency to call and mobility characteristics and show that the efficiency decreases with increasing call holding time and with decreasing cell sizes. This indicates that the growing importance of mobile multimedia services and emergence of micro and pico-cellular systems may also bring increased costs for cellular operators.

3.2 Related Work

The importance of protecting handoff attempts is shown in [4] by calculating handoff blocking probabilities without protection of handoff attempts and with two schemes of handoff prioritization. In particular, a scheme based on trunk reservation is analyzed where new call attempts in a cell are discarded unless a pre-defined number of channels are available and a modified version of this scheme is also evaluated where handoff attempts may be queued if no channels are available. It is found that both schemes reduce handoff blocking probability significantly compared to the unprioritized case and that the scheme with queueing performs better than the pure reservation scheme. The trunk reservation method is compared to some heuristic admission control algorithms in [5]. The authors show through simulation results that when the network load is known, none of the examined heuristic algorithms performs better than the trunk reservation method that is simplest because it requires no information on the state of cells other than where the call is originated. If the load varies in an unpredictable way then this scheme is outperformed by a proposed “hybrid” algorithm that combines heuristics with an extension of the trunk reservation scheme. While in trunk reservation a limit is specified for the number of calls in progress in the cell where a new call is originated, in this extension a limit exists for the weighted sum of the number of calls in cells within a certain distance from the originating cell.

In [6] it is assumed that the wired infrastructure is ATM-based and the concept of virtual connection trees is introduced. It is suggested that at call setup a virtual connection be established not only to the base station that the mobile is actually connected to but to a number of other base stations to which handoff may be performed soon. The connections to these base stations are organized in a tree to decrease the performance penalty. Though virtual connections are established to the unused base stations, these connections do not carry traffic which allows for their statistical multiplexing with other unused branches of the tree. The performance measure to be analyzed is then the probability of cell overload, that is the situation when the number of active mobiles in a cell exceeds the available capacity, under a given mobility pattern. A very similar idea is used in the strategy based on "branch connections" in [7]. In [8] pro-active and re-active handoff control schemes are distinguished. In the re-active control resource allocation must be performed at each handoff and there is no guarantee that the handoff attempt will be accepted. In the pro-active control the resource allocation at call setup must ensure that the call can maintain its quality throughout its duration even if a large number of handoffs are performed. As an example of pro-active control, the virtual connection tree concept is analyzed.

A similar approach is taken in the shadow cluster concept proposed in [3] where for each active mobile resources are allocated in the vicinity of its current location and in the direction of its motion. The amount of resources to allocate in these cells is based on estimates of the probability of the mobile actually entering a given cell.

In [9] services offered to the mobile user fall into three categories. Mobility Independent Guaranteed flows and Mobility Independent Predictive flows receive guaranteed and predicted services respectively, despite mobility, as long as the moves are in accordance with the mobility specification given at call setup time. Resources reserved for these classes but not in use can be used by Mobility Dependent Predictive service flows.

In [1] three different objective functions for cellular Call Admission Control policies are defined and analyzed. It is shown in a single-class Poissonian environment that the guard channel policy (trunk reservation) and the limited fractional guard channel policy are optimal admission control policies under the studied objective functions.

Call admission control policies to decrease handoff blocking probability are analyzed in [10] in an environment with dynamic channel allocation and a single class of calls. The call admission problem is formulated there as a Markov decision process. Still under the dynamic channel allocation assumption [11] proposes genetic algorithms to avoid the computationally inefficient Markov decision process and to find a good though not optimal solution. The approach is extended in [12] where call admission control decision in a cell must be based on local information defined as the state information of the cell itself and of cells within a certain distance.

In most of these approaches resource control strategies are proposed and the resulting new call blocking and handoff blocking probabilities evaluated. It is generally accepted that handoff blocking is more undesirable than the blocking of new call requests and that handoff attempts must be protected. In [12], the notion of "relative penalty factor" is used to give larger weight to handoff blocking in the overall performance measure than to new call blocking.

Most proposals also show that this "protection" results in lower efficiency of the cellular resources than what could be achieved if handoff attempts and new call requests were treated equally. The primary difference between this work and the existing approaches is that instead of advocating a resource control policy, we attempt to point at an intrinsic *efficiency limit of cellular mobile networks* and relate this limit to user mobility and call characteristics. In particular, we search for an upper bound of achievable efficiency, under any local resource control scheme, subject to a hard constraint on the ratio of calls blocked at a handoff attempt.

3.3 Model

We consider cellular mobile systems where connectivity to mobile users is provided by means of radio links to base stations that reside on a wired network. The geographic area served by the system is subdivided into cells, each of which is served by a single base station. We will assume that a mobile user is always located in exactly one of the cells and that handoff happens in zero time. Soft handoff and handoff queueing are excluded. Further, we assume that a cell i ($i = 1 \dots N$) is permanently assigned a set of logical channels which together give the cell's capacity D_i, unchanged during operation. Systems with dynamic channel allocation are outside the scope of the model.

Calls originated by a user on the wired network and destined towards a mobile user are not distinguished from calls generated by the mobile user towards a wired user. For convenience of discussion we will assume that all calls are generated by the mobile user. Mobile-to-mobile calls need not be addressed separately because from a cellular resource efficiency point of view they can be considered as two independent calls. Group calls are also outside the scope of the present discussion. We assume that the proportion of mobile-to-mobile calls is small, hence traffic generated by different mobile users can be considered independent.

Mobile users generate calls independently. Each user generates calls of type j $(j = 1 \ldots K)$ according to a Poisson arrival process with intensity λ_j. Unless earlier blocked at a handoff, the holding time for a call of type j is an exponentially distributed random variable with mean $1/\mu_j$. The call in progress occupies b_j out of the capacity of the cell where the mobile user is located. It is assumed that blocked calls are immediately cleared from the system.

We assume that cells are of hexagonal shape with radius R (defined as the largest distance from the center of the cell to any point at its edge). If ρ denotes the density of mobile users then new calls of type j are generated in a cell at rate $3\sqrt{3}R^2\rho\lambda_j/2$ [4]. We use an infinite population model meaning that the rate at which new calls are generated does not depend on the number of calls in progress. Though in a micro-cellular system the number of users at a time in a cell may be small, this assumption is still realistic assuming that users often migrate between cells.

Mobility is widely modelled in the literature by assuming that for a user the time remaining until the next handoff is exponentially distributed; see for instance [2], [6], [10], [12]. In [4] a more detailed mobility model is established where users move at a constant speed and direction within a cell but they randomly select a new speed and direction from uniform distributions $[0, V_{max}]$ and $[0, 2\pi]$, respectively, at cell borders. However, in the next step an exponential model is fitted to this model and is used in the analysis. Building partly on these results we assume here that the time spent in a cell is exponentially distributed with the mean proportional to the cell radius and inverse proportional to the average speed (v):

$$T_H = \eta \frac{R}{v} \tag{3.1}$$

where η is constant for the analysis. In the numerical examples $\eta = 2$ will be used. We assume that the system does not make predictions of the users' movements.

The efficiency of the cellular system will be defined as the average total used bandwidth in the system (i.e., throughput) normalized by the total available system capacity which represents the optimal throughput if there is no mobility or advance reservations are not applied:

$$\epsilon = \frac{1}{\sum_{i=1}^{N} D_i} \cdot \lim_{T \to \infty} \frac{1}{T} \int_T \sum_{k \in \mathbf{S(t)}} b_k dt \tag{3.2}$$

where $\mathbf{S(t)}$ is the set of active calls at time t and b_k is the bandwidth of the kth call. We will assume that a hard constraint P_f^{max} is specified for P_f, the probability that a once established call will be blocked at a handoff instead of being successfully completed and we will search for the maximum achievable ϵ under this constraint.

3.4 Analysis

In the following section we analyze the performance of cellular systems with resource reservation for future handoff attempts. As a reference case, we first consider a system where resources for a call must be reserved in advance in all cells that the mobile user will visit throughout the duration of the call. This can be regarded as the limiting case of all possible strategies because it provides zero handoff blocking probability. We analyze this special system in Section 3.4.1. In order for a strategy to be applicable, it must require little knowledge of the state of the system and no information about future behavior of mobile users. The strategies that are optimal in terms of implementation complexity are those that perform admission control based purely on local information; that is, to decide if a new call is admitted in a system, they only rely on information that is available in the cell where the new call requests admission. This set of strategies receives special attention in the literature. In Section 3.4.2 we consider strategies of this kind and determine the maximum resource efficiency that is achievable subject to a hard constraint on handoff blocking probability.

3.4.1 Efficiency of Systems with Deterministic Advance Reservation

The mean number of handoffs performed during a call of type j is calculated as the expected number of arrivals from a Poisson process (handoff arrival) during an exponentially distributed time interval (the call duration) and can therefore simply be written as

$$M_j = \frac{1}{\mu_j} \cdot \frac{1}{T_H} = \frac{v}{\eta R \mu_j} \tag{3.3}$$

using the notations introduced in Section 3.3. In the following we will approximate the mean number of cells visited during a call of type j by $M_j + 1$ and neglect the fact that the real number can actually be smaller if the mobile visits a cell more than once during the call. If resources are reserved in advance in all the cells that the mobile will visit during the call then the average total bandwidth reserved for a call of type j with bandwidth requirement b_j must be

$$B_j = b_j(M_j + 1) \tag{3.4}$$

during the entire duration of the call. Since λ_j/μ_j is the offered traffic per mobile user of call type j, the average total used (resp. reserved) bandwidth in the system $\overline{b}$ (resp. $\overline{B}$) can be written as

$$\overline{b} = \sum_{j=1}^{K} A\rho \frac{\lambda_j}{\mu_j} b_j \tag{3.5}$$

$$\overline{B} = \sum_{j=1}^{K} A\rho \frac{\lambda_j}{\mu_j} B_j \tag{3.6}$$

where A is the size of the covered area. As we can never reserve more capacity than what is available in the system $\overline{B} \leq \sum_{i=1}^{N} D_i$, and the maximum achievable efficiency is

$$\epsilon = \frac{\overline{b}}{\overline{B}} = \frac{\sum_{j=1}^{K} \frac{\lambda_j}{\mu_j} b_j}{\sum_{j=1}^{K} \frac{\lambda_j}{\mu_j} b_j (M_j + 1)} \tag{3.7}$$

Since $M_j = v/(\eta R \mu_j)$ this shows that the efficiency decreases hyperbolically with the decrease of cell radius or the increase of call holding time, both of which are trends to be expected. In the case of $K = 1$ for instance

$$\epsilon^{(1)} = \frac{1}{M+1} = \frac{\eta R \mu}{v + \eta R \mu} \tag{3.8}$$

Figure 3.1 illustrates $\epsilon^{(1)}$ as a function of $1/\mu$ and R. The efficiency decreases hyperbolically with increasing mobility.

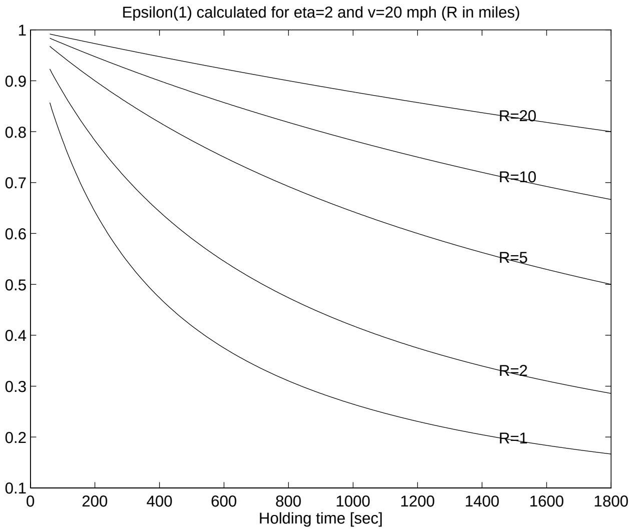

Figure 3.1: Efficiency as a function of call holding time. Single class of calls.

Further, there is a penalty for systems where calls with large bandwidth requirement are at the same time of typically longer holding time, as will normally be the case if the proportion of multimedia calls increases. To demonstrate this penalty we can calculate ϵ in the $K = 2$ case:

$$\epsilon^{(2)} = \frac{\eta R\overline{b}/(A\rho)}{\lambda_1 b_1 \frac{1}{\mu_1^2}(v + \eta R\mu_1) + \lambda_2 b_2 \frac{1}{\mu_2^2}(v + \eta R\mu_2)} \tag{3.9}$$

If the calls have parameters $\mu_1 = \mu/t, b_1 = bt, \lambda_1 = \lambda/2t^2$ and $\mu_2 = \mu t, b_2 = b/t, \lambda_2 = \lambda t^2/2$ then by varying the parameter t we can see the effect of an increasingly heterogeneous set of calls. The case $t = 1$ falls back to the single application system and with increasing t call type 1 will be more and more a "long holding time, large bandwidth but low call intensity" multimedia call and call type 2 a more "ordinary low bandwidth, high call intensity, short holding time" phone call, such that the average total used bandwidth $\overline{b}$ is unchanged. We will then have, after a few steps

$$\epsilon^{(2)} = \frac{\eta R\mu}{v\frac{t+1/t}{2} + \eta R\mu} \tag{3.10}$$

which is clearly inferior or equal to $\epsilon^{(1)}$ with the equality holding in the $t = 1$ case. Figure 3.2 illustrates the decrease of efficiency with t. For $t = 1$ the efficiency is that of a single-class system, and with increasingly heterogeneous call characteristics the efficiency decreases.

To understand the relationship between efficiency and the cell size a little further we recall that in cellular systems "bandwidth density" and hence the overall system capacity can be increased by decreasing the cell size. It is expected that due to the increasing demand for bandwidth-demanding mobile applications the cell sizes will further decrease in the future and micro-cell and pico-cell networks will be built. If γ denotes the cluster size (defined as the number of cells in the smallest area where the same logical channels can not be reused) and D_{total} is the amount of capacity that base stations can physically support then the capacity available in a single cell in

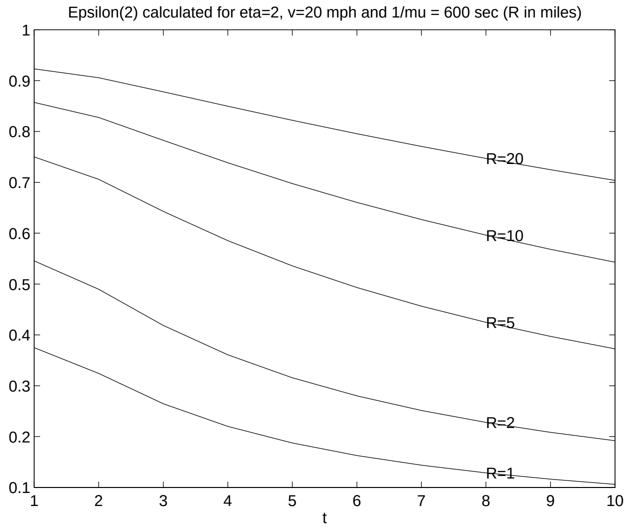

Figure 3.2: System efficiency vs. diversity of call characteristics

an equal sharing case is D_{total}/γ and the "bandwidth density" defined as the available bandwidth per area unit is

$$D = \frac{D_{total}}{\gamma \frac{3\sqrt{3}}{2} R^2} \tag{3.11}$$

In order to keep up with the increasing demand for capacity, the cellular operator will have to decrease the cell radius R such that $1/R$ will be proportional to the square root of the demand increase. Considering the relationships we have already established this means that an increase in the total offered traffic decreases system efficiency even if the structure of the traffic in terms of call types is unchanged. This is because the decrease of R increases $\overline{B}$. If, for instance, the arrival rate λ_j for all call types increases evenly to $\lambda'_j = t\lambda_j$, then efficiency will decrease to $\epsilon' = \frac{1}{\sqrt{t}}\epsilon$. Further, an increase in connection holding times decreases system efficiency not only through its direct effect on $\overline{B}$ that we have seen before but also through its effect on R which in turn also increases $\overline{B}$.

In the deterministic advance reservation case the following main phenomena were found:

- efficiency of cellular resources decreases with increasing call holding time and with decreasing cell size;
- efficiency decreases with increasing call diversity, i.e., if some calls are long holding time large bandwidth multimedia calls, others are traditional voice calls; and
- increasing offered traffic decreases efficiency through the necessary decrease of cell size.

3.4.2 Efficiency of Single Application Systems with Statistical Local Admission Control

In this section we focus on local admission control strategies and determine the strategy that ensures maximum efficiency subject to a hard constraint on the handoff blocking probability. Using this optimal strategy, we show that the system's resource efficiency decreases with increasing mobility regardless of the new call blocking probability.

Call failure probability

Let $P_{h,j}$ denote the probability that a handoff attempt of a call of type j is blocked. As pointed out in [4] this probability is not directly the service quality as perceived by the user. For the mobile user the important quality measure is the probability that a once successfully established call gets blocked at one of its handoff attempts. We will refer to this probability as "call failure probability" and will denote it $P_{f,j}$ for calls of type j. Of course, $1 - P_{f,j}$ gives the ratio of established type j calls successfully terminated when the conversation is finished.

If h is the number of handoffs throughout the duration of the call then

$$P_{f,j} = 1 - (1 - P_{h,j})^h \tag{3.12}$$

As h is itself a random variable, namely the number of arrivals from a Poisson process with rate $1/T_H$ during an exponentially distributed time with mean $1/\mu$, the call failure probability can be written as

$$\begin{aligned}
P_{f,j} &= 1 - \sum_{h=0}^{\infty} (1 - P_{h,j})^h P(h \textit{ handoffs during the call}) = && (3.13)\\
&= 1 - \sum_{h=0}^{\infty} (1 - P_{h,j})^h \int_0^{\infty} P(h \textit{ handoffs} \mid \textit{holding time is } x) \cdot \mu_j e^{-\mu_j x} dx = && (3.14)\\
&= 1 - \sum_{h=0}^{\infty} (1 - P_{h,j})^h \int_0^{\infty} \frac{(x/T_H)^h}{h!} e^{-x/T_H} \cdot \mu_j e^{-\mu_j x} dx && (3.15)
\end{aligned}$$

which evaluates as

$$P_{f,j} = \frac{vP_{h,j}}{\eta R \mu_j + vP_{h,j}} = \frac{P_{h,j}}{P_{h,j} + 1/M_j} \tag{3.16}$$

Handoff attempts into a cell

We will approximate the process of handoff attempts arriving to a cell by a Poisson process with a rate calculated as follows. The rate of handoff attempts to the ith cell is the sum of the rates of handoff attempts to the ith cell from its neighbouring cells, the set of which is denoted $\mathbf{N_i}$. If $\beta_{i,k}(t)$ is the rate at which calls perform handoff from the kth cell into the ith cell, then

$$\beta_i(t) = \sum_{k \in \mathbf{N_i}} \beta_{i,k}(t) \tag{3.17}$$

Denoting by $n_k(t)$ the number of active calls in the kth cell at time t the rate $\beta_{i,k}(t)$ would in a homogeneous system be $n_k(t)/6T_H$ if all calls were sooner or later to perform a handoff. Some terminals, however, terminate the call before performing a handoff. The rate $1/T_H$ can be considered as the handoff rate of a terminal during an active call (as used in the conditional expression of Equation 3.13) but it is not equal to the rate at which terminals in a cell perform handoff. For ease of explanation let us first assume that the holding times are equal for all call types. Then the rate of handoff attempts from the kth cell to the ith cell is approximated by

$$\beta_{i,k}(t) \approx \frac{n_k(t)}{6T_H} P_H \tag{3.18}$$

where P_H is the probability that a mobile user currently engaged in a call will perform a handoff before the call is terminated and can be calculated as[1]

$$\begin{aligned} P_H &= \int_0^\infty P(\textit{call remains active for at least x time}) \cdot \frac{1}{T_H} e^{-x/T_H} dx = & (3.19) \\ &= \int_0^\infty e^{-\mu x} \frac{1}{T_H} e^{-x/T_H} dx = & (3.20) \\ &= \frac{1}{1+\mu T_H} & (3.21) \end{aligned}$$

Hence

$$\beta_{i,k}(t) \approx \frac{n_k(t)}{6T_H} \cdot \frac{1}{1+\mu T_H} \tag{3.22}$$

We can release the assumption of equal holding times and get the rate of type-j call handoff attempts from the kth cell into the ith cell as

$$\beta_{i,k}^{(j)}(t) \approx \frac{n_k^{(j)}(t)}{6T_H} \frac{1}{1+\mu_j T_H} \tag{3.23}$$

where $n_k^{(j)}(t)$ is the number of type-j calls in the kth cell at time t.

Optimal Strategy

We will now consider a system of a single application ($K = 1$) where admission decision for new calls is based uniquely on the state of the cell where the call is generated. We will search for a new call admission control strategy that is optimal in the sense that it provides the highest system efficiency subject to a constraint on the handoff blocking probability. The algorithm will be searched for over the set of possible Fractional Guard Channel algorithms which represent the broadest set of possible local policies [1]. For simplicity we will assume $b_1 = 1$. The number of active calls in the ith cell at time t will be denoted by $n_i(t)$.

At time t in the ith cell

- new call requests are generated at rate $\alpha = \frac{3\sqrt{3}R^2\rho}{2}\lambda$
- calls terminate at rate $n_i(t)\mu$
- active calls leave the cell with handoff at rate $\frac{1}{T_H}n_i(t)$
- handoff attempts arrive at rate $\beta_i(t)$

[1] We note that using this probability we can calculate the mean number of handoffs throughout the duration of a call as the expected value of a geometrically distributed random variable, and will re-obtain, as expected, M_j. In addition, it can be used to obtain the expression for $P_{f,j}$.

As we are searching for an upper bound of achievable efficiency, we may assume that the offered load is higher than what could be taken by the system and that the call admission control has to drop some new call attempts. The dropping decision is now based only on local information which in this memoryless system is the number of active calls, $n_i(t)$. Let $\alpha'(n)$ denote the rate at which new call attempts are accepted in a cell when the number of active calls is n. We note that trunk reservation is a special case of this kind of policy. The admission policy where the number of active calls determines the rate at which new call requests are admitted into the network is referred to in [1] as "Fractional Guard Channel" policy.

It is important to differentiate between short and long term behaviour of the system. On a short term the rate of handoff attempts into the cell is independent of the state of the cell, the rates of call termination and handoffs out of the cell are proportional to the number of active calls and the rate of admitted new call requests is determined by the non-increasing $\alpha'(n)$ function. Handoff blocking being a result of temporary overload, its probability must be calculated based on these rates.

The efficiency of cellular resources, on the other hand, is related to the long term behaviour of the system. Looking at a long time scale we can assume that load in the system is uniform. Let us denote by n^* the long term average of active channels in a cell. Clearly, n^* is directly related to the efficiency because in the single class system with uniform cells

$$\epsilon = \frac{1}{\sum_{i=1}^{N} D_i} \cdot \lim_{T\to\infty} \frac{1}{T} \int_T \sum_{i=1}^{N} n_i(t)dt = \frac{n^*}{D} \tag{3.24}$$

Using Equation 3.22, the rate of handoff attempts into a cell is then

$$\beta(n^*) = \frac{n^*}{T_H} \cdot \frac{1}{1+\mu T_H} \tag{3.25}$$

and the long-term balance of the system is determined by the equation[2]

$$\alpha'(n^*) + \beta(n^*) = (\mu + \frac{1}{T_H})n^* \tag{3.26}$$

The system is in balance at n^* if

$$\alpha'(n^*) = (\mu + \frac{1}{T_H})n^* - \beta(n^*) = \mu n^* \frac{2+\mu T_H}{1+\mu T_H} \tag{3.27}$$

Figure 3.3 illustrates that if $\alpha'(n)$ is non-increasing, there is a single point of balance and that is stable because moving towards higher n the rate of accepted new calls will become inferior to the total departure rate less incoming handoffs.

Intuitively, of the set of non-increasing $\alpha'(n)$ functions with fixed n^* the one resulting in smallest handoff blocking is the one that admits the least number of new call requests for any n, denoted in the figure $\alpha'_{\text{step}}(n)$.

$$\alpha'_{\text{step}}(n) = \begin{cases} n^* \cdot \mu \frac{2+\mu T_H}{1+\mu T_H} & \text{if } n \le n^* \\ 0 & \text{else} \end{cases} \tag{3.28}$$

Looking for an upper bound on n^* subject to a constraint on handoff blocking probability but independent of the arrival process and of the new call blocking probability, we can assume that $\alpha'(n)$ is of the form $\alpha'_{\text{step}}(n)$. This actually is a special case of fractional guard channel

[2] We assumed here that the ratio of dropped handoff requests is negligible in terms of total load. Of course the dropped new call requests are not neglected.

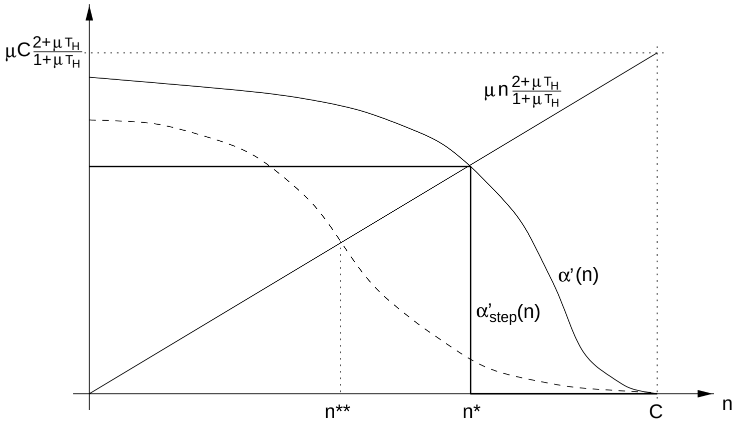

Figure 3.3: Long-term balance of the single class system

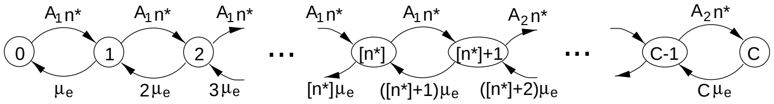

Figure 3.4: Markov chain of the single class system with $\alpha'_{\text{step}}(n)$

policy where the threshold is placed at the mean occupancy and the rate of calls admitted is proportional to the mean occupancy. The state changes in the ith cell can then be modelled by the Markov chain as shown in Figure 3.4 where

$$\begin{aligned} A_1 &= \frac{1}{T_H} \cdot \frac{1}{1+\mu T_H} + \mu \frac{2+\mu T_H}{1+\mu T_H} = \mu + \frac{1}{T_H} && (3.29) \\ A_2 &= \frac{1}{T_H} \cdot \frac{1}{1+\mu T_H} && (3.30) \\ \mu_e &= \mu + \frac{1}{T_H} && (3.31) \end{aligned}$$

and $[x]$ denotes the largest integer that is inferior or equal to x. This Markov process is similar to, but is more general than the process associated with a trunk reservation system. As the arrival process of handoff attempts is independent of the actual state of the cell and handoff attempts are blocked iff there are no available channels left, the probability of handoff blocking is equal to the probability $P(D)$ of being in state D of this Markov chain. The problem is now reformulated as searching for the largest n^* such that the probability $P(D)$ be inferior to the constraint on handoff blocking probability. The steady-state probability of being in state D of this birth-death process can be calculated as usual:

$$P(D) = P(0) \cdot \frac{A_1^{[n^*]+1} A_2^{D-[n^*]-1} (n^*)^D}{\mu_e^D D!} = \tag{3.32}$$

$$= \frac{1}{\mu_e^D D!} \cdot \frac{A_1^{[n^*]+1} A_2^{D-[n^*]-1} (n^*)^D}{\sum_{i=0}^{[n^*]+1} \frac{(A_1 n^*)^i}{\mu_e^i i!} + \sum_{i=[n^*]+2}^{D} \frac{A_1^{[n^*]+1} A_2^{i-[n^*]-1} (n^*)^i}{\mu_e^i i!}} \tag{3.33}$$

and the condition for n^* can be written as

$$P_h = \frac{\frac{(n^*)^D}{D!} \theta^{D-[n^*]-1}}{\sum_{i=0}^{[n^*]+1} \frac{(n^*)^i}{i!} + \sum_{i=[n^*]+2}^{D} \frac{(n^*)^i}{i!} \theta^{i-[n^*]-1}} \leq P_h^{max} \tag{3.34}$$

where

$$\theta = \frac{1}{(1+\mu T_H)^2} = (\frac{M}{M+1})^2 \tag{3.35}$$

takes its values from the interval (0,1) and is closer to 1 if the average number of handoffs during a call is higher. In accordance with intuition for $\theta \to 0$ the handoff blocking probability approaches zero and for $\theta \to 1$ the handoff blocking probability will be $E_D(n^*)$, that is equal to the blocking probability of an Erlang system with D channels and n^* offered traffic. This is because with very long holding times (or very frequent handoffs) the call attempts made in a cell will consist primarily of handoff attempts which then compete for all D channels as a single class of calls with Poisson arrival and the mean occupancy is n^*.

Numerical Results

We recall that n^* is defined as the mean number of active channels, that is, of established calls, in a cell and is therefore directly related to the efficiency. By searching for the largest n^* satisfying Equation 3.34 we determine an upper limit of the mean number of active channels per cell under any local call admission control policy. This is plotted in Figure 3.5, as a function of θ. For $\theta \to 0$ we see that n^* approaches D and for $\theta \to 1$ it reaches a value determined by the Erlang system. One might expect that for $P_h^{max} \to 0$ the efficiency approaches that of the system with deterministic reservation. This is, however, not the case because our statistical model fails for extreme low values of P_h^{max}. Loosely speaking, if we require zero handoff blocking, then the whole system should never take more connections than what can be supported by a single cell and this can not be achieved by any local policy.

However, the handoff blocking probability is of little meaning for a mobile user. In a real system the hard constraint is specified on the probability of a call being dropped at a handoff attempt instead of successfully terminating. The relation of this "call failure" probability, P_f, to the handoff blocking probability P_h is given by Equation 3.16. Using this relationship we can calculate from Equation 3.34 the maximum achievable efficiency $\epsilon = \frac{n^*}{D}$, as a function of the mean number of handoffs per call M, under a hard constraint on the call failure probability. The result is plotted in Figure 3.6 as a function of M and in Figure 3.7 as a function of P_f.

3.5 Discussion

In this chapter we have studied the relationship between the efficiency of a cellular mobile system and user mobility. Our attention was limited to cellular systems using fixed channel assigment, hard handoff and no handoff queueing. We further assumed that calls are independent (e.g., mass

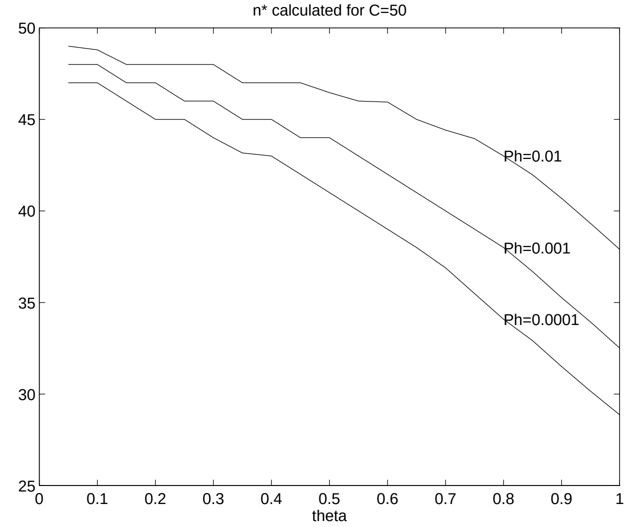

Figure 3.5: Upper limit of the mean number of active channels per cell

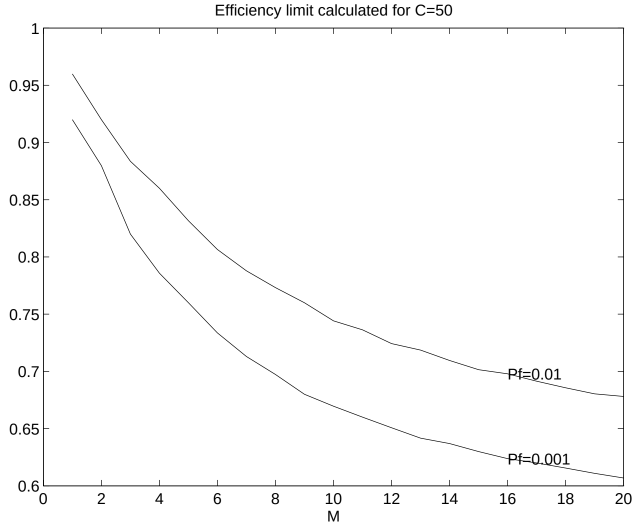

Figure 3.6: Maximum achievable efficiency vs. user mobility

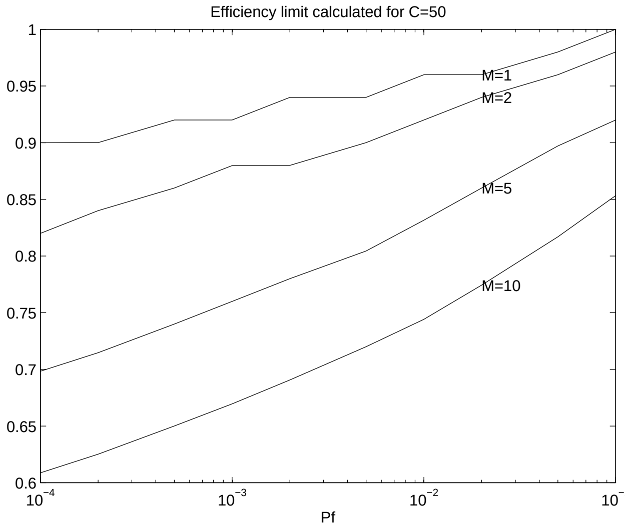

Figure 3.7: Maximum achievable efficiency vs. tolerated call failure probability

group calls are excluded). We have found that in these systems the achievable resource efficiency is bounded due to user mobility if admission control is applied for new calls in order to reduce handoff blocking probability. We have calculated the efficiency of a system with deterministic resource reservation which can be considered as a limiting case of possible admission control strategies. Next, we have studied systems with statistical local admission control. We have defined an admission control strategy that is optimal in the sense that it provides the highest system efficiency subject to a hard constraint on the handoff blocking probability. Assuming that the size of wireless cells allows for a statistical analysis, we have established a Markov model of a system using this strategy and have determined the system efficiency. This result can be interpreted as the maximum achievable efficiency of a cellular system under the set of possible local admission control strategies. We have found that this maximum depends on call and mobility parameters through M, the mean number of handoffs during a call. The dependency is plotted in Figure 3.6 and in Figure 3.7.

Figures 3.6 and 3.7 show that with increasing number of handoffs per call, which can be a result of increasing holding times or decreasing cell size, the cellular system's efficiency decreases. It is important to note that this efficiency limit is a result of user mobility only and it holds regardless of the tolerated new call blocking probability. If a constraint exists for new call blocking probability, as is typically the case in real systems, the efficiency must be below this limit.

We recall that the "optimal new call admission control strategy" we used to derive these results can not be regarded as a candidate strategy for real systems. This strategy is optimal only in the strict sense of providing highest efficiency. It is, however, not a suggested strategy for real networks because

- it does not provide any limit on the blocking probability of new calls;
- it assumes the knowledge of the system efficiency ϵ which is, a parameter to measure in a real system; and

- it assumes control over the rate at which new call attempts in the system arrive.

Though these properties inhibit using the optimal strategy as a real-time admission control algorithm, they do not, however, limit its application in network planning. In the planning phase, the future system efficiency is normally known with sufficient accuracy. Substituting the engineered efficiency for ϵ and the engineered handoff blocking probability for P_h, the optimal strategy as defined previously can be used to calculate the new call arrival rate that the system supports. This is a valuable input to determine the pricing structure that makes the given system operate optimally.

Though local strategies are the simplest and most attractive solutions to the advance reservation problem, there are other alternatives that were not addressed in this work. Simulation results in [5] indicate that there are heuristic algorithms that in some cases outperform the local strategy by taking into account the state of adjacent cells. These heuristics are based on the observation that the admission of a call may be advantageous even if it is generated in a highly loaded cell, providing that adjacent cells have large free capacity. Though a number of algorithms are proposed in the literature that rely on this observation, most of these use heuristics instead of analytical evaluation because of the complexity of the problem. An extension of the results presented in this chapter for non-local strategies is therefore also a challenging task.

Another limitation of the results presented in this chapter is related to the applied system and traffic models. For the sake of obtaining a tractable model, we needed to limit our attention to fixed channel allocation systems that exclude soft handoff, handoff queueing and other mechanisms that are now gaining acceptance in state of the art mobile systems. Handoff queueing, for example, was shown in [4] to improve the performance of the trunk reservation strategy. A natural extension of work presented here is to include these advanced techniques which will be addressed in the next phase of this work.

Chapter 4

An Architecture and Protocol for Cellular Wireless Data Networks

Recent initiatives to add mobility to the Internet and to increment cellular telephony systems with packet data services indicate that there is an increasing interest in wireless mobile packet data networks. Such networks could be used to provide wireless LAN services in an office or campus environment, by emerging wireless Internet Service Providers or in IP based mobile telephony networks. This diversity of applications requires a flexible and scalable solution. While Mobile IP provides a simple, scalable mobility scheme, it is not appropriate to support fast, seamless handoffs. Third generation cellular systems offer smooth mobility support, but are built on a complex infrastructure that lacks the flexibility expected by the Internet community. In this chapter, we identify the key requirements of mobility solutions for a future ubiquitous Wireless Internet and derive design principles for a cellular wireless IP access network. Building on these design principles we propose a cellular packet data network architecture and protocol that rely on cellular telephony concepts but implement them based on the IP paradigm.

4.1 Problem Statement

The development of commodity-based palmtop devices with built in high-speed packet radio access to the Internet will have a major impact on the mobile telecommunications industry and the way we communicate. The availability of cheap, ubiquitous and reliable wireless Internet access will shift the service base traditionally found in mobile telecommunication networks toward emerging wireless Internet Service Providers (ISPs). Large numbers of mobile users equipped with wireless IP enabled communicators will have access to a wide array of web based mobile multimedia services anywhere anytime. This environment places significant demand on existing and next-generation cellular and IP networking solutions. In this chapter we study an Internet host mobility concept that takes an alternative approach to that found in mobile telecommunications (e.g., General Packet Radio Service [13]) and in IP networking (Mobile IP [14]).

Recent initiatives to add mobility to the Internet mostly focus on the issue of address translation through introduction of location directories and address translation agents [15]. In these protocols (e.g., Mobile IP) packets addressed to a mobile host are delivered using regular IP routing to a temporary address assigned to the mobile host at its actual point of attachment. This approach results in simple and scalable schemes that offer global mobility. It is not appropriate, however, for fast mobility and smooth handoff because after each migration a local address must be obtained and communicated to a possibly distant location directory or home agent (HA).

Mobile IP, for example, requires that the mobile host's home agent be informed whenever the host moves to a new foreign agent (FA). During the update messaging phase, packets will be forwarded to the old location and will not be delivered hence disturbing active data transmission [19]. Similarly, during route optimization [26], [30], data transfer is disrupted while the correspondent host obtains a new binding. The effect of these delays grows with increasing handoff frequency. In addition, the update messages load both the Internet and the home agents even when the mobile host is idle while moving. This load is proportional to the number of mobile hosts and not to the generated data traffic. This may be a problem as host mobility becomes ubiquitous and cell sizes smaller.

Cellular mobile telephony systems are founded on a radically different concept from that of Mobile IP. Instead of aiming at global mobility support, cellular systems are optimized to provide fast and smooth handoffs in a restricted geographical area. In the area of coverage, mobile users have wireless access to the mobility unaware global telephony network. A scalable forwarding protocol interconnects distinct cellular networks to support roaming between them. However, there are fundamental architectural differences between cellular and IP networks that make the application of cellular techniques to IP challenging. Cellular telephony systems rely on a restrictive "circuit" model that requires connection establishment prior to communication. In contrast, IP networks perform routing on a per packet basis. In addition, today's cellular systems are based on hierarchical networks and use costly mobile-aware nodes (e.g., MSC). Though recent initiatives to add packet data service to cellular telephony networks (e.g., GPRS [13]) and third generation mobile systems depart from the traditional circuit oriented model in many aspects, they do not, however, offer the simplicity and flexibility expected in IP based solutions. In particular, both these systems preserve the strict hierarchical network structure and rely on expensive specialized hardware and software.

In this work we address host mobility in a scenario where a wireless connection to the Internet is typical, rather than as it is today, an exception. We therefore assume an environment where highly mobile hosts often migrate during active data transfer and expect the network to manage these handoffs with little or no disturbance to ongoing data sessions. While people rarely read text or watch video while walking or driving, they may wish to, however, download files, browse the web, or talk on the Internet phone while on the move. Leveraging experience from cellular telephony and following on from the hierarchical mobility management approaches in [19] and in [31], we assume a mobile Internet architecture where local wireless access networks handle local mobility while a mobility enabled Internet provides wide area mobility support as illustrated in Figure 4.1. In this case a mobile host's home agent is only informed when the host moves into a new access network and is unaware of the host's mobility within an access network. The main advantage of separating local and wide area mobility is that home agents need not be informed about local mobility within a wireless access network. We believe that this will become increasingly important as cells become smaller, host migration frequency faster and user population greater. By handling the majority of handoff control locally we can engineer faster handoffs and limit the impact of handoff on active data sessions while avoiding the exposure of local migration to distant home agents.

In this chapter, we will assume that Mobile IP is used in the Internet as a global mobility protocol and we will limit our attention to the wireless access network. Our objective is to create wireless access networks that leverage experience from cellular mobile telephony in order to achieve fast seamless handoff control, but implement these principles around the IP paradigm inheriting the simplicity and robustness of IP networks. Based on the vision of global wireless Internet access outlined above, we first establish a set of requirements for wireless access networks in Section 4.3. In Section 4.4, we use these requirements to derive wireless access network design principles. In Section 4.5 we outline an architecture and an associated protocol founded on the design principles and in Section 4.6 we compare its properties to our objective and vision. The

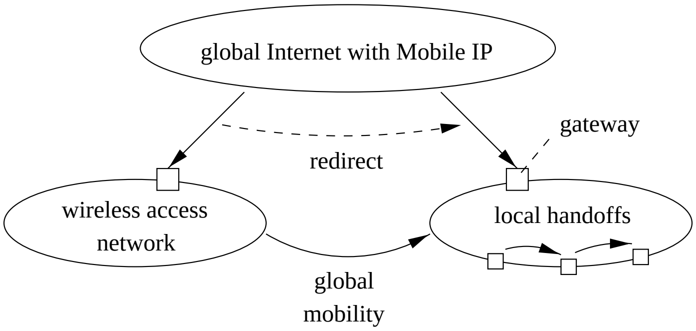

Figure 4.1: Wireless access networks and Mobile IP

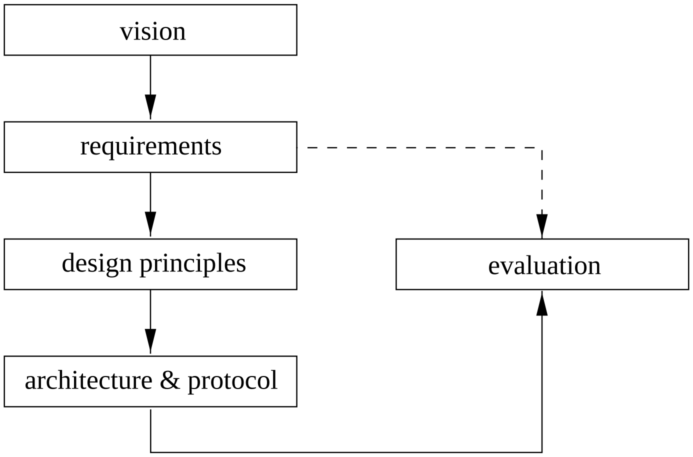

Figure 4.2: Wireless access network design methodology

process is schematically illustrated in Figure 4.2.

4.2 Related Work

To the author's knowledge, the present initiative represents one of the first attempts to create cellular mobile systems founded on IP networking principles. Our work, however, is not the only solution to provide IP data service to wireless mobile users. Second generation cellular telephony systems now provide packet data services and third generation systems are specifically designed to accommodate IP applications. (Overviews of second and third generation systems are presented in Section 1.2.) Second generation standards, however, are optimized for voice service and their infrastructure reflects this design decision. Wireless mobile telephony service can efficiently be provided in a large service area, typically on a metropolitan or country-wide scale. In these large systems, the hierarchical structure and expensive hardware that are specific to cellular telephony networks are justified by the need to cover a very large geographically area ubiquitously and with high radio resource efficiency. In contrast, wireless data services are often better to provide in a small geographical area because smaller wireless cells allow for higher user data rates. While covering very large areas with small cells is not an economical solution, in limited areas, such as a campus area or office building, small cells can be viable. These inherent characteristics of cellular wireless data services justify the concept of wireless overlay network concept described in [31]. In this concept, wireless packet data services are provided over various types of service areas and at various levels of access rates. A metropolitan area wireless ISP, for example, allows for roaming in a relatively large area but its base stations are scarcely spread and user data rates are low. Campus area systems provide higher data rates but offer connectivity in a limited area only. Systems operating inside offices and homes offer data rates comparable to wired networks, but have an even more limited footprint. In the wireless overlay network concept, these systems complement each other by offering the freedom of choice to the wireless mobile user. In the present work we adopt the concept of wireless overlay networks with the added requirement that the same protocol should be applicable from small to large areas and from low to high data rates. In addition to fulfilling the requirements of wireless overlay networks, this property can potentially allow for gradually changing small installations into large systems and it also simplifies the implementation of mobile hosts. These features, we believe, are keys to creating a truly scalable and ubiquitously available wireless data solution.

Internet and the IP protocol were originally designed for hosts with fixed point of attachment and mobility has only recently appeared as a requirement. As a natural extension to the fundamentally mobility unaware IP protocol, Internet host mobility proposals (e.g., Mobile IP) mostly focus on the requirement of global terminal portability and are less concerned about fast handoffs. This limits their applicability in cellular wireless environments and calls for extensions for the support of seamless handoffs. (An overview of Internet host mobility proposals is presented in Section 1.2.3.)

Recognizing Mobile IP's failure in supporting fast handoffs, recently a number of extensions to the base Mobile IP protocol have emerged to support local handoffs. In [17] and [18], foreign agents are structured in a hierarchy. The care-of-address known by the home agent is the address of the top of this hierarchy. Upon receiving a packet, the foreign agent at this address interacts with a local data base to determine which lower level foreign agent the packet should be forwarded to. This procedure may be repeated depending on the depth of the hierarchy. Similar ideas are adopted for campus and domain foreign agents discussed in [19] and for local registrations in the DREMIP proposal [20]. All of these proposals are extensions to Mobile IP that reduce the necessary registration messaging in the case of local handoffs. By implementing mobility related functions on top of a regular IP network, however, they are not alternatives of a cellular network where routing can be optimized for the mobile environment. In addition, location management in the hierarchical foreign agent proposals does not distinguish active and idle hosts. Foreign agents maintain data base entry for each host in the region and have to potentially search large data bases in order to route each packet. This limits scalability especially if mobile users keep their terminals switched on around the clock as is the norm today in cellular telephony.

A number of proposals have advocated the use of IP multicasting technique to achieve smooth local handoffs [21] or global mobility support [22]. In these protocols, the mobile host is identified by a multicast IP address. This multicast group is joined by base stations the host is connected or may be connected to after a handoff. In the latter case, packets are delivered to the new base station even before the host has migrated. This feature makes the multicasting based approach capable of supporting smooth handoff support, but the applicability of these solutions is restricted due to the complexity of necessary multicast enabled routers and to the shortage of multicast addresses.

The Columbia protocol [24] was designed to operate optimally in small area installations. In this protocol the base stations represent radio enabled routers of a campus area network.

Base stations broadcast search messages among each other in order to determine the location of a mobile host. By tunnelling packets between base stations, the Columbia scheme effectively creates a mobile overlay network on top of the wired campus network. This protocol works well for small number of mobile hosts but may encounter scalability problems because of the nature of its broadcast search algorithm. Using IP routers as base stations also limits scalability in terms of the number of wireless cells that can be supported. The local mobility protocol proposed by [19] uses workstations as base stations and hence is more appropriate for networks with small cells. This protocol, however, is similar to commercially available WLAN solutions [25] in the respect that it only provides mobility within the area covered by a local area network and is therefore not a possible candidate for a ubiquitous wireless Internet protocol.

4.3 Model

We assume a mobile Internet architecture where local wireless access networks handle local mobility while a Mobile IP capable Internet provides wide area mobility support. We depart from Internet host mobility solutions in the respect that we assume a wireless access network infrastructure that is dedicated for the support of cellular mobility rather than operating as an overlay over a regular IP network. Inside the access network we can use a routing algorithm optimized for cellular mobility. The access network is connected to a regular IP network, most typically to the Internet, through a *gateway* router. In order to support global mobility, that is, migrations between distinct wireless access networks, the gateway router implements Mobile IP foreign agent functionality [14].[1] Migrations from one access network to another are then handled as described in the Mobile IP specification [14] with the difference that migrations between foreign agents effectively mean migrations between wireless access networks. Upon entering an access network, mobile hosts register with their respective home agents and specify the local gateway's IP address as their care-of-address. Packets addressed to these hosts will then be tunnelled to the gateway. The gateway removes the tunnelling header and forwards packets toward the mobile host through the wireless access network. Packets in the reverse direction are routed from the gateway to the destination address using regular IP routing. Figure 4.3 illustrates the path of packets to/from a mobile host with IP address **X**. Here **BS** denotes base stations and **R** the gateway router. In what follows, we identify some key requirements for wireless access networks. A wireless network reference model is illustrated in Figure 4.4.

Universal Building Blocks

In cellular telephony networks functionality is split between various hierarchy levels and network elements (e.g., Mobile Switching Center, Base Station Controller) are optimized to operate at a certain position in the network. Mobility management and resource control functions are implemented through the interworking of all these network elements. In order for wireless access networks to operate efficiently in very small installations and still scale up to large networks, they need to depart from this approach and rely on a universal building block, as is the case in IP networks. Like regular IP routers, this universal *node* must equally be capable of operating in isolation or in a large network. A node operating in isolation does not, of course, support handoff, but can serve as a wireless access point to a regular IP network. Used in a network, nodes shall not only serve as wireless access points (i.e., base stations), but interwork to handle migrations and handoffs.

[1] In this description we assume Mobile IPv4 [14] and no route optimization [26]. The operation is just slightly different if Mobile IPv6 [30] or route optimization is used.

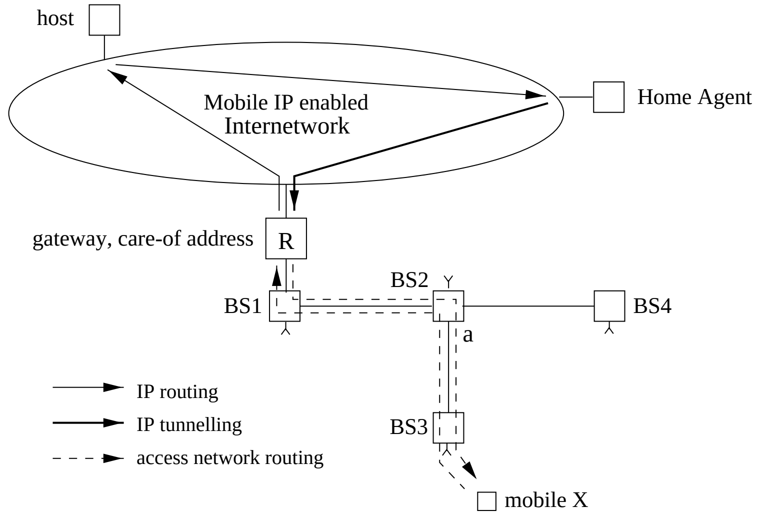

Figure 4.3: Routing to/from wireless mobile hosts

Plug-and-Play Solution

In our vision of ubiquitous wireless IP access, cellular networks may need to be built in an incremental way and network configuration may often change on the fly. System installation can not be preceded by a thorough planning phase and network operation may not require specially trained personnel. We expect wireless access networks to function in a plug-and-play manner. An arbitrary network of nodes should automatically operate as a cellular wireless access network without prior configuration. Nodes must manage their own radio resources and learn about adjacent nodes to support migrations and handoffs. Reconfigurations and changes of topology (e.g., due to element failures) should be smoothly handled without network management interaction.

Performance Scalability

In metropolitan area systems base stations will be scarcely spread and will have to cover large wireless cells. These base stations can be implemented on top of large capacity IP routers or other high-end switch hardware. In contrast, indoor systems will contain large numbers of nodes in a small geographical area and wireless cells will be small. These base stations will use short range radio devices and achieve high per user data rate. True performance transparency between different types of access networks, as defined in [33], is therefore not feasible in a wireless environment. Instead, we define performance scalability, that is the ability to use the same protocol in distinct types of environments. In order to allow smooth interworking between the different types of access networks and to support a gradual evolution of networks, the vision of ubiquitous wireless IP access assumes that the same protocol can be used in a wide range of cellular access network scenarios. A wireless ISP will likely use higher capacity nodes and more powerful radio devices than a campus system. Nodes used in these various types of networks, however, should implement the same protocol and should be capable of seamless interworking, as is the case today with high and low end IP routers. The protocol must lend itself to a simple and low cost implementation ensuring that it remains an affordable solution even in an indoor environment having a dedicated base station in each office.

Minimal Footprint in User Terminal

Notebook users represent the most typical consumer base of today's cellular wireless packet data services. With the development of commodity palmtop devices and PDAs, this base can rapidly expand and the importance of wireless Internet connection can further increase. Mobile telephones, pagers, personal communicators, and intelligent badges can all use the same ubiquitous access technology making it potentially even more powerful and affordable. This vision, however, can only become reality if user terminals are simple commodity devices performing only elementary functions in support of mobility. Ideally, a user terminal should be memoryless in the sense that it always communicates with the locally available base station but does not notice when it moved to a new base station. Though truly memoryless mobile terminals may not be a feasible solution, migrations and handoffs should impose little load on the mobile host's processor. This is especially important in systems where migration frequency is high and the end user device needs to be small. Wireless access networks may need to operate in harsh conditions where the radio channel is poor. Radio channel black-outs need also to be handled without complicated control processes on the mobile host side.

Passive Connectivity

In a wireless access network the number of users can grow to a point where using fast lookups in per user location data bases is no longer viable. In addition, mobility management requires mobile hosts to send registration information after migrations. The resulting signalling overhead has significant impact on the performance of the wireless access network. To overcome these problems, cellular telephony systems require mobiles to register every migration only when they are engaged in 'active' calls. In contrast, 'idle' mobile hosts send registration messages less frequently and as a result can roam large areas without loading the network and the mobility management system. In this case, the location of idle mobiles is only approximately known to the network. To establish a call to an idle mobile, the mobile must be searched for (i.e., paged) in a limited set of cells. This feature of *passive connectivity* allows the cellular network to accommodate a very large number of users at any instance without overloading the network with large volumes of mobility management signalling information. In order to scale to large user populations, wireless access networks need to provide passive connectivity similar to cellular telephony systems. A number of major differences between voice and data access networks make the adaptation of this concept challenging. In particular, while telephony users are always either in active or in idle state, users of IP networks can not easily be partitioned into such categories. The access network must address this issue without introducing excessive control mechanisms to the system.

4.4 Design Principles

In this section we examine the requirements identified previously and use them to derive a set of design principles for wireless access networks. It must be noted that in accordance with the qualitative definition of our requirements, the derivation of design principles is based on heuristics. The feasibility and validity of the derived principles will be proven by creating an experimental system based on these principles and comparing its properties to our vision and requirements. These steps will be described in Sections 4.5 and 4.6, respectively.

1. We expect the cellular infrastructure to consist of universal building blocks (i.e., *nodes*) that are fully operational individually, but can also operate in a network, without prior

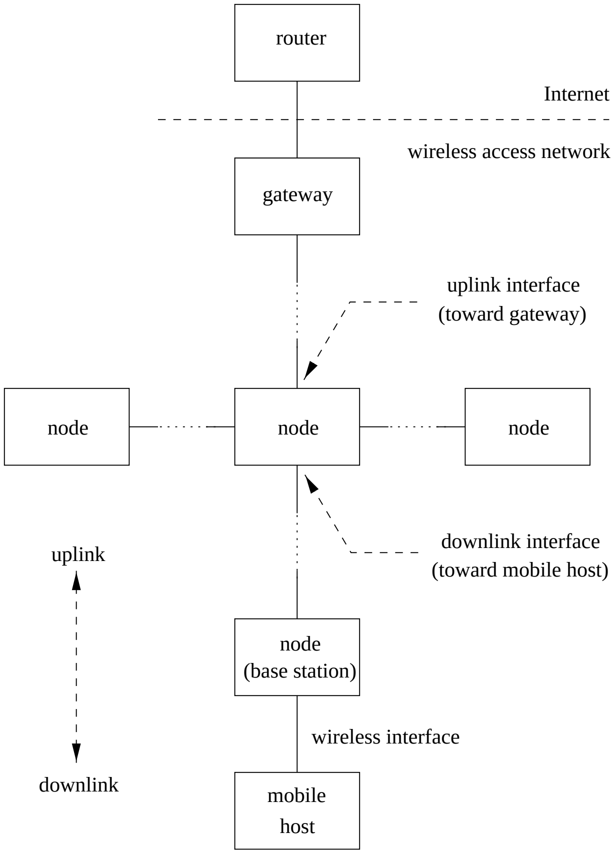

Figure 4.4: Wireless access network reference model

configuration. Since location management data bases must be integrated in the nodes which, on the other hand, must be simple and affordable for a small scale installation, there is no place for a centralized location data base. As centralized mobility management is ruled out, location information must be distributed among the nodes. To avoid creating potentially very large location data bases and data base consistency problems, nodes can, however, not have a full view of the location of all mobile hosts in the service area. We therefore conclude that location management must take a distributed form such that each node of the wireless access network has only a limited local view of the location of mobile hosts.

2. In order to reduce node complexity and to allow for fast routing, it is necessary to depart from existing IP local mobility solutions (e.g., hierarchical foreign agents [17]) which use mobility enabled IP routers as base stations. To eliminate the need for IP routing functionality and to increase routing speed, we propose the integration of location management with routing. In the integrated solution the location information stored in a location data base can directly be the route toward the destination instead of the network address of another node. This allows to route packets in a single step because contacting the location data base does not need to be followed by a lookup in the local routing table.

3. Signalling and registration functions load the mobile host and conflict with the requirement of simple, low cost user terminals especially in environments of high migration frequency. The integration of routing with location management, on the other hand, makes it possible to use regular IP packets to relay location information. Packets originated by the mobile host carry location information by the route they take and make explicit handoff signalling unnecessary as long as the host is actively transmitting data. When the host has no data to transmit then this *implicit signalling* can not be relied on, but explicit migration notifications can still be transmitted in the form of regular IP packets, as *inband signalling messages.*

4. To explicitly clear after a handoff the location information related to a mobile host's old position, either the mobile host can send a release message before performing the handoff or nodes can interact to distribute the information related to migration. While the former solution increases the functionality of the user terminal, the latter assumes that nodes are aware of the network topology and exchange control messages. In order to make wireless networks a plug-and-play solution and nodes self-sufficient, we need to eliminate inter-node messaging. Release messages are unnecessary if nodes store location information as *soft states.* Location information related to a mobile host then remains valid for a system specific time and needs regular refreshing to renew validity.

5. To efficiently route data packets to a mobile host, the host's location must be exactly known by the location management system. Maintaining exact location information about all mobile hosts, however, conflicts the requirement of passive connectivity. Mobile hosts may sometimes remain idle for long periods of time during which their location management messaging must not impose heavy load on the network. In order to provide passive connectivity, it is necessary for wireless access networks to handle location management of idle and active mobile hosts separately.

It is worth pointing out that the derived design principles reflect our intention of combining cellular telephony concepts with the IP paradigm. While the use of simple peer nodes, inband signalling and soft states are characteristic of IP networks, integration of location management with routing and separation of idle from active mobile hosts are more in line with the cellular approach.

4.5 Cellular IP

In the previous sections we established a model and a set of requirements for wireless access networks. Based on the requirements, we derived system design principles. In this section we describe *Cellular IP*, a wireless access network architecture and protocol based on the identified design principles.

4.5.1 System Overview

Cellular IP inherits cellular technology principles for mobility management, passive connectivity and handoff support, but implements these around the IP paradigm. The universal component of a Cellular IP network is the *node* that serves as wireless access point but at the same time routes IP packets and integrates cellular control functionality traditionally found in Mobile Switching Centers (MSC) and Base Station Controllers (BSC). The nodes are built on regular IP forwarding engine, but IP routing is replaced by Cellular IP integrated routing and location management. The Cellular IP network is connected to the Internet via a *gateway* router. Mobile hosts attached to the network use the IP address of the gateway as their Mobile IP care-of address.

In accordance with commercially available wireless LAN solutions [25], Cellular IP assumes that a random access protocol covers the wireless link. Instead of a centralized radio resource management, base stations manage their own radio resources on a per packet basis. Random access may result in potentially lower radio resource efficiency than what centralized management can achieve, but tends to adapt better to IP applications. In addition, the lack of centralized resource management makes handoffs simpler which is a major requirement in systems of high migration frequency.

In Cellular IP, *uplink* packets are routed from the mobile host to the gateway on a hop-by-hop basis. The path taken by these packets is cached in nodes on route. To route *downlink* packets addressed to a mobile host the path used by recent packets coming from the same host is reversed. When the mobile host has no data to transmit then it sends empty IP packets to the gateway to maintain its downlink routing state. Following the principle of passive connectivity mobile hosts that have not received packets for some period allow their downlink routes be cleared. In order to efficiently route packets to these idle hosts, Cellular IP uses a paging mechanism leveraging experience from cellular telephony. In what follows we present a brief overview of the Cellular IP functions followed by a more detailed description of the protocol.

Routing

The Cellular IP gateway periodically broadcasts a beacon packet that is flooded in the network. Nodes record the interface they last received this beacon through and use it to route packets toward the gateway. All packets transmitted by mobile hosts regardless of the destination address are routed to the gateway using these routes.[2]

As uplink packets pass each node on route to the gateway, their route information is recorded as follows. Each base station maintains a *routing cache*. When a data packet originated by a mobile host enters a base station the local routing cache stores the IP address of the source mobile host and the interface over which the packet entered the node. In the scenario illustrated in Figure 4.3 data packets are transmitted by a mobile host with IP address **X** and enter **BS2** through its interface **a**. In the routing cache of **BS2** this is indicated by a mapping **(X,a)**. This mapping remains valid for a system specific time *route-timeout* and its validity is renewed by each data packet that traverses the same interface coming from the same mobile. As long as the

[2] In normal operation these routes rarely change. The periodic beacons are useful, however, to quickly and automatically respond to a change of topology after a network element failure or management action.

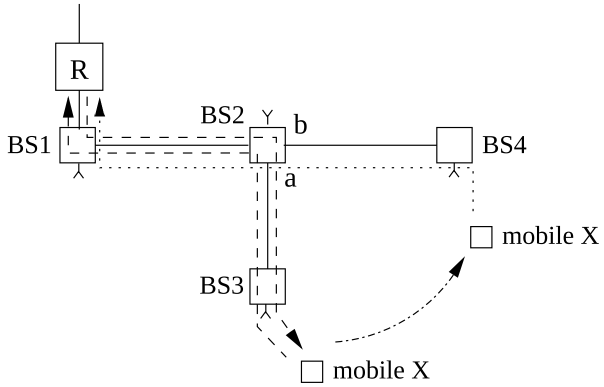

Figure 4.5: Cellular IP handoff scenario

mobile host is regularly sending data packets nodes along the path between the mobile's actual location and the gateway maintain valid entries in their routing cache forming a soft-state route between the mobile and gateway nodes. Packets addressed to the same mobile host are routed on a hop-by-hop basis using the established routing cache.

A mobile host may sometimes wish to maintain its routing cache mappings even though it is not regularly transmitting data packets. A typical example for this is when the host is the receiver of a stream of UDP packets and has no data to transmit. To keep its routing cache mappings valid the mobile host transmits *route-update packets* at regular intervals called *route-update time* that should be smaller than route-timeout. These packets are empty data packets addressed to the gateway. Route-update packets have the same effect on routing cache as normal data packets; however, they do not leave the Cellular IP network.

Handoff

Cellular IP handoff is founded on a simplistic approach by allowing some packet loss in exchange for minimizing handoff messaging instead of trying to guarantee zero loss. Handoff is initiated by mobile hosts. Hosts listen to beacons transmitted by base stations and initiate handoff based on signal strength measurement. To perform a handoff a mobile host has to tune its radio to the new base station and transmit a route-update packet. This will create routing cache mappings on route to the gateway hence configuring the downlink route to the new base station. The mappings associated with the old base station are not cleared at handoff, rather, they timeout as the associated timers expire.

A handoff scenario is illustrated in Figure 4.5. Here, a mobile host with IP address X moves from **BS3** to **BS4** during an active data session. When Layer 2 measurements indicate that radio connection with **BS4** is better than with **BS3** (including some hysteresis), the mobile host tunes its radio to the channel used by **BS4** and transmits a route-update packet. The path of this route-update packet is shown by dotted line in Figure 4.5. This packet first configures a new route cache mapping in **BS4** which has formerly not had mapping for this particular mobile host. Next, it configures a new mapping in **BS2**. This node will now have two mappings for the same mobile host, one associated with its old and one with its new location. During a time determined by the timeout of routing cache mappings, the two mappings will coexist and packets addressed to mobile host **X** will be delivered through both base stations. After this time, the old mapping will be cleared and downlink packets will be forwarded using the mapping **(X, b)**. At the same time, the mapping cached in **BS3** associated with the same mobile host will also be cleared. The mapping in **BS1** remains unchanged during the handoff process.

Paging

Cellular IP defines an *idle mobile host* as one that has not received data packets for a system specific time *active-state-timeout.* Idle mobile hosts let their respective soft-state routing cache mappings time out. These hosts transmit *paging-update packets* at regular intervals defined by *paging-update-time.* The paging-update packet is an empty IP packet addressed to the gateway that is distinguished from a route-update packet by its IP type parameter. The mobile host sends its paging-update packets to the base station that has the best signal quality. Similar to data and route-update packets, paging-update packets are routed on a hop-by-hop basis to the gateway. Nodes may optionally maintain *paging cache.* A paging cache has the same format and operation as a routing cache except for the following differences. First, paging cache mappings have a longer timeout period called *paging-timeout.* Second, paging cache mappings are updated by any packet sent by the mobile hosts including paging-update packets. In contrast, routing cache mappings are updated by mobile originated data and route-update packets only. This results in idle mobile hosts having mappings in paging caches but not in routing caches. In addition, active mobile hosts will have mappings in both types of cache. Packets addressed to a mobile host are normally routed by routing cache mappings. Paging occurs when a packet is addressed to an idle mobile host and nodes find no valid routing cache mapping for the destination. If a node has no paging cache, it will forward the packet to all its interfaces except the one the packet came through. Paging cache is used to avoid broadcast search procedures. Nodes that have paging cache will only forward the paging packet if the destination has a valid paging cache mapping for the mobile hosts and only to the mapped interface(s). Without any paging cache the first packet addressed to an idle mobile is broadcast in the network. While the packet does not experience extra delay it does, however, load the access network. Using paging caches, the network operator can restrict the paging load in exchange for memory and processing cost.

Idle mobile hosts that receive a packet move from idle to active state, start their active-state-timer and immediately transmit a route-update packet. This ensures that routing cache mappings are quickly established potentially limiting any further flooding of messages to the mobile host.

4.5.2 Protocol Details

Protocol Parameters

Cellular IP is designed to efficiently operate in a wide range of environments from small indoor systems to large area installations. These systems may operate in largely different mobility and traffic conditions. A mobile office network will typically offer higher access rates than a metropolitan area wireless ISP, but its users will migrate less frequently. The open parameters in the protocol allow the network operator to adapt the system to local conditions. In Table 4.1 we list the open parameters in Cellular IP. For each parameter, a representative value is also shown. These values, however, are listed for information only and shall largely vary depending on the actual environment.

Base Station Beacon

Cellular IP base stations must periodically transmit beacon signals to allow for mobile hosts to identify an available base station. Information elements carried by the beacon signal are:

Name	Meaning	Representative value
route-update-time	Inter-arrival time of route-update packets	100 ms
route-timeout	Validity of routing cache mappings	300 ms
paging-update-time	Inter-arrival time of paging-update packets	10 sec
paging-timeout	Validity of paging cache mappings	30 sec
active-state-timeout	Time the host remains active without incoming data	10 sec

Table 4.1: Open parameters in Cellular IP

- base station identifier;
- Layer 2 parameters related to the base station;
- the Cellular IP network identifier; and
- the IP address of the gateway.

Addressing and Packet Types

As mobile host addresses have no location significance inside a Cellular IP network, any space of unique host identifiers could be used to identify mobile hosts. In Cellular IP, mobile hosts are addressed by their home IP addresses.

Route-update and paging-update packets are also regular IP packets of which

- the source address is the sending mobile host's IP address;
- the destination address is the gateway's IP address; and
- the IP protocol type field indicates the packet type (route-update or paging-update).

The payload of route-update and paging-update packets may be empty. Optionally, control information may be carried in these packets' payload, encoded in the type-length-value format specified in [D1]. Control information carried in the route-update or paging-update packets can contain user authentication, charging information or other system specific messages.

Node Algorithm

Nodes of a wireless access network need not contain IP routing functionality. By recording through which interface they last received the gateway's beacon (see Section 4.5.1), nodes know the route toward the gateway. This interface is called *uplink interface*. Other interfaces are used to reach the mobile hosts. These are called *downlink interfaces*. By maintaining this elementary routing information nodes effectively create and maintain a logical tree topology over a possibly meshed cellular network infrastructure. An IP packet arriving to the node through the uplink interface is assumed to be addressed to a mobile host and invokes the downlink routing algorithm illustrated in Figure 4.6. The algorithm first checks the local routing cache to search for information about the destination address. If this lookup returns a valid result, the packet is forwarded as dictated by the routing cache. If valid routing cache entries are not found, the packet is routed by the paging cache or is broadcast if a paging cache is not available.

Packets arriving through a downlink interface are assumed to be coming from a mobile host and invoke the uplink routing algorithm shown in Figure 4.7. Before forwarding the packet over the uplink interface, routing and paging cache mappings are refreshed. As described in

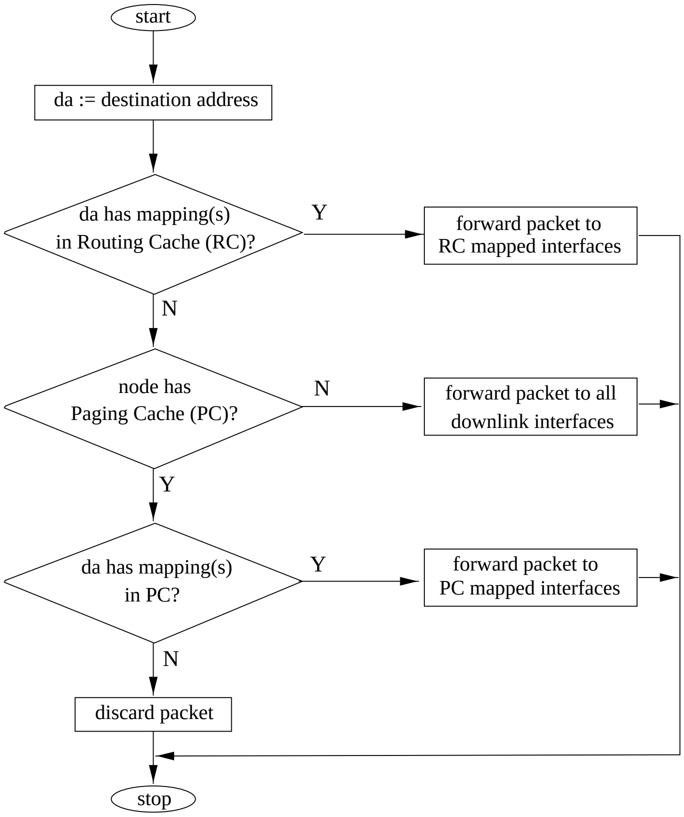
start
da := destination address
da has mapping(s) in Routing Cache (RC)?
Y
forward packet to RC mapped interfaces
N
node has Paging Cache (PC)?
N
forward packet to all downlink interfaces
Y
da has mapping(s) in PC?
Y
forward packet to PC mapped interfaces
N
discard packet
stop

Figure 4.6: Downlink routing algorithm

Section 4.5.1, paging-update packets update paging cache mappings only, while route-update and data packets update both paging and routing cache mappings.

Mobile Host Algorithm

Mobile hosts can be modelled as a simple two-state state machine as illustrated in Figure 4.8. It is important to note that though the names of the two states may suggest that they refer to whether or not the host is transmitting data, they are more related to incoming data. The host should be in "active" state when it expects incoming data and in "idle" state otherwise. In most cases data will be expected when there are also data or acknowledgment packets to transmit which justifies the names of the two states.

A mobile host moves from idle to active state when it receives any IP packet. If it does not receive more packets, it remains in active state for a time duration defined by the system parameter *active-state-timeout*. Any IP packet received in active state restarts the active state timer. When the timer finally elapses, the host returns to idle state.

When the mobile host moves from idle to active state, it transmits a route-update packet. Further route-update packets are transmitted with an inter-arrival time defined by the system parameter route-update-time. Whenever the active mobile host transmits a data packet, however, the route-update packet timer is reset and the next route-update packet is scheduled after a time

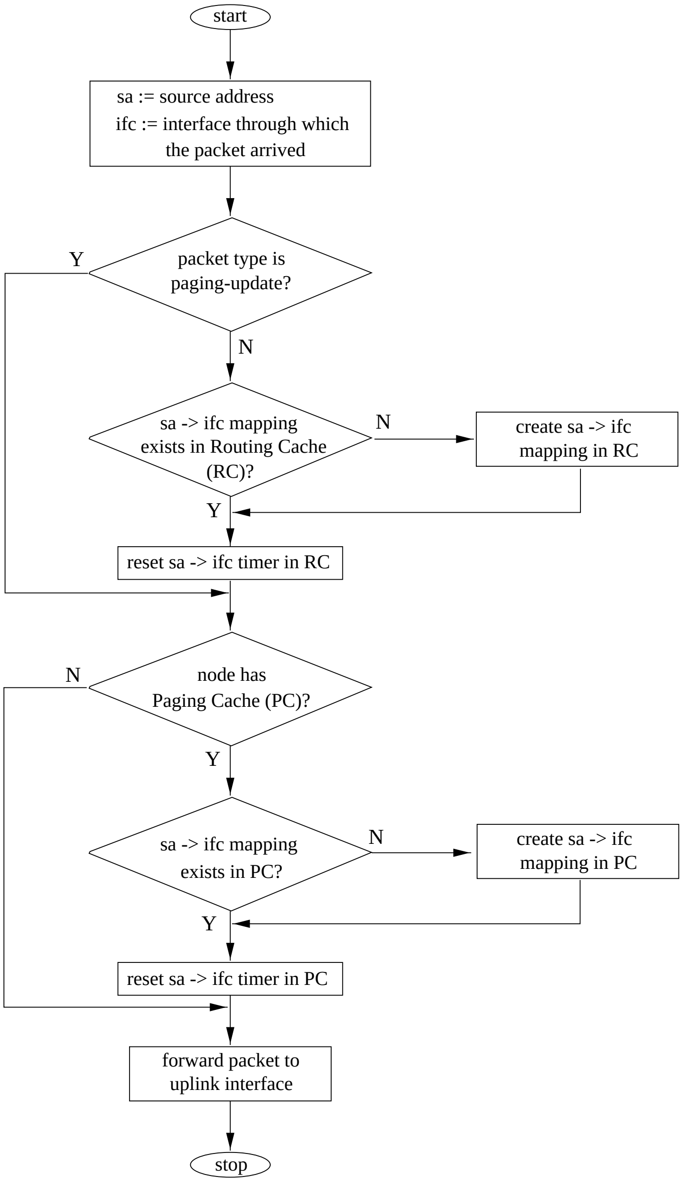
start
sa := source address
ifc := interface through which the packet arrived
packet type is paging-update?
Y
N
sa -> ifc mapping exists in Routing Cache (RC)?
N
create sa -> ifc mapping in RC
Y
reset sa -> ifc timer in RC
node has Paging Cache (PC)?
N
Y
sa -> ifc mapping exists in PC?
N
create sa -> ifc mapping in PC
Y
reset sa -> ifc timer in PC
forward packet to uplink interface
stop

Figure 4.7: Uplink routing algorithm

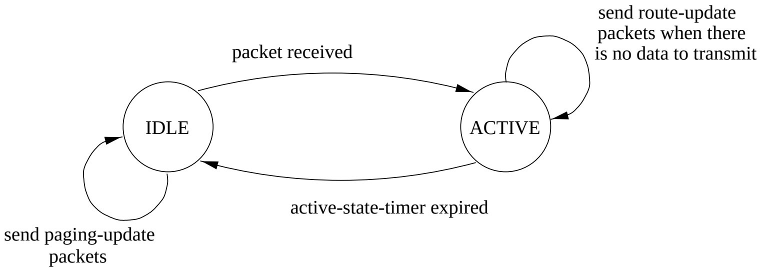

Figure 4.8: Mobile host state machine

equal to route-update time. This ensures that while the mobile host is in active state, packets are transmitted with inter-arrival times no longer than route-update time and that as long as data packets are transmitted at sufficient frequency, route-update packets are not generated.

In idle state the host transmits paging-update packets periodically, at intervals defined by paging-update-time. Similar to route-update packets, the transmission of paging-update packets is suspended when data packets are being transmitted. (We recall that transmitting a packet does not make the mobile move to active state.)

Regardless of the state of the mobile host, it must immediately transmit an IP packet whenever it moved to a new base station. This typically happens at migrations, but is also the case after a wireless channel black-out or when the host enters a new Cellular IP network. The packet transmitted after migration is a route-update packet if the host is in active state and a paging-update packet if it is idle.

4.5.3 Implementation

As a validation of the concept and design principles, the Cellular IP architecture and protocol was implemented in a research project at the Center for Telecommunications Research, Columbia University, New York. In this section we present an overview of the implementation and some measurement results. A more detailed description of the experimental implementation (including validation results) is provided in [32].

Cellular IP is implemented using modular software design on FreeBSD 2.2.6 software platform in user space. The system includes two protocol modules, namely the node and the mobile host modules. Both protocol modules rely on the same system module. The system module filters IP packets from the physical medium to move them to user space and delivers packets processed in the user space to the required network interface as illustrated in Figure 4.9. The system module uses the Berkeley Packet Filter's Packet Capture library (PCAP) [28]. PCAP was designed to capture packets for statistical purposes but can also be used to forward packets on a network interface. In what follows, we describe the node and mobile host protocol modules separately.

Node Module

As described previously, the Cellular IP node serves as wireless access point, router and location manager. In addition, our implementation allows a node to also implement the gateway functionality relying on the kernel's IP routing function. Figure 4.9 illustrates the reference model of the node implementation and Figure 4.10 shows the functional model of the routing module in a Cellular IP node. In these figures, the path of uplink packets is shown as dashed arrows and the path of downlink packets by solid arrows. Dotted lines represent control information. Figure 4.10 shows that all uplink packets update the paging cache, but only a subset of uplink packets update routing cache. This subset is filtered by the packet classifier. Downlink packets are routed by the routing cache if a mapping is available and by the paging cache if a routing cache mapping was not found. Note that in our implementation, the uplink and downlink interfaces are configured by network management instead of gateway beacon messages as specified in the protocol. Along with the routing and (optional) paging cache, the most important functions of the node include:

- a paging update function, which maintains the paging cache by updating it for each uplink packet and by clearing expired mappings;
- a classifier, which parses uplink packets and selects those that should update the routing cache (data and route-update packets);
- a route update function, which maintains the routing cache by updating it for each packet selected by the classifier and clears expired mappings;
- a routing cache lookup function, which parses downlink packets and searches the routing cache for mapping(s) associated with the destination mobile host;
- a paging cache lookup function, which searches the paging cache for mappings if a routing cache mapping was not found; and
- a forwarding engine, which forwards downlink packets to the interface selected by routing cache lookup in the first instance and by paging cache lookup if no route was found.

Note that the protocol module is not aware whether a downlink interface is a wireless interface or a wired connection to another node. Our implementation uses 2 Mbps 2.4 GHz WaveLAN [25] radio devices, but the protocol module can transparently interwork with other radio interfaces. The node may have multiple air interfaces or no air interface at all if it serves as a concentrator node only.

In addition to the functions described above, the node contains beacon generators for each wireless interface.

Mobile Host Module

The mobile host module is implemented as a daemon that, in our experimental testbed runs in user space. This module has very small footprint making it suitable in PDAs, palmtops or cellular IP phones. The standard IP protocol stack is not touched by the Cellular IP daemon and applications are unaware of mobility. The main elements of the daemon are as follows:

- a handoff controller, which keeps statistics of measured beacon strengths and decides and performs handoffs. A handoff mainly consists of setting the radio frequency and changing the IP default route to the new base station's address.
- a protocol state machine, which has two states: active and idle. In idle state, any incoming packet triggers a transition to active state. At the same time, a timer is initiated that is reset by each incoming packet. The expiration of the timer triggers the transition to idle state.
- a control packet generator, which periodically transmits route-update or paging-update packets as required by the state machine. In addition, the packet generator monitors outgoing packets to stop generating control packets when data is being transmitted.

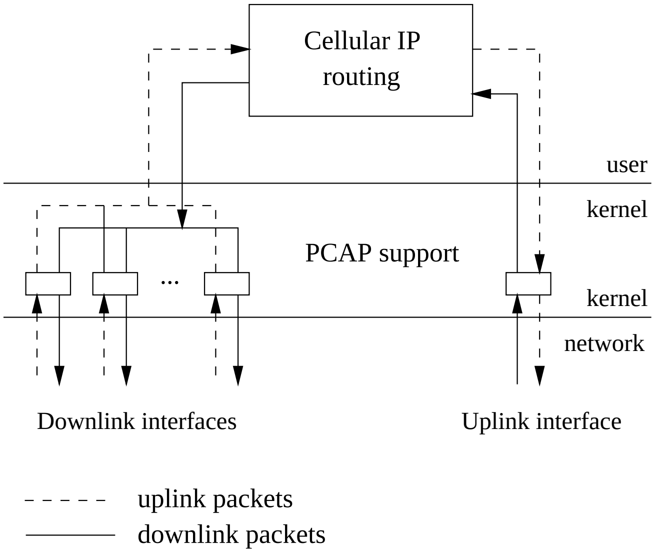

Figure 4.9: Node implementation reference model

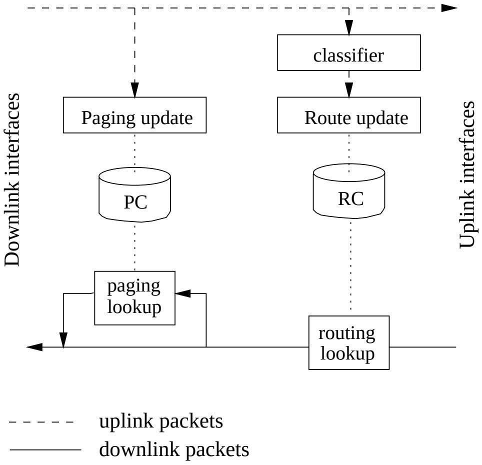

Figure 4.10: Routing module in Cellular IP nodes

4.6 Discussion

The implementation of the experimental Cellular IP system proved the feasibility of the vision outlined in Section 4.1 and provided a testbed for detailed performance studies. Experimental performance analysis results will be described in Chapter 5. Here we summarize the main conclusions we derived from a qualitative analysis of the implementation experiences.

Most important, the experimental system proved the feasibility of creating cellular networks out of simple, off-the-shelf hardware. Our base stations are 300 MHz Pentium PCs which can efficiently handle the mobility management of several thousand mobile users at a time. (The exact number depends on mobility and traffic characteristics. See Chapter 5 for the numerical results.) This high capacity is made possible by the elementary simplicity of Cellular IP mobility management. In particular, it can be attributed to the elimination of explicit signalling and to the concept of passive connectivity. By eliminating explicit signalling and adopting the idea of passive connectivity Cellular IP successfully combines fundamental properties of cellular telephony systems and IP networks. This is a significant improvement compared to existing IP mobility schemes that handle active and idle mobile hosts equally. We note at the same time that replacing explicit signalling with inband implicit signalling also has proved to have a few disadvantages. Since location information in Cellular IP is carried by data packets, each mobile originated data packet triggers a location update function in the nodes it traverses on route to the gateway. Even though the update function consists of very simple steps compared to location management actions of commercial cellular systems, the high frequency at which these actions are performed impacts the throughput of Cellular IP nodes. In particular, we have found that node throughput defined as the maximum number of packets handled per time unit is approximately 10% lower using Cellular IP than by regular IP routing in the same hardware. (A comparison with alternative IP mobility schemes requires implementations of these schemes on the same hardware. This comparison is the subject of future work.) We note, however, that part of this decrease can be attributed to the fact that our implementation resides in user space and is not yet optimized for speed. We conclude that by substituting implicit for explicit cellular signalling, we trade raw throughput for implementation simplicity and for an increase of supported user population. This decision can be justified, we believe, by the fact that cellular throughput will in most installations be limited by the wireless capacity and node throughput is unlikely to become a bottleneck even if decreased by the location management penalty.

In order to provide passive connectivity, Cellular IP relies on a soft state based distinction between active and idle mobile hosts. The improvement gained by passive connectivity depends on the transition rates between these two states and on the mean time spent in active state in a typical user population. This property is a function of both user behaviour and application characteristics. A numerical analysis of this phenomenon is provided in Section 5.5.3 in three different network scenarios.

In the current phase of the experimental implementation the gateway does not transmit beacon messages and uplink interfaces must be manually configured by network management. The self configuring feature of the protocol could therefore not be tested. This property will be analyzed using a future version of the testbed. However, the topology independence and self sufficient nature of Cellular IP nodes is demonstrated by the implementation. In accordance with our requirements, the node algorithm is independent of the size and topology of the network it operates in. The node's view of the network is limited to its own interfaces and our Cellular IP node is not aware whether its downlink interface is a wireless link or a wired connection to a next node. This radically departs from cellular telephony networks that rely on a strict hierarchical network architecture but it also is a significant simplification compared to IP based solutions such as Mobile IP [14], VIP [23] or the Columbia protocol [24] where mobility related functions are implemented as an overlay on top of regular IP routing. In addition, our nodes do not exchange any control messages other than forwarding regular IP packets. These properties of the Cellular IP implementation prove its ability to efficiently operate in small scale environments and are indications about its potential to scale up to large network installations. Numerical results concerning the scalability of Cellular IP networks are presented in Chapter 5.

The integration of location management with routing makes it possible for Cellular IP to identify mobile hosts by their home addresses unlike in other IP based solutions. This property is related to the fact that Cellular IP is itself a Layer 3 routing protocol rather than a mobility extension of the IP routing protocol. We have found that this simplifies implementation because packets can be forwarded without address conversion or tunnelling. Packets transmitted by a mobile host are carried over Cellular IP nodes without modification and are forwarded to the Internet unchanged. Packets arriving from the Internet need to pass Mobile IP detunnelling function in the gateway to support global migration. Once detunnelled, however, these packets are forwarded through the network unchanged and arrive to the destination mobile host in the form they were transmitted by the correspondent host.

Cellular IP uses a distributed soft state location management system. We have found that by adjusting a limited set of soft-state parameters the behaviour of location management can be freely tuned in a wide range without changing the protocol itself. This flexibility is similar to but is greater than GSM's flexibility in determining the size of location areas [16]. This feature will become advantageous in an environment of ubiquitous wireless Internet access if we wish to use the same protocol in various types of access networks. A detailed analysis of Cellular IP's adaptability to mobility and traffic characteristics is provided in Chapter 5.

The experimental implementation also revealed, however, a number of shortcomings of the system in its present form. We have found that the assumption of having only a single gateway may conflict with the requirement of robustness. A similar rigidity also appears in the hierarchical and regional foreign agent solutions [17], [18] (where the higher level foreign agent is a single point of failure in the hierarchy of foreign agents) but not in VIP [23] or the multicasting based scheme [21] . An extension that allows the Cellular IP protocol to operate in networks that have multiple connections to the Internet is therefore required. Although in the experimental implementation we did not address issues of security and user authentication, we have realized that adding these functions may encounter difficulties due to the distributed nature of the protocol. The lack of explicit signalling messages may also become a problem if admission control needs to be applied for handoff attempts or new call requests. These problems must be overcome if Cellular IP is to be used for guaranteed quality, for example voice services. The extension beyond best effort service will also need to improve upon Cellular IP's handoff algorithm. The current solution that takes a simplistic approach and is optimized for best effort service has been found to result in packet losses, particularly if the system is heavily loaded. Numerical results concerning the performance of Cellular IP handoff are presented in Chapter 5.

Chapter 5

Performance Evaluation of Cellular IP Networks

In the previous chapter we have derived a set of design principles for cellular wireless access networks and around these principles we have designed the Cellular IP architecture and protocol. Cellular IP leverages experience from cellular telephony but is firmly based on the IP paradigm. The system consists of simple peer nodes that can be interconnected in an arbitrary topology to automatically form a cellular access network. In accordance with IP principles, Cellular IP takes a simplistic approach to location management, routing and handoffs. These properties make Cellular IP an ideal candidate to build simple, cheap cellular networks for the provision of ubiquitous wireless Internet access. In this chapter we provide a performance evaluation of Cellular IP networks based on a combination of analytical, simulation based and experimental techniques.

5.1 Problem Statement

The objective of this chapter is to evaluate the performance of the Cellular IP architecture and protocol. In the design of Cellular IP we envisioned an environment where wireless mobile Internet access is the norm rather than, as it is today, an exception. We assumed that wireless access networks provide mobility and handoff support in "local" areas of various scale and character and that a global mobility protocol supports roaming between access networks. This vision motivated the design of Cellular IP, a wireless access technology that provides a cheap and flexible solution for wireless IP access networks ranging from small indoor systems to large area networks. In this environment a property of outmost interest is the solution's ability to adapt to a wide range of mobility and traffic conditions. Our attention in this chapter will be focused on this question. Based on a combination of analytical, simulation and experimental studies we will analyze Cellular IP from four key aspects related to performance and adaptability.

A fundamental design objective of Cellular IP was implementational and functional simplicity. To reduce complexity, we omitted explicit location registrations and replaced them by implicit inband signalling. As a result, nodes of the access network need not be aware of the network topology or of the mobility of hosts in the service area. This design choice deliberately trades off performance for simplicity, potentially letting packets to be lost at handoff rather than explicitly buffering and redirecting packets as the mobile host moves. In Section 5.5.1 we analyze handoff performance in Cellular IP and quantify the performance penalty associated with Cellular IP handoff simplicity.

In Section 5.5.2 we look at the 'cost' of mobility management by investigating the trade-offs involved in setting the following Cellular IP system parameters:

- route-update time;
- paging-update time;
- route-timeout;
- paging-timeout; and
- active-state-timeout.

Determining the Cellular IP mobility management cost is important because different cellular system installations may operate in largely different mobility conditions. A mobile office network will typically offer higher access rates to mobile users than would be found in a metropolitan area wireless ISP, but its users will migrate less frequently. The network operator is free to set the Cellular IP system parameters to adapt the protocol performance to local conditions with the goal on minimizing mobility management cost.

Being a "universal building block" of Cellular IP networks, the node is central in the system's behaviour. Nodes serve as base stations, packet forwarding points and location management entities. The design of Cellular IP was in part motivated by the need to use commodity hardware to support cellular network elements such as the node. In Section 5.5.3 we investigate the performance limits of our node implementation using off the shelf hardware. These results are used as a basis for investigating the ability of Cellular IP to scale to support different cellular systems installations. Three mobile networking scenarios are chosen that provide insight into the ability of Cellular IP to be customised to meet a wide range of mobility, application and network conditions.

By taking a simplistic approach to mobility and handoffs, Cellular IP follows the IP concept of providing best effort service over a low cost, robust infrastructure. Most IP based applications were designed to operate efficiently in a harsh environment and can handle packet loss or largely varying delay. Some applications, however, require predictable conditions to operate efficiently. In the Internet a number of recent initiatives address this issue and attempt to increment IP with service quality guarantees [27], [80]. In Section 5.5.4 we investigate Cellular IP from this perspective and outline alternative handoff algorithms that improve service performance in exchange for added functionality in nodes and mobile hosts.

5.2 Related Work

Despite the large amount of literature about cellular networks, little is available regarding their performance evaluation. An extensive study of signalling procedures in third generation cellular mobile systems is provided in [61] and IMT-2000 networking aspects are partly covered by [J1], [67] and [34]. Internet access through narrow band cellular systems has been addressed in [63] and through GPRS in [83]. The performance of a CDMA wireless hop in an integrated voice-data environment is studied in [62]. None of the above studies are applicable, however, to Cellular IP because of the fundamental architectural differences between traditional cellular systems and Cellular IP.

A better basis for comparison of Cellular IP peformance is provided by the now emerging IP mobility schemes. Most of these proposals, however, are in the phase of conceptual design and performance results are not yet available. In what follows, we briefly outline performance evaluation methods and results concerning two alternative local mobility schemes.

An experimental evaluation of the local IP mobility protocol proposed by Caceres in [19] is provided in [69]. Similar to our Cellular IP implementation, the measurements are performed in a Unix environment using WaveLAN radio devices. The analysis is focused on handoff performance using a two-cell testbed. The time it takes for the packet flow to be restored after the mobile host has connected to the new base station is measured to be around 10 ms. Packet loss at handoff is measured using a packet audio example. Measurements indicate that the number of lost packets depends on the period of beacons transmitted by the base station. For a typical beacon period of 100 ms an average of 3.7 packets get lost at each handoff. Unlike Cellular IP, the Caceres proposal uses buffering to reduce packet loss at handoff. It is shown in [69] that buffering can result in duplicate packets seen by the application. The analysis of the trade-off between packet loss and duplicates shows that the optimal number of packets buffered at handoff is between 2 and 6. The buffer size and base station beacon period can be varied to adapt the system to latency and reliability requirements. Finally, TCP performance as a function of the base station beacon period is measured using `ttcp`. Throughput is found to decrease rapidly if the beacon period is less than 50 ms but to be almost independent of the beacon period above 50 ms.

The handoff performance of the multicasting based mobility solution [21] is analyzed in [70] in an experimental environment also using WaveLAN and a two-cell setup. Unlike Cellular IP, the multicasting approach is designed to provide very smooth handoffs with virtually zero packet loss. This is achieved by assigning multicast addresses to mobile users and delivering packets to both the old and new base stations during handoff. In addition, packets may be buffered in the base stations to eliminate packet loss. Measurements reported in [70] show that multicasting combined with buffering entirely eliminates packet loss at handoff while multicasting without buffering still allows 2 to 4 packets to be lost, depending on the distance between the old and new base stations. Handoffs without multicasting result in 3 to 5 packet losses. In addition, a series of measurements indicate that the throughput of a TCP session is hardly affected by handoffs using the multicasting approach. The throughput achieved at a handoff frequency of 1/sec shows less than 1% degradation compared to that achievable by a static host.

The performance of TCP in a wireless environment has recently been widely studied in the literature. TCP is optimized to operate in a wired environment where bit errors are rare and packet loss is typically due to congestion. In wireless mobile systems, however, packets can get lost due to bit errors over the wireless link or at handoff. TCP flow control interprets these packet losses as signs of congestion and reduces transmission rate. To overcome the problem, in [77] the end-to-end TCP connection is split into two separate connections, one over the wireless "last hop" and another over the wired section. In [72] a fast retransmission scheme is proposed and evaluated in an experimental testbed. A wide array of other proposals are presented in [73], [74], [75], [76] and in references therein. An overview and comparison of existing solutions is presented in [71]. Most proposals, however, rely on maintaining per flow information (e.g., packet buffer) in the base stations. This requirement conflicts with Cellular IP design decision of keeping nodes unaware of data sessions. In addition, the flow specific information needs to be relocated at handoff which further increases node complexity. The applicability of the above mentioned solutions in a Cellular IP network is therefore limited. In this chapter we focus on the evaluation of handoff performance in the Cellular IP base protocol and exclude errors over the wireless interface. Solutions to augment Cellular IP with protection against such errors are the subject of future research.

5.3 Model

In the present analysis, we limit our attention to the performance of a Cellular IP network in isolation. Interworking with global mobility protocols, particularly with Mobile IP is the subject of future phases of this work. We assume a Cellular IP system where cells partially overlap allowing the mobile host to immediately connect to a new base station after leaving the old one. In systems where this is not the case, the time it takes for the mobile host to move from the old cell to the new one adds to the handoff delay calculated and measured in this chapter.

In order to focus on the networking aspects of Cellular IP, we ignore possible errors over the wireless link. Radio errors are not modelled in the simulator and are not generated in the testbed. (Due to the proximity of the base stations to mobile hosts in the testbed, transmission errors are rare.) In accordance with real systems, the wireless link represents, however, a throughput bottleneck in our experiments. This phenomenon is particularly interesting when the performance of TCP during handoffs is studied.

The analysis is focussed but not limited to best effort environments. Our examples are taken from basic IP applications, but packet loss is measured and the impact of handoff delays on performance is also analyzed. Though the current version of the protocol is optimized for best effort services, we believe that this analysis can serve as a springboard for augmenting Cellular IP with service quality provisioning techniques.

5.4 Methodology

The analysis of complex systems calls for a combination of analytical, simulation and experimental methods. Since an analytical approach gives the deepest insight into a system's behaviour, in this chapter we use analytical tools whenever a meaningful and still tractable model can be established. Analytical methods prove to be particularly powerful in the analysis of mobility management cost (Sections 5.5.2). Issues beyond the limits of analytical methods are studied using a combination of measurement and simulation techniques. We use measurements in most cases where the Cellular IP prototype system allows for a meaningful experiment. The advantage of measurement techniques compared to both analytical and simulation methods is the lack of a modeling phase that inevitably results in the loss of some details. However, due to the limited size of the currently available Cellular IP testbed, some aspects of protocol performance can not be measured. In addition, measurements are time consuming and inefficient in cases where major system configuration parameters need to be varied. In these cases we rely on simulations. In the following two sections we describe the simulation environment and the experimental setting used for the analysis.

5.4.1 Simulation Environment

The Cellular IP simulator is an extension of the `ns-2` network simulator [79]. `ns-2` is a public domain simulator written in C++ and Tcl [85] programming languages and is widely used to analyze IP networks, in particular the TCP protocol. The Cellular IP extension is written in Tcl language. The simulator supports Cellular IP networks of arbitrary topology but the model contains a few limitations compared to real systems. The most important limitations are the following:

- An "ideal wireless interface" is used. Packets transmitted over the wireless interface encounter no delay, bit error or loss. Congestion over the air interface can not be modelled.
- The beacon messages transmitted by a Cellular IP gateway are not modelled. The network is configured when the simulation session is initiated and the topology remains constant during simulation.
- Wireless cells are assumed to overlap and mobile hosts move from one cell to another in zero time. (We point out that this does not limit the simulator's ability of studying packet loss

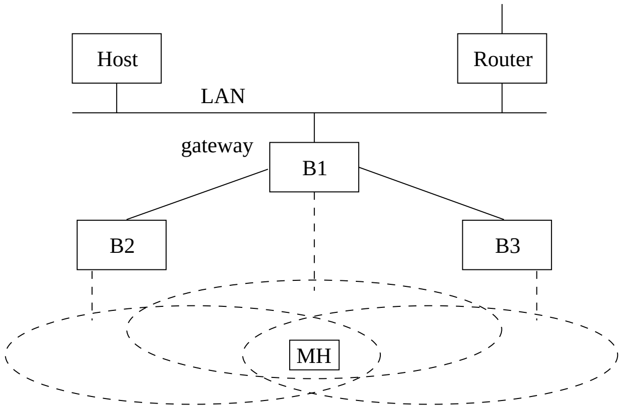

Figure 5.1: Cellular IP testbed

at handoff because that is caused by misrouted packets rather than by the mobile host's being hidden during handoff.)

5.4.2 Experimental Setting

All of the experiments described in this chapter were conducted using the configuration illustrated in Figure 5.1. This Cellular IP network consists of three nodes (denoted by **B1**, **B2**, and **B3**), all multihomed 300 MHz Pentium PCs. All three nodes implement the Cellular IP protocol as defined in [D1] and described in Section 4.5. The protocol software is written in C programming language and resides in user space.

One of the nodes also serves as gateway router. As illustrated in Figure 5.2., this node also implements the Cellular IP node software. In addition, it relies on the operating system's built in IP routing function to interface a regular IP network. The gateway is connected to a 100 Mbps Ethernet Local Area Network (LAN).

The gateway node is connected to two other nodes through 100 Mbps full duplex links. Nodes are equipped with WaveLAN 2.4 GHz radio interfaces. These devices implement the IEEE 802.11 protocol that is optimized for wireless packet data services, particularly wireless LANs [25]. WaveLAN radio interfaces appear to the operating system as regular network interfaces (e.g., Ethernet), which makes it attractive for LANs and for Cellular IP.

The mobile host (**MH**) is a 300 MHz Pentium PC notebook. The Cellular IP mobile host functions, including the state machine (see Section 4.5) are implemented as a daemon running in user space. The mobile host is also equipped with a WaveLAN 2.4 GHz radio interface. Unlike the nodes that operate at statically assigned frequencies, the mobile host can dynamically select one out of the eight frequencies supported by this product. At any time, the mobile host is tuned to be able to communicate with exactly one of the three nodes.

The WaveLAN radio devices operate over distances up to 50 meters. The testbed's nodes are close to one another and throughout the experiments the mobile host is in the overlapping region of the three cells. This gives us full control over handoffs. We extended the mobile host's implementation with a utility that can periodically trigger handoffs regardless of the signal strength. A handoff initiated by this utility is identical to a handoff triggered by signal strength measurements.

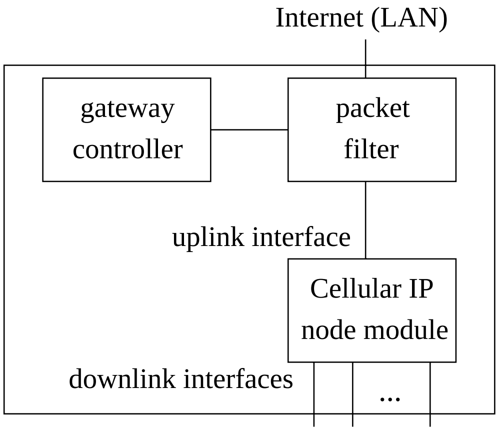

Figure 5.2: Gateway node architecture in the Cellular IP testbed

5.5 Analysis

5.5.1 Handoff Performance

Handoffs are central to the performance of a cellular access network especially in systems with small wireless cells and fast moving hosts. Cellular IP is designed to operate efficiently even at very high handoff frequencies. In accordance with the design goals, a lightweight handoff algorithm is used that avoids explicit signalling messages (used for example in cellular telephony systems and in Mobile IP) and buffering or forwarding of packets (proposed in [21] and [24]). By this design decision, however, performance is traded for simplicity. Explicit registrations, and packet buffering or forwarding reduce or eliminate the disturbance handoff means to active data sessions. In Cellular IP, packets can be lost at handoffs and these losses must be dealt with at higher protocol layers (e.g., TCP). In this section we analyze the performance of Cellular IP handoff to determine the performance penalty we pay for a simpler implementation and operation.

Handoff Delay

The disturbance that handoff means to ongoing sessions is commonly characterized by the *handoff delay*. Handoff delay is usually defined as the time it takes to resume normal traffic flow after the host performs a handoff. Though this does not fully determine the performance seen by the applications, it is a good indication of handoff quality. In [19] handoff delay is further decomposed to *rendezvous time* and *protocol time*. Rendezvous time refers to the time it takes for a mobile host to attach to a new base station after it leaves the old one. This time is related to wireless link characteristics, particularly to the inter-arrival time of beacons transmitted by base stations. Protocol time refers to the time it takes for the traffic flow to be restored once the mobile host has received the beacon from the new base station. In the present analysis we assume that the rendezvous time is small and handoff performance is determined by the protocol time. Instead of adopting the notations proposed in [19], we therefore define handoff delay as the time it takes for a mobile host to receive the first packet through the new base station after it moved from the old to the new base station — which we assume to take zero time.

In Cellular IP, handoff delay and packet loss are consequences of the time it takes for the distributed routing state to follow host mobility. As described in Section 4.5, immediately after handoff, mobile hosts transmit a route-update packet to reduce this time to a minimum. The route-update packet travels from the new base station to the gateway and configures a new downlink route to the mobile host. The old and new downlink routes both originate in the gateway but while the former routes packets to the old base station, the latter leads to the base station the host has just moved to. A handoff scenario with the old and new routes is illustrated in Figure 4.5. The node where the old and new routes join (**BS2** in Figure 4.5) is referred to as the *cross-over node*. The new downlink route becomes operational when the first route-update packet transmitted through the new base station reaches the cross-over node. The time period while the mobile host is not receiving packets after handoff is therefore the time it takes for the route-update packet to reach the cross-over node plus the time it takes for the first downlink packet to travel from the cross-over node to the base station. Handoff delay, as defined previously, is therefore equal to the round-trip time between the new base station and the cross-over node.

Packet Loss at Handoff

Application level quality, however, is more related to the number of packets lost at handoff than to the handoff delay. To determine handoff packet loss, let us assume that a periodic stream of packets is being transmitted from the Internet to a mobile host. Before handoff, the packets are routed along the old route. In the following calculation, we will assume that the cross-over node knows in advance which of the stream of packets will be the last one to reach the mobile host at its old location. Let us assume that the cross-over node marks this packet. Upon receiving the marked packet, the mobile host performs a handoff and immediately transmits a route-update packet through the new base station. Downlink packets routed by the cross-over node after the marked packet but before the arrival of the route-update packet are routed to the old base station and get lost. This time interval is equal to the sum of the time it takes for the marked packet to propagate from the cross-over node to the mobile terminal and the time it takes for the route-update packet to reach the cross-over node. The loss of packets at handoff is therefore related to the "handoff loop time" defined as the transmission time from the cross-over node to the mobile host's old location plus the transmission time from the mobile host's new location to the cross-over node. Specifically, the number of lost packets at handoff n_{loss} is equal to the number of packets arriving to the cross-over node during the handoff loop time T_L, that is

$$n_{loss} = wT_L \tag{5.1}$$

where w is the rate of downlink packets. Since the average handoff loop time is equal to the average handoff delay, the expected number of packets lost at handoff can equally be calculated using the handoff delay (which is easier to measure) and in what follows we do not differentiate between these two values.

Measured handoff packet loss values are plotted in Figure 5.3. During these measurements the mobile host received 100 byte UDP packets at rates of 25 and 50 packets per second (pps) while performing handoffs every 5 seconds. Each point on the graph was obtained by averaging loss measurements over 50 consecutive handoffs. To vary the round-trip time between the mobile host and the cross-over node, we emulated an increasing load which results in increasing buffering of downlink packets. Under these experimental conditions hard handoff results in at least 1 packet loss for small mobile to gateway round-trip delays and up to 4 packet losses for delays of 80 ms. This result is comparable to handoff packet loss results reported about the Caceres protocol (3.7 packets lost per handoff) and about the multicasting approach (2 to 4 lost packets if buffering is not used) as we have cited in Section 5.2. This comparison does not reveal, however, the fundamental differences between Cellular IP handoff and the handoff schemes proposed in the cited approaches. In Cellular IP handoff does not differ from normal operation. Neither

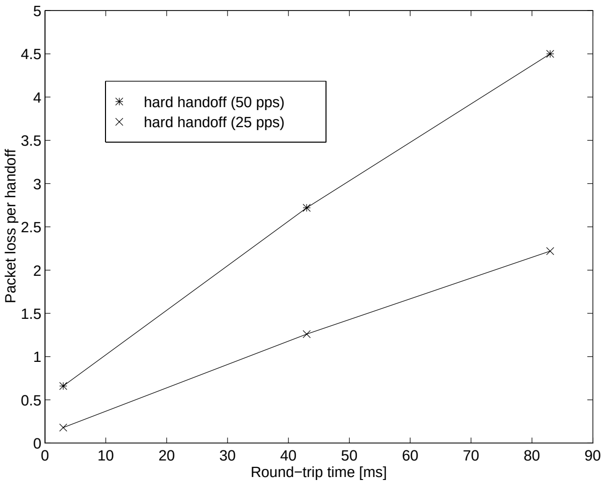

Figure 5.3: Downlink packet loss at handoff (measurement)

mobile hosts nor base stations have special states associated with handoff. In exchange for this simplicity, however, handoff performance is dependent upon the traffic conditions. In a highly loaded network the handoff delay will be higher and more packets will get lost.

Real time Internet applications, for example voice over IP, are sensitive to packet delay and can not retransmit lost packets. For these applications, the number of lost packets fully characterizes handoff performance. Other applications, however, use end-to-end flow control to respond to network and traffic conditions and retransmit packets and/or reduce transmission rate if errors occur. In what follows, we will focus on TCP performance in the presence of handoffs. TCP is selected because it represents the most typical traffic type over today's Internet which carries World Wide Web, file transfer, remote login and other applications. Investigating TCP performance is also important because its flow control has been shown to operate sub-optimally in a wireless environment (see Section 5.2 on related research).

TCP Behaviour at Handoff

We will first use simulation to look at the behaviour of a TCP session at handoff. The simulated configuration is identical to the experimental testbed shown in Figure 5.1. In the first example TCP is used to download data to a mobile host. The TCP packet size is 1000 bytes and a mobile user has up to 5 Mbps downlink bandwidth, that is, the downlink packet rate w is 625/sec. Packet transmission time between nodes in the simulated configuration is 2 ms, resulting in a handoff delay of 4 ms.

Figure 5.4 shows the sequence numbers of downlink data packets and uplink acknowledgments seen in the gateway at a handoff while TCP Tahoe flow control is used. (Configuration file: `handoff_loss_tcp_trace`.) Handoff is performed at 4 seconds simulated time. In accordance with Equation 5.1 three consecutive packets get lost as indicated by the three consecutive missing acknowledgments. After the handoff delay packets continue to arrive at the mobile host. These packets, however, are out of sequence and cause the receiver to generate duplicate acknowledgments as indicated by the horizontal line of acknowledgment sequence numbers. The duplicate acknowledgments inform the TCP transmitter about the losses and cause it to retransmit the

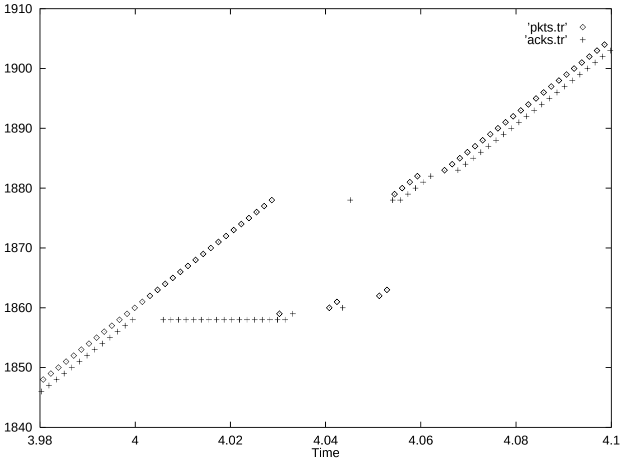

Figure 5.4: TCP sequence numbers at handoff (downlink case, simulation)

lost packets. The first retransmitted packet arrives approximately 20 ms after the handoff. Using Tahoe flow control, the transmitter remains silent until this packet is acknowledged and increases its transmission window size as further acknowledgments arrive. (A detailed description of TCP flow control is presented in [78].) Full speed is not regained until approximately 4.07 sec simulated time.

We conclude that a Cellular IP handoff is interpreted by a transmitter in the wired IP network as congestion and causes it to reduce transmission rate. Using Tahoe flow control the handoff triggers a slow start which increases the performance impact of handoff packet loss. In the studied circumstances, normal operation is resumed approximately 70 ms after handoff. We will show the impact of this disturbance on TCP throughput in a series of experiments later in this section.

In the next simulation session TCP is used to carry data *from* the mobile host. In this case handoff packet loss affects acknowledgments instead of data packets. Figure 5.5 shows simulation results in a configuration identical to the previous case. Before handoff the TCP sender (in the mobile host) uses its maximum window size of 20 which is reflected in the difference between data packet and acknowledgment sequence numbers. At 4 sec simulated time the mobile host performs a handoff and stops receiving acknowledgments for a period of approximately 4 ms, that is the handoff delay. During the handoff delay the sender does not transmit any packets since its window size is used up and it needs incoming acknowledgments to advance its transmission window.

In the simulation session shown in Figure 5.5 handoff occurs when the TCP session is in a stabilized phase and acknowledgments keep arriving to the mobile host in a continuous flow. After the handoff delay, these acknowledgments are routed to the mobile host's new location. Due to the cumulative nature of TCP acknowledgments, the first acknowledgment arriving to the mobile host after handoff informs the sender that all its transmitted packets have arrived to the receiver (up to the sequence number shown in the acknowledgment). This causes it to advance its transmission window and continue transmitting at the maximum available data rate. In the simulation example this rate is slightly higher than the rate dictated by TCP flow control (which is the long term average capacity) so the curve of data packet sequence numbers is somewhat steeper after handoff delay than outside the handoff area. Normal operation is resumed quickly and handoff represents little disturbance to the data session.

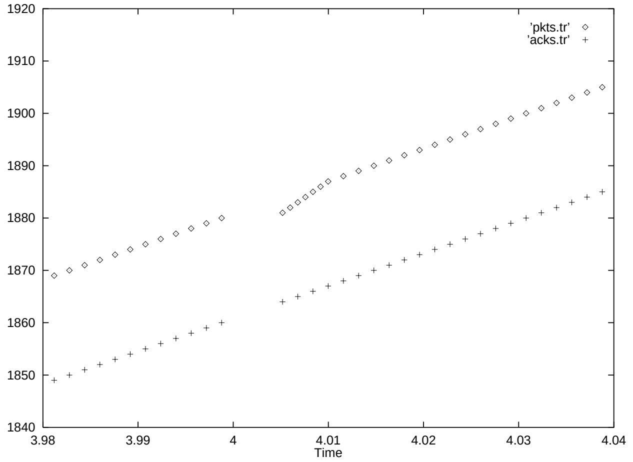

Figure 5.5: TCP sequence numbers at handoff (uplink case, simulation)

The behaviour is different, however, if handoff occurs when the TCP session is in its initial, slow start phase and acknowledgments are not regularly arriving to the mobile host. In this case the new downlink route is established after the handoff delay but no acknowledgments arrive to the sender. If at this point the sender has used all its transmission window and is waiting for acknowledgments then TCP can suffer a delay equal to the sender's retransmission timer. Mechanisms to avoid this problem are for further study.

TCP Throughput

Next, we study the impact of handoff performance on TCP Reno throughput in the experimental testbed. The mobile host performs handoffs between **B2** and **B3** at fixed time intervals. (See the experimental setup in Figure 5.1.) We measure TCP throughput using `ttcp` by downloading 16 MBytes of data from the correspondent host to the mobile host.

The results are plotted in Figure 5.6 where each data point is an average of 6 independent measurements. The throughput measured at zero handoff frequency (i.e., no handoffs) is lower that the 1.6 Mbps we could achieve using standard IP routing in the same configuration. This difference between IP and Cellular IP forwarding is attributed to the fact that IP is implemented in the kernel and Cellular IP in user space. In addition, Cellular IP uses PCAP to forward packets which is optimised for monitoring rather than IP forwarding. We observe that the performance of TCP degrades as the handoff frequency increases due to packet loss. The shape of the throughput curve changes around 5 handoffs per minute. This is attributed to the fact that as the handoff rate increases TCP has less time to recover from losses and at this point it starts operating continuously below its optimal point. Further increasing the handoff frequency results in a significant drop in performance approaching 550 kbps as the handoff rate moves toward one per second.

5.5.2 Mobility Management Cost

In what follows we formulate the cost of mobility management for routing and paging and compare our analytical and observed results.

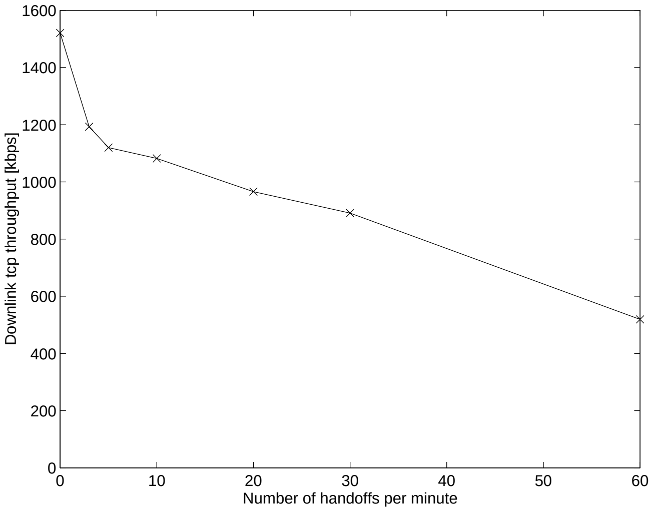

Figure 5.6: Throughput of TCP download (measurement)

Route Maintenance Overhead

The network operator will typically set the route-timeout to be a small multiple of the route-update time. This ensures that the mobile host's routing cache mappings remain valid even if a few route-update packets get lost. Let T_{ru} denote the route-update time and αT_{ru} the route-timeout where α is a small integer. To choose an optimal value for T_{ru}, the following trade-off should be observed. After an active host performs a handoff, its old routing cache mappings remain valid for a duration dictated by route-timeout. During this time, packets addressed to this host continue to be delivered to the old base station increasing the network load and reducing network performance. A small value of T_{ru} should be used to minimize this condition. On the other hand, an active host that has no data to send must transmit route-update packets at a rate of $1/T_{ru}$. This load increases with decreasing T_{ru}. Let the cost of carrying a packet to or from the mobile host be defined as the size of the packet in bits. This model neglects eventual differences in uplink and downlink cost due to different traffic conditions but is sufficient to characterize the T_{ru} trade-off. Consider a mobile host that is receiving data at a constant rate of r bps (including headers) and let p denote the fraction of the time when it is not sending packets and is forced to transmit route-update packets instead. (We note that in some typical IP applications downlink traffic is considerably higher than uplink traffic. This, however, does not necessarily cause p to be high if at least acknowledgments are transmitted in uplink.) The cost of transmitting route-update packets during time T is $R_{ru}pT/T_{ru}$ where R_{ru} is the size of a route-update packet in bits. During this time the mobile performs T/T_H handoffs where T_H (dwelling time) is the mean time spent in a cell. After each handoff, the old route remains active for at most αT_{ru}, the exact value depending on when it was last updated before handoff. Hence the mean cost of sending packets along the old route after handoff is $rT_{ru}(\alpha - 1/2)$ and the total cost of misrouted packets during time T is $rTT_{ru}(\alpha - 1/2)/T_H$. The optimal route-update time $\hat{T}_{ru}$ is the one that minimizes the sum of these costs and is calculated as

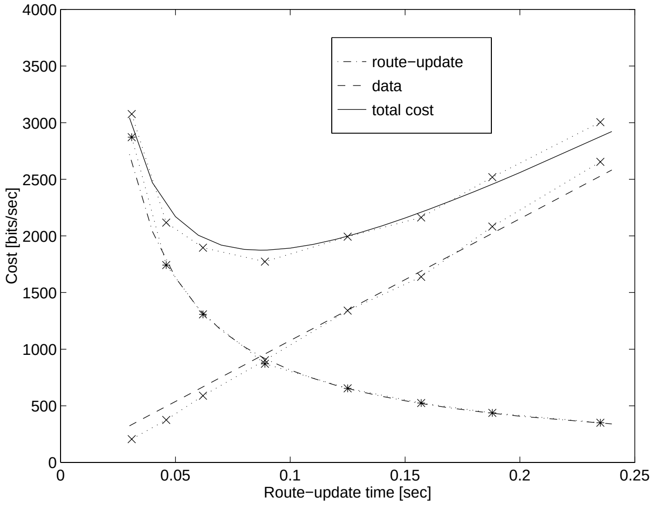

Figure 5.7: Location management cost vs. T_{ru} (dotted lines: measurement)

$$\hat{T}_{ru} = \sqrt{\frac{pR_{ru}T_H}{r(\alpha - 1/2)}}$$

This theoretical result is compared with measurements in Figure 5.7. The mobile host performs handoffs every 30 seconds while it is receiving data at a rate of 128 kbps. The size of route-update packets is 102 bytes, α is 3 and p is 0.1. The theoretical and measured cost associated with route-update packets and misrouted data match well showing that the optimal route-update time in the described scenario is 87 ms. We also plotted the sum of these costs which can be interpreted as the total cost associated with the mobility of an active host and is calculated as

$$C_a = \frac{pR_{ru}}{\hat{T}_{ru}} + \frac{r\hat{T}_{ru}(\alpha - 1/2)}{T_H} = \sqrt{\frac{4pR_{ru}r(\alpha - 1/2)}{T_H}}$$

This cost is not proportional to the migration frequency but to its square root. In keeping with the original design goals, this shows Cellular IP's efficiency in supporting highly mobile hosts. Note that the mobility cost increases with increasing user data rate. This property applies to most mobility schemes (e.g., when data must be forwarded from one base station to another after handoff) but is more apparent in Cellular IP. This is related to the soft-state nature of Cellular IP. Since there is no explicit signaling during handoff, which makes handoff transparent to the base stations, the base station is unaware that mobile hosts move into or out of its cell. Transmitting data to mobile hosts that have left the cell adds to the cost of mobility.

Paging Overhead

The paging-update time T_{pu} is subject to a trade-off similar as T_{ru}. A selected value that is too small will result in very frequent paging-update packets being sent by idle mobile hosts. On the other hand, considering that the paging-timeout is a small multiple of the paging-update time, increasing T_{pu} will result in an increase in the number of cells that an idle mobile host is paged in.

Paging is initiated when a new data session starts by a downlink packet, for instance a TCP connection is initiated *to* the mobile host. Let λ_P denote the arrival rate of such sessions and R_P the mean amount of traffic (bits) sent in paging packets. The paging packets are delivered to all the cells to which the mobile has valid paging cache mappings. Let us first assume that all base stations have paging cache and that the probability of immediately revisiting a cell is negligible. Paging occurs in the 'primary' cell that the target mobile host resides in plus any other 'secondary' cells where the mobile host has valid paging cache mappings. Secondary cells represent cells that the mobile host has recently visited and that have valid paging cache for the target mobile host. Paging secondary cells is a waste of transmission resources and reflects the cost of our paging scheme. The mean number of secondary cells paged is $(\beta - 1/2)T_{pu}/T_H$, where β is the ratio between the paging-timeout and the paging-update time. The optimal paging-update value $\hat{T}_{pu}$ is the one that minimizes the sum of paging-update traffic and wasted paging traffic and is obtained as

$$\hat{T}_{pu} = \sqrt{\frac{R_{pu}T_H}{\lambda_P R_P(\beta - 1/2)}}$$

where R_{pu} is the size of paging-update packets in bits. Using this optimal paging-update time, the total cost C_i associated with the mobility of an idle host is

$$C_i = \sqrt{\frac{4R_{pu}\lambda_P R_P(\beta - 1/2)}{T_H}}$$

These results take a similar form to those obtained for route-update time. However, the downlink data rate r that is an important parameter in the route-update time trade-off, is now replaced by $\lambda_P R_P$ which is the rate at which data arrives to the mobile host in paging packets. This rate depends largely on the application but will be in most cases orders of magnitude lower than r which justifies selecting a higher paging-update time than route-update time. This also accounts for the fact that the cost C_i associated with the mobility of idle hosts is significantly lower than the mobility cost of active users which is the basis of passive connectivity.

In Figure 5.8, we show paging traffic rate measured in the testbed using a set of typical Internet applications. It must be noted that these values depend heavily on the user behaviour and this data is presented for illustrative purposes only. Our measurements involved 5 minute sessions where T_{ru} was 100 ms and $\beta = 3$. In the `telnet` and WWW sessions the mobile host was the client. We collected data from two sets of measurements for two different active-state-timeout values. As discussed previously, this parameter determines the time until a mobile host maintains its routing cache mappings after receiving a packet. In other words, the active-state-timeout reflects the expectation that one downlink packet may with high probability be soon followed by another and that it is worth keeping up-to-date routing information for some time, despite the cost associated with transmitting route-update packets. Indeed, as Figure 5.8 shows, taking a larger active-state-timeout decreased the paging traffic for all applications. The difference between applications is also intuitive. An interactive application (e.g., `telnet`) sends one or more packets to the server triggering some action on the server side. This in turn results in new packet(s) being sent back. In the case of a local server paging only occurs when the time to process the request exceeds the active-state-timeout. This is rare, hence the low paging rate for local `telnet` sessions. The packet round-trip time adds to the server processing time for remote sessions. Figure 5.8 shows that in some cases the total response time exceeded 1 sec in

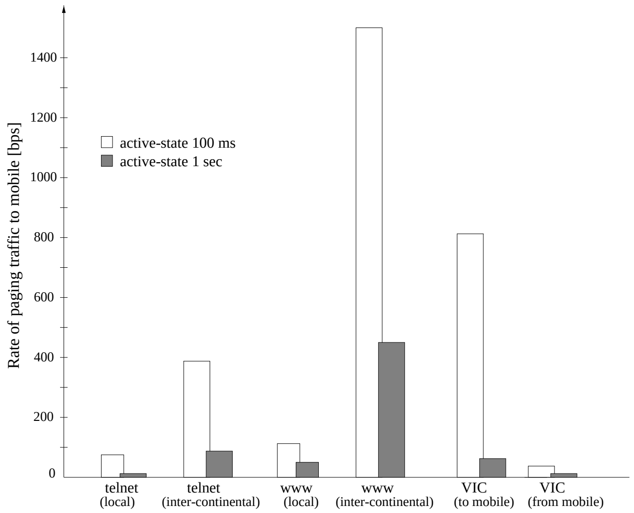

Figure 5.8: Paging traffic rate generated by some applications (measurement)

this example.

A similar difference can be observed between the local and remote WWW sessions. In this case paging occurred when the response time exceeded the active-state-timeout. We observe that this is rare when accessing a local server but frequent in the case of remote communication sessions. In addition, when large amounts of data are downloaded TCP sometimes stalled which also resulted in paging being triggered for the pauses that exceeded the active-state-timeout. Our final example application was `vic`. With the mobile host being the receiver of the video stream, we noted that the state machine at the mobile host rarely moves to idle state since each incoming packet resets the active-state-timeout. Because of the relatively large data rate (500 kbps) in our experiment the paging rate is significant, however. We observed that when the mobile host transmits packet video downlink packets only carry quality of service information and paging rate is negligible.

In the calculations we have assumed for the sake of simplicity that each node has a paging cache. Paging caches, however, are optional in Cellular IP nodes and a network is fully operational without paging cache at all. In a network with no paging cache, paging packets are broadcast and reach every base station. This does not impact the paging response time but represents a large load in the network. Paging caches are used to avoid broadcast paging in exchange for added node complexity and memory and processing cost. The importance of limiting paging traffic increases with increasing rate λ_P of sessions initiated from outside the Cellular IP network.

In the following set of simulation results we investigate the network's behaviour depending on the number and position of paging caches. The simulated configuration is illustrated in Figure 5.9 where **BSi** denotes base stations, **CC** denotes a concentrator node that does not have radio device and **GW** denotes the gateway router. In a series of simulation sessions we initiated data sessions to mobile users and measured the paging load in the network. The location of paging caches is subject to a trade-off between processing cost and network utilization. By creating more paging caches, the location information maintained about idle terminals is made more accurate thus reducing the number of cells paged in vain. Storing and updating paging cache mappings, on the other hand, represents a cost in terms of processing which in turn impacts node throughput.

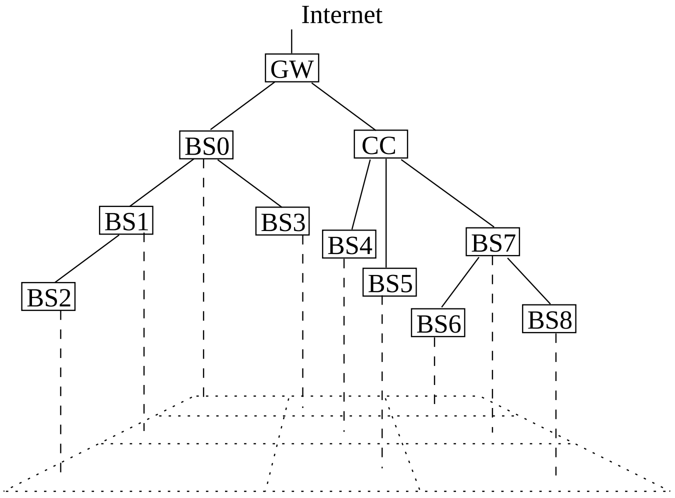

Figure 5.9: Network configuration for paging trade-off simulation

A numerical analysis of this trade-off needs to be based on costs associated with these effects and is hence necessarily specific to an implementation. In order to illustrate, however, the trade-off involved in configuring paging caches, in what follows we assign hypothetical cost values to paging a mobile terminal in vain and to updating a paging cache. We will analyze the network's behaviour using two different cost structures. In cost structure CS-1, we assume that the cost of paging a mobile terminal in a cell is equal to the cost of 10^3 cache update operations. In cost structure CS-2, the paging cost is equal to 10^4 cache update operations.

During these simulation sessions, a mobile host moved randomly among wireless cells. The time spent in a cell was an exponentially distributed random variable with 30 sec expected value. Correspondent hosts randomly established communication sessions toward mobile hosts and each session length was a random variable. The session inter arrival time and the session length were both exponentially distributed random variables such that the mobile host was active in 12.5% of the time. This way, rows of Table 5.1 which correspond to different mean session length values remain comparable in terms of paging cost. Columns of the table correspond to different paging cache setups. The first column contains results from simulation sessions where only the gateway had a paging cache. In the second column the gateway and another three centrally positioned nodes, namely **CC**, **BS0** and **BS7** had paging cache. Results shown in the third column were taken in a network where all nodes had paging cache.

Mean session length	gateway only	CC, BS0, BS7	all nodes
2 sec	12079	**4236**	4365
5 sec	5787	**2642**	2896
20 sec	1867	**1600**	1959
200 sec	**646**	1205	1615

Table 5.1: Aggregated cost (paging and PC updates) assuming CS-1 (simulation)

Table 5.1 summarizes simulation results using cost structure CS-1. Each reported value represents the aggregated cost of paging and cache updates during a 3000 sec simulation session. The absolute values hold no meaning, but their relations illustrate the behaviour of the network. In each row, we marked the lowest cost value by bold fonts. The Table shows that in the case of very long communication sessions, paging was rare and maintaining a large number of paging caches was unnecessary. In this scenario, it is advantageous to have paging cache in the gateway only and let all other nodes broadcast paging packets. In the case of shorter and more frequent data sessions, the significance of paging cost is larger and limiting the impact of paging packets in the network becomes more important. This requires configuring more paging caches. At this cost structure, however, even very frequent and very short data sessions do not justify configuring paging caches in every node because the cost associated with updating these caches is high.

Mean session length	gateway only	CC, BS0, BS7	all nodes
2 sec	11583	3084	**2836**
5 sec	5291	1487	**1354**
20 sec	1385	530	**497**
200 sec	183	**162**	194

Table 5.2: Aggregated cost (paging and PC updates) assuming CS-2 (simulation)

In Table 5.2 we report simulation results from exactly the same scenarios but assuming cost structure CS-2. In this case the paging cost is higher compared to the cache update cost which makes it more important to limit paging packet broadcasting. We observe that configuring paging cache in at least a few centrally positioned nodes is justified in all the scenarios studied. In the case of session lengths of 20 sec or less, it is in fact reasonable to place paging caches in all nodes. We conclude that the optimal setting of paging caches in a Cellular IP network depends on the costs associated with paging and with cache updates, but generally more paging caches should be used if the traffic typically consists of frequent short data sessions and less caches if long sessions are dominant.

5.5.3 Scalability

The design of Cellular IP was motivated by the vision of ubiquitous wireless Internet access where the same protocol may be used in small indoor systems up to metropolitan area wireless ISPs. This role can only be fulfilled by a protocol that adapts to a wide range of network sizes and shapes. In this section we study Cellular IP scalability through the scalability of our experimental implementation. Nodes are the universal building blocks of a Cellular IP network determining its performance and scalability. We will therefore start by looking at our node implementation and determining the performance limits of a Cellular IP node using off-the-shelf hardware. This is important because Cellular IP design promotes the use of commodity hardware to support network elements such as the node. Based on the performance measured in the node implementation we will next study three mobile networking scenarios that provide insight into the ability of Cellular IP to be customised to a wide range of mobility, application and network conditions.

Node Performance

In the first experiment we investigate the performance of our node implementation using a multi-homed 300 MHz Pentium PC. The solid line in Figure 5.10 shows the node throughput measured

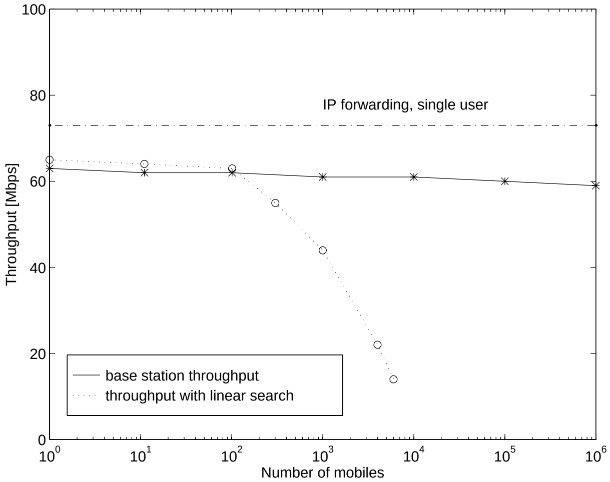

Figure 5.10: Node throughput (measurement)

using `ttcp` and 1500 byte packets. In these measurements we substituted a 100 Mbps Ethernet connection for the radio interface and created routing cache mappings for random IP addresses to emulate a large active user population. The fact that the throughput curve is hardly decreasing with increasing routing cache size suggests that in the scenarios that we have studied the performance bottleneck was not the cache lookup time. To verify this assumption we also measured the throughput that a single traffic stream achieved using standard IP forwarding through the same multihomed PC. As shown in Figure 5.10, the Cellular IP node throughput is somewhat below the standard IP throughput due to the additional packet processing involved with PCAP and additional packet copies across kernel and user space domains. In our implementation, the routing cache is stored in a binary tree to achieve fast lookups. We also measured the performance of the search algorithm and found that (in the range shown in Figure 5.10) the maximum packet rate can be well approximated as

$$\frac{2,500,000}{1+2\log(m)}$$

where m is the number of mappings in the cache. This explains that the performance bottleneck was found to be the network interface throughput rather than the search time over the range measured. For significantly larger user populations the cache lookup would likely become a bottleneck too. We did not, however, verify this thesis due to memory size constraints and unavailability of network interfaces that operated in excess of 100 Mbps. To illustrate this phenomenon we also experimented, however, with linear search instead of the binary tree. The throughput measured for the linear search is plotted with dotted line in Figure 5.10 and shows that when the user population exceeds 100 the cache lookup becomes the bottleneck. When the user population reaches 5000 the obtainable throughput is drastically diminished.

Application Scenarios

Building on the measurement results reported above we investigate the maximum number of users that our implementation can support for a number of network configurations. In particular, we study three different network scenarios that illustrate Cellular IP's ability to efficiently scale to meet local conditions by setting suitable values for the system parameters. The first scenario assumes a wireless IP telephony network demonstrating the importance of passive connectivity. In this scenario we show the large difference in mobility cost for a population of active and idle mobile hosts. Following this we evaluate the ability of Cellular IP to support a campus area wireless IP access network with high per user data rates. The final scenario shows how Cellular IP would operate if configured for a pico cellular network. In this scenario a large number of fast moving low data rate radio devices are supported.

In our *wireless IP telephony network* we assume that wireless subscribers generate 0.04 Erlang traffic and the mean call duration is 3 minutes. Speech is transmitted in 100 byte packets at constant 20 ms intervals, but no packets are transmitted in silent periods which account for 40% of the connection time in each direction. Active and idle users cross cell boundaries every 30 seconds. Let us first compute the optimal route-update time. Because of the silence suppression algorithm, the mean downlink data rate r is 24,000 bps and the parameter p is 0.4. This yields an optimal T_{ru} value of 404 ms using the same α and R_{ru} as before. The network load overhead associated with an active user C_a is 1616 bps. To compute the optimal paging-update time we need to calculate the rate of paging traffic to the mobile. Assuming that 50% of all calls are initiated to the mobile and that a connection setup message is 200 bytes long, the $\lambda_P R_P$ product is 0.18 bps. The optimal T_{pu} parameter is then 233 seconds and the cost C_i associated with an idle user is 7 bps. The average traffic generated by a subscriber, including uplink and downlink user data and mobility overhead, can be computed as the weighted sum of traffic generated by active and idle subscribers and is obtained as $0.04 \cdot (48,000 + 1616) + 0.96 \cdot 7 = 1991$ bps. Since in a Cellular IP network the gateway carries the aggregate traffic from the entire service area, the 65 Mbps throughput (see the evaluation above) limits the subscriber population to 30,000. According to the traffic model, out of this population the number of active users at any time is around 1200 which means that, according to our measurements shown previously, the cache search time does not become a bottleneck. Note that this implies that the supported population could increase by using base stations built on faster commodity hardware. The system configuration and engineering data related to the wireless IP telephony scenario are summarized in Table 5.2.

In the *campus wireless IP network* example we will take a simplified traffic model, assuming that wireless hosts engage in active data sessions 5% of the time and that the radio interface limits per user data rate to 500 kbps downlink and to 50 kbps uplink. We will further assume that the dominating traffic type is TCP but in 10% of the time the mobile host is receiving packets while it has no data to transmit forcing the transmission of periodic route-update packets. Based on our measurements, we estimate that $\lambda_P R_P$ is 400 bps and we assume that active and idle users cross cell boundaries every five minutes. From this, we can calculate the optimal timer values and the cost of mobility. The results are shown in the second column of Table 5.2. The number of supported users will be limited by the aggregate throughput, because the mobility overhead is negligible at this high data rate. The mean traffic generated by a user at any time instance is 28 kbps meaning that our base stations can support a maximum of 2300 users. As in the case of the telephony network example, the small number of active users at any time allows for quick cache lookups.

The final scenario addresses a *pico cellular wireless network* used by a computer assisted airport baggage tracking system. We select the tracking system because it represents a special wireless mobile application where the number of mobile units is very high and low cost is important. We assume that a small radio device ("badge") is attached to each bag as it proceeds through the service area. The device receives and sends messages during the sorting process. To allow for low cost badges, the transmission range is very small and handoffs occur every 2 seconds. Messages sent to the badge are 800 bytes long and each triggers a response of the same size. The transmission rate is 2 kbps in both directions. A message exchange happens every 30 seconds on the average. In this application, the mobile is never sending and receiving data at the same time, so the p parameter is 1. When computing the optimal paging-update time, we assume that the messages are sent in 50 byte packets and the first packet of each downlink message is sent as paging. Results from the analysis are shown in Table 5.2 and indicate that the cost of mobility is not negligible for this application even in the case of idle mobiles. Assuming that the badge radio is in active state only while it is receiving a message, the mean aggregate traffic (including uplink and downlink) will be 920 bps. Based on our node performance measurements we conclude that the maximum number of bags handled is in the order of 70,000. Because of the frequent messages and small packet size, routing caches contain mappings for approximately 10% of all badges at any one time. The total rate of routing cache lookups is 1.6 per second for each badge meaning that in this network the cache lookup algorithm is heavily loaded. However, according to measurements shown before, the current implementation can still handle this load using the commodity hardware and software support of our testbed.

	wireless telephony	campus network	pico cellular
data rate (r)	24 kbps	500 kbps	2 kbps
handoff frequency (T_H)	30 s	5 min	2 s
activity time (%)	4%	5%	10%
route-update time ($\hat{T}_{ru}$)	404 ms	138 ms	565 ms
paging-update time ($\hat{T}_{pu}$)	233 s	16 s	7 s
active mobility cost (C_a)	1616 bps	1180 bps	2890 bps
idle mobility cost (C_i)	7 bps	104 bps	230 bps
maximum population	30,000	2300	70,000

Table 5.3: Comparison of network scenarios

5.5.4 Service Quality Provisioning

Cellular IP deliberately trades handoff performance in exchange for implementation simplicity motivated by the desire to provide a cheap and robust solution primarily for best effort service. In Section 5.5.1 we studied handoff performance and showed the performance penalty of handoff simplicity. As a final step of our evaluation methodology we will now depart from the original design decision of minimum complexity and investigate Cellular IP's capability to extend toward the support of services beyond best effort.

Cellular IP can be considered as a modification of IP networking where routing state is dynamically modified in response to user mobility. The packet forwarding unit present in Cellular IP nodes is fundamentally identical to an IP router's packet forwarding engine. This similarity suggests that recent efforts to extend regular IP networks with service quality provisioning schemes (e.g., the differentiated service concept [27], [84]) will likely be adaptable to a Cellular IP network. The added difficulty of supporting service quality in Cellular IP stems from the handoffs. In the remainder of this section we will therefore focus on improving handoff performance.

Advance Binding

The Cellular IP base protocol's handoff algorithm (described in Section 4.5.1) is founded on a simplistic approach that allows some packet loss in exchange for minimizing handoff messaging instead of trying to guarantee zero loss. The loss of downlink packets during the handoff delay is investigated in Section 5.5.1.

In Cellular IP the routing information associated with a mobile host's old location is not cleared at handoff, rather, it times out as the associated timers expire. Before the timeout a period exists when both the old and new downlink routes are valid and packets are delivered through both base stations. This feature is used in the *advance binding* procedure that significantly improves handoff performance but still fits in the lightweight nature of the base protocol. Advance binding provides a probabilistic improvement instead of fully eliminating packet loss. An important feature of the advance binding handoff is that it improves handoff performance without any modification in Cellular IP nodes. The necessary changes are limited to the mobile host state machine where a single (temporary) state must be added to the base protocol's state machine. Nodes remain fundamentally stateless and unaware of handoff.

The purpose of the advance binding algorithm is to reduce handoff latency. To this avail, the routing cache mappings associated with the new base station must be created before the actual handoff takes place. When the mobile host decides handoff, it sends an *advance binding packet*, which is technically a route-update packet, to the new base station but immediately returns to listening to the old base station. While the host is still in connection with the old base station, the advance binding packet configures routing cache mappings associated with the new base station. After an *advance binding delay* (T_a), the host can perform a regular handoff.

In the case of advance binding handoff, the handoff delay period is started at the time when the mobile hosts transmits the advance binding packet. If the mobile host performs the actual handoff later than the expiration of the handoff delay, the routing state associated with the new base station will have been established and packets will continue to arrive to the mobile host immediately after handoff. Simulation results plotted in Figure 5.11 show that advance binding reduces and eventually eliminates packet loss at handoff. The figure shows the number of packets lost at handoff in a simulation configuration where handoff delay was 10 ms. (Configuration file: `handoff_loss_cbr_semisoft_Finterval`.) The three curves correspond to downlink rates of 20, 50 and 100 packets per second. By varying the advance binding delay T_a from 0 (corresponding to the original handoff scheme) to 14 ms we gradually improved handoff performance by first shrinking and then eliminating handoff delay. Figure 5.11 shows that as the advance binding delay reaches 10 ms, packet loss is eliminated regardless of the downlink packet rate. By further increasing the advance binding delay the performance is unchanged. If, however, the advance binding delay reaches the value of route-timeout (not shown in the Figure) then the routing state created by the advance binding packet will be cleared before the actual handoff takes place and the advance binding algorithm brings no improvement.

We conclude that the advance binding delay should be larger than the handoff delay but less than the system's route-timeout parameter. Since in most networks the handoff delay will be in the order of milliseconds and route-timeout around 100 ms, the selection of a proper advance binding delay is easy.

The advance binding handoff is attractive because of its simplicity. It requires only minor modification in the mobile host state machine and no modification at all in the nodes. Experimental results plotted in Figure 5.12 show, however, that advance binding does not always eliminate handoff packet loss. In this figure we plotted the throughput (measured in the testbed) of a downlink TCP session in the presence of hard and advance binding handoffs with solid and dotted lines, respectively. In contrast to simulation results shown previously, this figure shows that TCP throughput decreases with increasing handoff frequency despite advance binding. Investigating the difference between simulation and experimental results we have found that advance binding fails to provide seamless handoff if transmission delays from the cross-over node to the old and new base stations are very different. In such cases the data session is disturbed at handoff despite the elimination of handoff delay because the packet streams transmitted through the two base stations will not be synchronized. The mobile host continues to receive packets immediately after handoff, but does not necessarily receive them in the correct order. If the

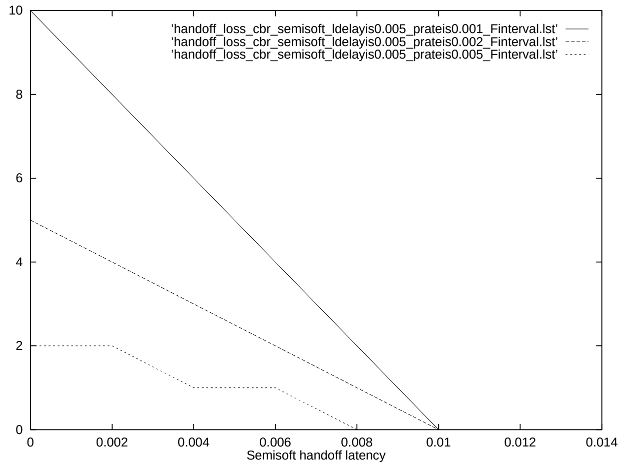

Figure 5.11: Packet loss at advance binding handoff (simulation)

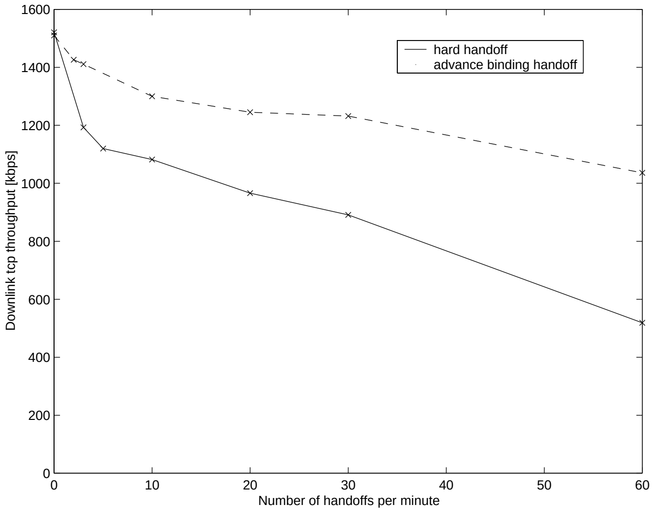

Figure 5.12: TCP throughput at hard handoffs and advance binding handoffs (measurement)

new base station is “behind” the old one, the mobile host will receive duplicate packets at handoff. IP based applications tolerate sporadic duplicate packets but can be disturbed by multiple duplicates of the same packet. Multiple duplicates can be generated if multiple handoffs occur within a short time period. To avoid such handoff ping-ponging, signal strength based handoff control can involve hysteresis. If, however, the new base station is “ahead” compared to the old one, packet loss can occur at handoff. These losses can not be compensated for because Cellular IP base stations do not buffer packets. The decline of TCP throughput with increasing handoff frequency shown in Figure 5.12 is a result of packet losses due to this condition. We conclude that advance binding decreases handoff packet loss in exchange for very little added complexity but does not eliminate it, especially when handoff occurs between base stations that operate in different traffic conditions (resulting in different delays).

Semisoft Handoff

While advance binding ensures that the mobile host continues to receive packets immediately after handoff, experimental results shown in Figure 5.12 have demonstrated that it alone does not always provide smooth handoff. This observation motivates the design of Cellular IP *semisoft handoff*. Semisoft handoff has two components. First, similar to advance binding handoff, it involves establishing the downlink route toward the new base station before handoff actually takes place. Unlike in advance binding handoff, however, semisoft handoff uses a special packet called *semisoft packet* for this purpose. A semisoft packet is transmitted through the new base station before handoff and it propagates through the network to create Routing Cache mappings in nodes on the way.

The second component of the semisoft handoff procedure is based on the observation that perfect synchronization of the data streams through the old and new base stations is needless. The condition resulting in packet loss at advance binding handoffs can be eliminated by temporarily introducing into the new path a constant delay sufficient to compensate for the time difference between the two streams. The delay will ensure that the new base station is behind, rather than ahead of the old one. This will result in duplicate packets instead of packet loss. Converting packet loss into duplicate packets is advantageous because these duplicate packets can be eliminated at the mobile host which is naturally aware of a semisoft handoff being in progress.

Unlike advance binding handoff, semisoft handoff requires support of Cellular IP nodes in introducing the said temporary delay. Semisoft handoff does still not require, however, any interaction between nodes, base stations and the mobile host. The delay is introduced in the new data stream by the cross-over node. This node knows that a semisoft handoff is in progress from the fact that a semisoft packet arrives from a mobile host that has mapping to another interface. The mapping created by the semisoft packet in the cross-over node has a flag to indicate that downlink packets routed by this mapping must pass a delay element before transmission. We recall that similar to other mappings, this information is stored as soft state and is cleared after the route-timeout. In normal conditions it is actually cleared before the expiration of the soft-state timer because mobile hosts send a route-update packet immediately after handoff. This packet will update the mapping created by the semisoft packet and clear the flag explicitly. Clearing the flag causes all packets eventually stored in the delay device to be forwarded to the mobile host. We point out that base stations only need a small pool of delay buffers since very few mobile hosts will simultaneously be in semisoft position.

Figure 5.13 shows handoff packet measurement results at hard and semisoft handoff. The experimental conditions for hard and semisoft handoff were identical. The mobile host received 100 byte UDP packets at rates of 25 and 50 packets per second (pps). Each point on the graph was obtained by averaging loss measurements over 50 handoffs. In these experiments, the new downlink packet stream at semisoft handoffs was delayed in the cross-over node by a buffer

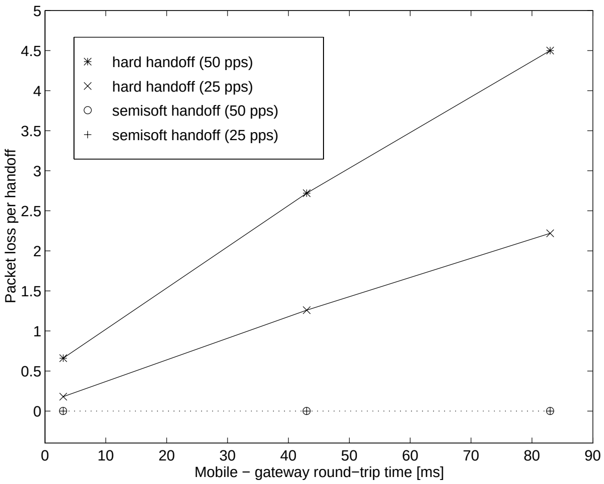

Figure 5.13: Packet loss at hard and semisoft handoff (measurement)

holding each packet until the arrival of the next downlink packet. When the semisoft handoff is completed, the last packet is cleared from the buffer and is sent to the mobile host. Figure 5.13 shows that semisoft handoff eliminated packet loss entirely. Note that buffering a single packet in the delay element was sufficient to eliminate loss even in the case of a large round-trip time when hard handoff resulted in the loss of up to four packets. This is because the semisoft buffering is only to compensate for the *difference* between the transmission times along the old and new paths and not for the entire round-trip time between the mobile and the cross-over point.

The next experiment investigated the improvement in TCP throughput gained using semisoft handoff. The experimental conditions for the semisoft, advance binding and hard handoff measurements were identical. The semisoft delay element in this case was an 8-packet circular buffer. The dashed curve in Figure 5.14 shows TCP throughput to a mobile host that performs hard, advance binding or semisoft handoffs at increasing rates. We observe that using semisoft handoff, packet loss is entirely eliminated and a slight disturbance only remains due to the transmission delay variations encountered at handoff. We point out that even for one handoff per second, the throughput is almost identical to that observed for a static host.

5.6 Discussion

In this chapter we have evaluated the performance of Cellular IP focusing on handoff performance, mobility management cost and scalability. In particular, we have studied the performance penalty of the simplistic approach taken in Cellular IP handoff and have evaluated the service quality seen by IP applications at handoff. We have quantified the cost of host mobility management and have determined the optimal setting of protocol parameters associated with mobility management. We have analyzed the impacts of configuring the optional paging cache units in Cellular IP networks. We have looked at the possibility of extending Cellular IP with service quality support and have analyzed improved handoff schemes.

We recall that one primary intention in designing the Cellular IP architecture and protocol was to investigate the possibility of building cellular wireless networks around the IP paradigm

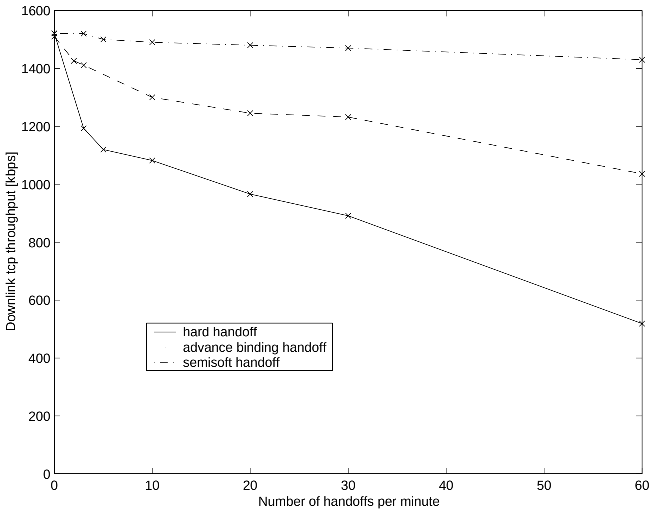

Figure 5.14: TCP throughput at hard, advance binding and semisoft handoffs (measurement)

and to locate the problem areas in such networks. The performance analysis results shown in this chapter indicate the feasibility of the concept. Cellular IP has been found to operate smoothly even in highly mobile environments despite the fact that in contrast to traditional cellular systems it does not have explicit signalling mechanisms, hard states or rigid network structure. The distributed nature of the protocol, on the other hand, makes service quality provisioning more difficult than in the case of centralized approaches such as traditional cellular systems. However, initial results on advanced handoff techniques, also reported in this chapter, indicate that improving service quality is feasible without radically departing from the simplistic approach of Cellular IP.

A number of approaches founded on ideas similar to the Cellular IP concept have recently been published. The Hawaii protocol [87] operates in a network largely similar to Cellular IP, but it requires base stations to run IP protocol stacks in addition to the mobility specific functionalities. Hawaii uses this property to exchange explicit signalling messages between base stations which makes the protocol more complex than Cellular IP but may potentially allow for higher service quality by reducing packet loss at handoffs. Further new approaches to local mobility are the Thema [89] and Dynamics [88] protocols that both use IP tunnels to mobile hosts.

Chapter 6

Conclusions

6.1 Contribution of Thesis

In this dissertation we have addressed some of the challenges that existing and future cellular mobile systems will meet in an environment of ubiquitous wireless Internet availability. We have argued that the development of small palmtop computers and of web based services demands the development of efficient and flexible cellular wireless access technologies. These technologies will build on cellular telephony technology on one hand and will incorporate IP concepts on the other hand.

Cellular access networks in this new environment will face new challenges in terms of network throughput requirements and service quality constraints. In order to plan and manage cellular networks that serve a variety of applications providing the expected service to each of them, new performance analysis methods are required. In Chapter 2, we have proposed and evaluated hybrid-hierarchical simulations, a new simulation architecture that can provide significant increase in simulation performance in the analysis of mobile communications systems. Results reported in this chapter are also found in [C6] and [J3].

The increasing user population and data rates call for increasing cellular system throughputs. Cellular network operators can increase system throughput by decreasing the cell size which, however, results in increased handoff frequency. To avoid the frequent dropping of connections at handoff, the operator must then reserve an increasing amount of capacity in wireless cells for future handoff attempts. In Chapter 3 we have studied the impact of such advance reservations on system efficiency (that is, throughput). We have analyzed the efficiency of cellular systems where resources are reserved for future handoff attempts in a deterministic way, and have calculated the decrease of efficiency in response to increasing user speed, decreasing cell size and increasing application heterogeneity. We have studied systems with statistical resource reservation policies and have calculated the upper bound of system efficiency as a function of the tolerated handoff blocking probability, regardless of what admission control and reservation policy are used. These results are also reported in [C1] and [J4].

Third generation cellular mobile systems and Internet host mobility proposals both address the issue of wireless IP host mobility, but do it from two different directions. In Chapter 4 we have argued that both approaches have drawbacks and that a ‘third way’ is required that uses cellular telephony techniques but implements them around the IP paradigm. We have established requirements for a cellular access networking solution and have derived network design principles. Based on these principles, we have designed Cellular IP, a new architecture and protocol that allows for providing Internet access to wireless mobile users in a simple, flexible and scalable manner. Cellular IP is described in details in [J2], in [C2] and in [D1].

Finally, in Chapter 5 we have analyzed Cellular IP from the aspects of handoff quality, mobility management cost, scalability and service quality provisioning. For the analysis, we have used a combination of analytical, simulation and experimental methods. We have found that Cellular IP performs well even in environments of very high mobility and have quantified the performance penalty of the simplistic approach taken in Cellular IP design. We have evaluated the impact of network planning decisions such as protocol parameter settings and paging cache distribution and have derived suggestions on setting these in an optimal way. We have also identified a few shortcomings of the Cellular IP protocol, primarily in the area of service quality provisioning. Results of the performance analysis are also found in [T1] and in [C3].

6.2 Future Research Directions

The work reported in this dissertation opens the way for continued research in a number of directions. These are described in detail in each chapter. A particularly interesting area for future research is the design and enhancement of protocols in support of wireless mobile Internet access. A number of protocols based on ideas similar to Cellular IP have recently been proposed [87], [89], [88]. Most of these proposals combine cellular telephony techniques with IP principles. More work is required in this area to create the technology that truly becomes the 'wireless pendant' of IP providing a ubiquitous coverage for wireless hosts over a wide range of heterogeneous environments. We believe that similar to Cellular IP, such a solution must take a fundamentally simplistic approach but it must also allow for added functionality in support of service quality, security and advanced location management techniques. It must operate smoothly over wireless technologies ranging from short range radios such as Bluetooth [90] up to long range, possibly satellite radios. The work reported in this dissertation will be continued to answer these challenges.

Bibliography

[1] Ramachandran Ramjee, Ramesh Nagarajan, Don Towsley, “On Optimal Call Admission Control in Cellular Networks,” *Wireless Networks Journal*, Vol. 3, No. 1, 1997, pp. 29-41.

[2] Aylin Yener, Christopher Rose, “Local Call Admission Policies for Cellular Networks Using Genetic Algorithms,” *CISS’95*, March 1995.

[3] D.A. Levine, I.F. Akyldiz, M. Naghshineh, “The Shadow Cluster Concept for Resource Allocation and Call Admission in ATM-Based Wireless Networks,” in *Proc. 1st Annual International Conference on Mobile Computing and Networking*, November 1995.

[4] Daehyoung Hong, Stephan S. Rappaport, “Traffic Model and Performance Analysis for Cellular Mobile Radio Telephone Systems with Prioritized and Nonprioritized Handoff Procedures,” *IEEE Trans. Veh. Technol.*, Vol. VT-35, No. 3, August 1986, pp. 77-92.

[5] Arak Sutivong, Jon M. Peha, “Call Admission Control Algorithms For Cellular Systems: Proposal and Comparison,” *Working paper*, http://www.ece.cmu.edu/ peha/cellular_admission.ps

[6] Anthony S. Acampora, Mahmoud Naghshineh, “An Architecture and Methodology for Mobile-Executed Handoff in Cellular ATM Networks,” *IEEE J. Select. Areas Commun*, Vol. 12, No. 8, October 1994, pp. 1365-1375.

[7] K. Lee, “Adaptive Network Support for Mobile Multimedia,” in *Proc. 1st Annual International Conference on Mobile Computing and Networking*, November 1995.

[8] Mahmoud Naghshineh, Mischa Schwartz, Antony S. Acampora, “Issues in Wireless Access Broadband Networks,” *IBM Research Report*, RC 19980 (11/16/94).

[9] Anup K. Talukdar, B.R. Badrinath, Arup Acharya, “On Accomodating Mobile Hosts in an Integrated Services Packet Network,” in *Proc. IEEE INFOCOM*, 1997.

[10] Mohammad Saquib, Roy Yates, “Optimal Call Admission to a Mobile Cellular Network,” in *Proc. IEEE Vehicular Technology Conf.*, July 1995.

[11] Aylin Yener, Christopher Rose, “Near-Optimal Call Admission Policies for Cellular Networks Using Genetic Algorithms,” *Wireless’94*, Calgary, July 1994.

[12] Aylin Yener, Christopher Rose, “Genetic Algorithms Applied to Cellular Call Admission Problem: Local Policies,” *IEEE Trans. Veh. Technol.*, February 1997.

[13] G. Brasche, B. Walke, “Concepts, Services and Protocols of the New GSM Phase 2+ General Packet Radio Service,” *IEEE Commun. Mag.*, pp. 94-104, August 1997.

[14] Charles Perkins, editor, “IP Mobility Support,” *Internet RFC* 2002, October 1996.

[15] Pravin Bhagwat, Charles Perkins, Satish Tripathi, “Network Layer Mobility: an Architecture and Survey,” *IEEE Personal Commun.*, Vol. 3, No. 3, pp. 54-64, June 1996.

[16] M. Mouly, M-B. Pautet, “The GSM System for Mobile Communications,” *published by the authors*, ISBN 2-9507190-0-7, 1992.

[17] Charles Perkins, "Mobile-IP Local Registration with Hierachical Foreign Agents," *Internet Draft*, draft-perkins-mobileip-hierfa-00.txt, February 1996, Work in Progress.

[18] S.F. Foo, K.C. Chua, "Regional Aware Foreign Agent (RAFA) for Fast Local Handoffs," *Internet Draft*, draft-chuafoo-mobileip-rafa-00.txt, November 1998, Work in Progress.

[19] Ramon Cáceres, Venkata N. Padmanabhan, "Fast and Scalable Handoffs for Wireless Internetworks," in *Proc. ACM Mobicom*, 1996.

[20] M.C. Chuah, Y. Li, "Distributed Registration Extension to Mobile IP," *Internet Draft*, draft-chuahli-mobileip-dremip-00.txt, October 1997, Work in Progress.

[21] H. Balakrishnan, S. Seshan, R. Katz, "Improving Reliable Transport and Handoff Performance in Cellular Wireless Networks," *ACM Wireless Networks* 1(4), December 1995.

[22] Jayanth Mysore, Vaduvur Bhargavan, "A New Multicasting-based Architecture for Internet Host Mobility," in *Proc. ACM Mobicom*, Budapest, October 1997.

[23] Fumio Teraoka, "Virtual Internet Protocol version 2," *Internet Draft*, draft-teraoka-mobileip-vip-01.txt, December 1995, Work in Progress.

[24] John Ioannidis, Dan Duchamp, Gerald Q. Maguire Jr., "IP-Based Protocols for Mobile Internetworking," in *Proc. ACM SIGCOMM'91*, pp. 234-245, September 1991.

[25] "WaveLAN Air Interface," Data Manual, *AT&T Corporation*, Doc. No. 407-0024785 Rev. 2 (draft), July 11, 1995.

[26] David B. Johnson, Charles Perkins, "Route Optimization in Mobile IP," *Internet Draft*, draft-ietf-mobileip-optim-07.txt, November 1998, Work in Progress.

[27] D. Clark, J. Wroclawski, "An Approach to Service Allocation in the Internet," *Internet Draft*, draft-clark-diff-svc-alloc.00.txt, July 1997, Work in Progress.

[28] Steve McCanne, Van Jacobson, "The BSD Packet Filter: A New Architecture for User-level Packet Capture," in *Proc. USENIX 93* San Diego, 1993.

[29] Zygmunt J. Haas, "Networking Issues in Wireless Technologies," tutorial presentation, *IEEE Workshop on Multiaccess, Mobility and Teletraffic for Wireless Communications (MMT'98)*, October 21-23, 1998, Washington DC.

[30] David B. Johnson, Charles Perkins, "Mobility Support in IPv6," *Internet Draft*, draft-ietf-mobileip-ipv6-07.txt, November 1998, Work in Progress.

[31] Mark Stemm, Randy H. Katz, "Vertical Handoffs in Wireless Overlay Networks," *ACM Mobile Networking (MONET)*, Special Issue on Mobile Networking in the Internet, Summer 1998.

[32] Javier Gomez, Sanghyo Kim, András Valkó, Andrew Campbell, "Cellular IP Implementation and Validation," *Technical Report*, February 1999, Work in Progress.

[33] Andrew Myles, David Skellern, "Comparing Four IP Based Mobile Host Protocols," *Computer Networks and ISDN Systems*, pp. 349-356, November 1993.

[34] G. Eneroth, M. Johnsson, "ATM Transport in Cellular Networks," in *Proc. Int. Switching Symp. (ISS'97)*, Toronto, September 1997.

[35] A. Magi, "Performance Evaluation Algorithms for Telephone Networks Taking into Account the Holding Time of Unsuccessful Calls," *MSc thesis*, Technical Univ. Budapest, Dept. Telecom. Telematics, Budapest, 1996.

[36] G.S. Fishman, "Accelerated Accuracy in the Simulation of Markov Chains," *Operat. Res.*, Vol. 31., 1983.

[37] G. Kesidis, A. Singh, D. Cheung, W.W. Kwok, "Feasibility of Fluid Event-Driven Simulation for ATM Networks," *IEEE Global Telecommunications Conference* (Globecom '96), 1996, pp. 2013-2017.

[38] Zs. Haraszti, I. Dahlquist, A. Farago, T. Henk, "PLASMA - An Integrated Tool for ATM Network Operation," in *Proc. Int. Switching Symp. (ISS '95)*, 1995.

[39] *B-ISDN ATM Adaptation Layer Type 2 Specification*, Draft ITU-T Recommendation I.363.2

[40] M. Frater, B. Bitmead, R. Kennedy, B. Anderson, "Fast Simulation of Rare Events Using Reverse-Time Models," in *Proc. ITC Specialist Seminar*, Adelaide, Paper No. 9.1, 1989.

[41] M. Coppola, "A Cosimulation Environment with Opnet and VHDL," *13th UK Workshop on Performance Engineering of Computer and Telecom. Systems (UKPEW '97)*, Ilkley, UK, 1997.

[42] K. Hines, G. Borriello, "Dynamic Communication Models in Embedded System Co-Simulation," *34th ACM Design Automation Conference*, Anaheim, CA, 1997.

[43] A.M. Law, W.D. Kelton, "Simulation, Modeling and Analysis," *McGraw-Hill*, New York, 1991.

[44] M. Villén-Altamirano, J. Villén-Altamirano, "RESTART: A Method for Accelerating Rare Event Simulation," in *Proc. 13th ITC, Queueing, Performance and Control in ATM*, pp. 71-76, 1991.

[45] C. Bisdikian, J.S. Lew, A.N. Tantawi, "The Generalized $D^{[X]}/D/1$ queue: A flexible computer communications model," *Telecommunications Systems*, Vol. 6, No. 2, 1996.

[46] P.M. Lin, B.J. Leon, C.R. Stewart, "Analysis of Circuit-Switched Networks Employing Originating Office Control with Spill," *IEEE Trans. Commun.*, Vol. COM-26, No. 6, pp. 754-765, 1978.

[47] R.Y. Rubinstein, R. Marcus, "Efficiency of Multivariate Control Variates in Monte Carlo Simulation," *Operat. Res.*, Vol. 33, 1985.

[48] S.S. Lavenberg, T.L. Moeller, P.D. Welch, "Statistical Results on Control Variables with Application to Queueing Network Simulation," *Operat. Res.*, Vol. 30, 1982.

[49] V.S. Frost, W.W. Larue, K.S. Shanmugam, "Efficient Techniques for the Simulation of Computer Communication Networks," *IEEE J. Select. Areas Commun.*, Vol. 6, No. 1, pp. 146-157, 1988.

[50] Performance evaluation and design of multiservice networks, *COST 224*, 1992.

[51] M.C. Jeruchim, "Techniques for Estimating the Bit Error Rate in the Simulation of Digital Communication Systems," *IEEE J. Select. Areas Commun.*, Vol. SAC-2, No. 1, pp. 153-170, 1984.

[52] S.B. Weinstein, "Estimation of Small Probabilities by Linearization of the Tail of a Probability Distribution Function," *IEEE Trans. on Commun. Technology*, Vol. COM-19, 1971.

[53] I. Berberana, "A Method for Estimating Loss Probabilities in Networks of Queues: the Extreme Value Theory Approach," in *Proc. 3rd IEEE Intern. Workshop on CAMAD of Networks and Links*, Toronto, Paper No. 2.2, 1990.

[54] V. Dijk, E. Aanen, H. van der Berg, J.M. van Noortwijk, "Extrapolating ATM Simulation Results using Extreme Value Theory," in *Proc. 13th ITC, Queueing, Performance and Control in ATM*, 1991.

[55] E. Gustafsson, R. Ronngren, "Fluid Traffic Modelling in Simulation of a CAC Scheme for ATM Networks," in *Proc. IEEE MASCOTS '97*, Haifa, Israel, 1997.

[56] P.J.P. O'Reilly, J.L. Hammond, "An Efficient Simulation Technique for Performance Studies of CSMA/CD Local Networks," *IEEE J. Select. Areas Commun*, Vol. SAC-2, No. 1, 1984.

[57] S.S. Lavenberg, P.D. Welch, "Using Conditional Expectation to Reduce Variance in Discrete Event Simulation," in *Proc. 1979 Winter Simulation Conference*, 1979.

[58] C.D. Pham, S. Fdida, "Perspectives in Performance Evaluation of Large ATM Networks," in *Proc. 5th IFIP Workshop on Performance Modeling and Evaluation of ATM Networks*, Ilkley, UK, 1997.

[59] R.M. Fujimoto, D. Nicol, "Parallel Simulation Today," *Annales of Operations Research: Simulation and Modeling* (ed. O. Balci), Vol. 53, 1994.

[60] K.S. Shanmugan, P. Balaban, "A modified Monte Carlo simulation technique for the evaluation of error rate in digital communication systems," *IEEE Trans. Commun.*, Vol. COM-28, 1980.

[61] Brian G. Marchent, "Third Generation Mobile Systems for the Support of Mobile Multimedia based on ATM Transport," tutorial presentation, *5th IFIP Workshop*, Ilkley, July, 1997.

[62] R. Fantacci, L. Zoppi, "A CDMA Wireless Packet Network for Voice-Data Transmissions," *IEEE Trans. Commun.*, Vol. 45, No. 10, October, 1997.

[63] M. Kojo, K. Raatikainen, M. Liljeberg, J. Kiiskinen, T. Alanko, "An Efficient Transport Service for Slow Wireless Telephone Links," *IEEE J. Select. Areas Commun*, Vol. 15, No. 7, September, 1997.

[64] Richard D. Carsello, Reuven Meidan, Stephen Allpress, Fran O'Brien, Joseph A. Tarallo, Norm Ziesse, Arun Arunachalam, Jose M. Costa, Ermanno Berruto, Richard C. Kirby, Allan Maclatchy, Fumio Watanabe, Howard Xia, "IMT-2000 Standards: Radio Aspects," *IEEE Personal Commun.*, Vol. 4, No. 4, August 1997.

[65] Raj Pandya, Davide Grillo, Edgar Lycksell, Philippe Mieybégué, Hideo Okinaka, Masami Yabusaki, "IMT-2000 Standards: Network Aspects," *IEEE Personal Commun.*, Vol. 4, No. 4, August 1997.

[66] Ermanno Berruto, Giovanni Colombo, Pantelis Monogioudis, Antonella Napolitano, Kyriacos Sabatakakis, "Architectural Aspects for the Evolution of Mobile Communications Toward UMTS," *IEEE J. Select. Areas Commun*, Vol. 15, No. 8, October 1997, pp. 1477-1487.

[67] M.J. McTiffin, A.P. Hulbert, T.J. Ketseoglou, W. Heimsch, G. Crisp, "Mobile Access to an ATM Network Using a CDMA Air Interface," *IEEE J. Select. Areas Commun*, Vol. 12, No. 5, June 1994.

[68] F. Leite, R. Engelman, S. Kodama, H. Mennenga, S. Towaij, "Regulatory Considerations Relating to IMT-2000," *IEEE Personal Commun.*, Vol. 4, No. 4, August 1997.

[69] R. Cáceres and V.N. Padmanabhan, "Fast and Scalable Wireless Handoffs in Support of Mobile Internet Audio," *ACM MONET Journal*, Vol. 3, No. 4, December 1998.

[70] S. Seshan, H. Balakrishnan, R.H. Katz, "Handoffs in Cellular Wireless Networks: The Daedalus Implementation and Experience," *Kluwer International Journal on Wireless Communication Systems*, January 1997.

[71] H. Balakrishnan, V.N. Padmanabhan, S. Seshan, R.H. Katz, "A Comparison of Mechanisms for Improving TCP Performance over Wireless Links," *IEEE/ACM Trans. on Networking*, December 1997.

[72] Ramon Cáceres, Liviu Iftode, "Improving the Performance of Reliable Transport Protocols in Mobile Computing Environments," *IEEE J. Select. Areas Commun*, Vol. 13, No. 5, June 1995.

[73] Raj Yavatkar, Namrata Bhagawat, "Improving End-to-End Performance of TCP over Mobile Internetworks," in *Proc. Workshop on Mobile Computing Systems and Applications (Mobile '94)*, December 1994.

[74] Jorge A. Cobb, Prathima Agrawal, "Congestion or Corruption? A Strategy for Efficient Wireless TCP Sessions," *1st IEEE Symposium on Computers and Communications (ISCC '95)*, Alexandria, Egypt, June 1995.

[75] A. Bakre, B. Badrinath, "I-TCP: Indirect TCP for Mobile Hosts," *Technical Report* DCS-TR-314, Dept. Computer Science, Rutgers University, October 1994.

[76] Elan Amir, Hari Balakrishnan, Srinivasan Seshan, Randy H. Katz, "Efficient TCP over Networks with Wireless Links," in *Proc. Fifth Workshop on Hot Topics in Operating Systems (HotOS-V)*, Orcas Island, WA, May 1995.

[77] B. R. Badrinath, A. Bakre, T. Imielinski, R. Marantz, "Handling Mobile Clients: A Case for Indirect Interaction," in *Proc. IEEE Fourth Workshop on Workstation Operating Systems*, Aigen, Austria, October 1993.

[78] Richard W. Stevens, "TCP/IP Illustrated," *Addison-Wesley*, ISBN 0-201-63346-9, 1994.

[79] "UCB/LBNL/VINT Network Simulator - ns (version 2)", ns home page, http://www-mash.cs.berkeley.edu/ns/ns.html.

[80] R. Braden, L. Zhang, S. Berson, S. Herzog, S. Jamin, "Resource Reservation Protocol (RSVP)," *Internet Draft*, September 1997.

[81] F.P. Kelly, "Effective bandwidths at multi-class queues," *Queueing Systems*, 9, pp. 5-16, 1991.

[82] T. Zimmerman, L. Boemer, J. Eichinger, B. Gunzelmann, W. Liegl, "CDMA and ATM — Ideal partners," *Telcom Rep. Int.* 18, No. 3, pp. 22-25, 1995.

[83] S. Hoff, M. Meyer, A. Schieder, "A Performance Evaluation of Internet Access via the General Packet Radio Service of GSM," in *Proc. IEEE Vehicular Technology Conf.*, pp. 1760-1764, 1998.

[84] K. Nichols, V. Jacobson, L. Zhang, "A Two-bit Differentiated Services Architecture for the Internet," *Internet Draft*, draft-nichols-diff-svc-arch-00.txt, November 1997, Work in Progress.

[85] John Ousterhout, "Tcl and the Tk Toolkit," *Addison-Wesley*, ISBN 0-201-63337-X, May 1998.

[86] http://www.ericsson.com/wireless/products/mobsys/gsm/gsm.shtml

[87] R. Ramjee, T. La Porta, S. Thuel, K. Varadhan, "IP micro-mobility support using HAWAII," *Internet Draft*, draft-ramjee-micro-mobility-hawaii-00.txt, February 1999, Work in Progress.

[88] Dynamics home page, http://www.hut.fi/ kmustone/dynamics/

[89] P. McCann, T. Hiller, J. Wang, A. Casati, C. Perkins, P. Calhoun, "Transparent Hierarchical Mobility Agents (THEMA)," *Internet Draft*, draft-mccann-thema-00.txt, March 1999, Work in Progress.

[90] Jaap Haartsen, "Bluetooth – The Universal Radio Interface for Ad Hoc, Wireless Connectivity," *Ericsson Review*, No. 3, 1998.

Publications

[J] **Journal Papers and Book Chapters**

[J1] A. Valkó, A. Rácz, G. Fodor, "Voice QoS in 3rd Generation Mobile Systems," *IEEE J. Select. Areas Commun*, Vol. 17, No. 1, January 1999, pp. 109-123.

[J2] A. Valkó, "Cellular IP: A New Approach to Internet Host Mobility," *ACM SIGCOMM Computer Communication Review*, Vol. 29, No. 1, January 1999, pp. 50-65.

[J3] A. Valkó, A. Rácz, G. Fodor, L. Westberg, "An Efficient Simulation Environment for 3rd Generation Cellular Networks," *IFIP Book on Performance of ATM Networks, Kluwer Academic Publishers*, Vol. 4, 1999, to appear.

[J4] A. Valkó, A. Campbell, "An Efficiency Limit of Cellular Mobile Systems," *Computer Communications Journal*, Special Issue on Recent Advances in Mobile Communications Networks, 1999, to appear.

[D] **Internet Drafts**

[D1] A. Valkó, A. Campbell, J. Gomez, "Cellular IP," *Internet Draft*, draft-valko-cellularip-00.txt, November 1999.

[C] **Conference and Workshop Papers**

[C1] A. Valkó, A. Campbell, "An Efficiency Limit of Cellular Mobile Systems," in *Proc. IEEE MMT'98 Workshop on Multiaccess, Mobility and Teletraffic for Wireless Communications*, October 21-23, 1998, Washington DC.

[C2] A. Valkó, "Cellular IP - A Local Mobility Protocol," in *Proc. IEEE 13th Annual Computer Communications Workshop*, September 11-14, 1998, Oxford, Tennessee.

[C3] A. Valkó, J. Gomez, S. Kim, A. Campbell, "Performance of Cellular IP Access Networks," submitted to *Sixth IFIP International Workshop on Protocols For High-Speed Networks (PfHSN '99)*, 1999.

[C4] A. Valkó, "Cellular IP," *First Workshop on IP Quality of Service for Wireless and Mobile Networks*, Aachen, Germany, April 1999.

[C5] Andrew T. Campbell, Javier Gomez, András G. Valko, "An Overview of Cellular IP," invited paper, *IEEE Wireless Communications and Networking Conference*, New Orleans, September 1999, to appear.

[C6] A. Valkó, L. Westberg, "A Simulation Environment for ATM-Based Cellular Networks," in *Proc. 5th IFIP Workshop on Performance Modelling and Evaluation of ATM Networks*, July 21-23, 1997, Ilkley.

[C7] B. Eged, I. Novák, A. Valkó, "Behavioural Time Modeling and DSP Simulation of Interconnections," *COST 229* WGs 3+5+6 Workshop, September 7-9, 1993, Budapest.

[C8] B. Eged, I. Novák, P. Bajor, A. Valkó, "Investigation of Interconnections in Digital Systems by Circuit Simulation Software," in *Proc. CAMP 93, CAD/CAM and Multimedia Conference*, September 28-30, 1993, Budapest.

[C9] B. Eged, I. Novák, P. Bajor, A. Valkó, “New Time Modeling for Analysis of Digital Systems,” in *Proc. CAMP 93, CAD/CAM and Multimedia Conference*, September 28-30, 1993, Budapest.

[C10] J. Németh, A. Valkó, T. Ben Meriem, M. Mary, “Handling the Interleave Code-Matrix and Parallel Processing in ATM Communication of type AAL1,” in *Proc. IEEE BSS International Workshop on Broadband Switching Systems*, April 19-21, 1995, Poznan.

[T] **Technical Reports**

[T1] A. Valkó, “Simulation Studies of Cellular IP Networks,” *Technical Report*, New York, November 1998.