Bayerische Julius-Maximilians-Universität Würzburg
Institut für Informatik
Lehrstuhl für Verteilte Systeme
Prof. Dr. P. Tran-Gia

Models and Algorithms for Demand-oriented Planning of Telecommunication Systems

Kurt Tutschku

Würzburger Beiträge zur
Leistungsbewertung Verteilter Systeme

Bericht 2/99

Würzburger Beiträge zur Leistungsbewertung Verteilter Systeme

Herausgeber

Prof. Dr. P. Tran-Gia
Universität Würzburg
Institut für Informatik
Lehrstuhl für Verteilte Systeme
Am Hubland
D-97074 Würzburg

Tel.: +49-931-888-5510
Fax.: +49-931-888-4601
email: trangia@informatik.uni-wuerzburg.de

Satz

Reproduktionsfähige Vorlage vom Autor.
Gesetzt in LaTeX Computer Modern 9pt.

ISSN 1432 – 8801

Models and Algorithms for Demand-oriented Planning of Telecommunication Systems

Dissertation zur Erlangung des
naturwissenschaftlichen Doktorgrades
der Bayerischen Julius–Maximilians–Universität Würzburg

vorgelegt von

Kurt Tutschku

aus

Würzburg

Würzburg 1999

Eingereicht am: 20.5.1999
bei der Fakultät für Mathematik und Informatik
1. Gutachter: Prof. Dr.-Ing. P. Tran-Gia
2. Gutachter: Prof. Dr. R. Mathar
Tag der mündlichen Prüfung: 21.7.1999

Dank

An erster Stelle gilt der Dank dem Betreuer dieser Arbeit, Herrn Prof. Dr.-Ing. P. Tran-Gia, für die vielen wissenschaftlichen Diskussionen und Impulse, die meine Arbeit fachlich stets sehr befruchtet haben. Besonders möchte ich sein stetiges Engagement auf nationaler und internationaler Ebene hervorheben, das dazu beigetragen hat, daß diese Arbeit wachsen und reifen konnte. Bedanken möchte ich auch bei Herrn Prof. Tran-Gia für sein großes Vertrauen und Verständnis, das er mir die Jahre hindurch, die ich an seinem Lehrstuhl verbringen durfte, entgegen gebracht hat.

Herrn Professor Dr. R. Mathar danke ich für die Übernahme des Zweitgutachtens und die vertrauensvolle und aufbauende Zusammenarbeit im Laufe der vergangenen Jahre.

I would like to thank Dr. K. Basu and Dr. Sairam Subramanian of Nortel Networks, Richardson Tx., for their great support in developing the set cover algorithms.

Meinen Kollegen am Lehrstuhl für Verteilte Systeme, die mich im Laufe der Jahre begleitet haben, gilt ebenfalls ein besonderer Dank: Mathias Dümmler, Dr. Thomas Fritsch, Dr. Notker Gerlich, Frank Heister, Stefan Köhler, Kenji Leibnitz, Dr. Rainer Müller, Dr. Michael Ritter, Dr. Oliver Rose, Dr. Alexander Schömig, Dirk Staehle, Dr. Thomas Stock, Norbert Vicari und Patricia Wilcox. Durch die vielen angeregenden wissenschaftlichen Diskussionen und nicht so sehr wissenschaftlichen Espresso-Runden, hat meine Arbeit sehr viel gewonnen.

Ein weiterer Dank geht an Dr. Michel Mandjes, Dr. Thomas Niessen, Steffen Reith und Michael Wolfrath für die wertvollen fachlichen Diskussionen die ich mit ihnen führen durfte.

Den Studenten Markus Greger, Marius Heuler, Titus Leskien, Peter Liebler, Dirk Schäfer, Uwe Schäfer, und Christian Schloter, die mir mit Diplomarbeiten oder als wissenschaftliche Hilfskräfte zuarbeiteten danke ich für die unterstützende Zusammenarbeit.

Bei meinen Freunden, die mir gezeigt haben das es auch noch außer-

halb der Universität und der Studientätigkeit ein beeindruckendes Universum voller Leben gibt, möchte ich mich von ganzen Herzen bedanken.

Zum Schluß geht mein ganz besonderer Dank geht an meine Eltern, Josef und Wilma Tutschku. Ihre unermüdliche, aufopfernde und liebevolle Unterstützung während der Studien- und Promotionsjahre gaben mir immer wieder Kraft und Ausdauer an dem gesteckten Ziel festzuhalten. Danke.

Contents

Part I

Planning Concepts for Telecommunication Networks

1 Introduction

The continuous deregulation of the telecommunications market in North America and Europe has a significant impact on the design and deployment of new telecommunication networks. Before deregulation, public telecommunication systems were predominately run by government affiliated authorities. The main task of these operators was to provide basic and reliable communication services to all customers in their country. Thus, the main engineering objective was to establish connectivity of the users to the network. The service cost only was of secondary importance since most network operators were monopolists and their users had to accept the imposed prices. On the one hand, this resulted in unjustified high tariffs for telecommunication services and, on the other hand, in less efficient network configurations.

The deregulation invited and provoked competition into the telecommunication market. Commercial companies are now allowed to apply for network licenses and thus to offer comparable services to customers. They can now choose between old and new operators according to their own requirements. Due to the competition in the market, supply and demand are determining the price of the service. Hence, the tariffs began to fall while the networks became more efficient. Moreover, the radical decline of the prices stimulated the demand for telecommunication services. Thus, today's networks are required to be even more efficient.

At the same time deregulation measures were introduced, the global exchange of commercial and personal information gained more and more importance. In particular, new telecommunication services such as video, multimedia, data, and personal communication services are becoming increasingly popular towards the end of the nineties. Again, the high popularity again increases the demand for these services.

In addition, the past years have seen the emergence of new information networking environments comprising new technologies such as *integrated services digital network (ISDN)*, *broadband ISDN*, the *In-*

ternet, or *wireless communication systems*. Some of the new techniques severely challenge the old telecommunication network design paradigms of separated planning procedures for architectural design and teletraffic engineering. For instance, the coverage area of a *code division multiple access (CDMA)* transmitter depends on both the radio wave propagation and user density, cf. Veeravalli et al. (1997).

The deregulation process, the commercialization of the telecommunication and the recent technological innovations led to a fundamental change in the network design paradigms. The focus of engineering has moved away from the pure technical capability of the network and is now mainly emphasizing *efficiency* objectives. In order to fulfill these efficiency objectives, new *models and algorithms for demand-oriented planning of telecommunications systems* are required.

This monograph is organized into two parts. The first part introduces the basic planning concepts for telecommunication systems. Chapter 2 outlines some fundamental functional models for wireline and wireless telecommunication systems. Chapter 3 provides an overview and attempts a classification of the conventional and new design objectives for state-of-the-art wireline and wireless telecommunications networks. Chapter 4 outlines the conventional engineering methods for communication networks and introduces a new forward engineering and integrated planning concept for fast, efficient and demand-oriented cellular system design. The second part of this monograph deals with the implementation of demand-oriented design methods. Chapter 5 is devoted to a new framework for the estimation and characterization of the expected teletraffic in mobile communication networks. Chapter 6 introduces algorithms for demand-oriented radio network design. Chapter 7 discusses the engineering of efficient call handling mechanisms in cellular mobile communication systems. Chapter 8 presents a method for *available bit rate (ABR)* service engineering in large scale *asynchronous transfer mode (ATM)* networks. Finally, Chapter 9 summarizes this monograph and provides an outlook into the future.

2 Functional Models of Telecommunication Networks

The purpose of this chapter is to give an overview on the basic terms and components of telecommunication networks. It introduces some fundamental functional models for wireline and wireless telecommunication networks and outlines the basic planning and design tasks in these systems. Particular focus is laid on their use in demand-oriented network design.

The chapter is organized as follows. Section 2.1 outlines general network model which is common to the most public telecommunication networks. Section 2.2 introduces the specific features of wireline networks and Section 2.3 is devoted to the characteristics of public mobile and wireline telecommunication systems.

2.1 Basic Telecommunication Network Model

A communication network is a spatially distributed arrangement of hardware and software that allows users to exchange information. It consists of a set of nodes that are interconnected to permit the exchange of information. These nodes are distinguished into *terminal nodes* and *communication nodes*. A *connection* commences and terminates at terminal nodes which also provide the interface to the user. The task of communication nodes is to establish *connectivity* among nodes and to *relay* information. The communication nodes are interconnected by *links* or

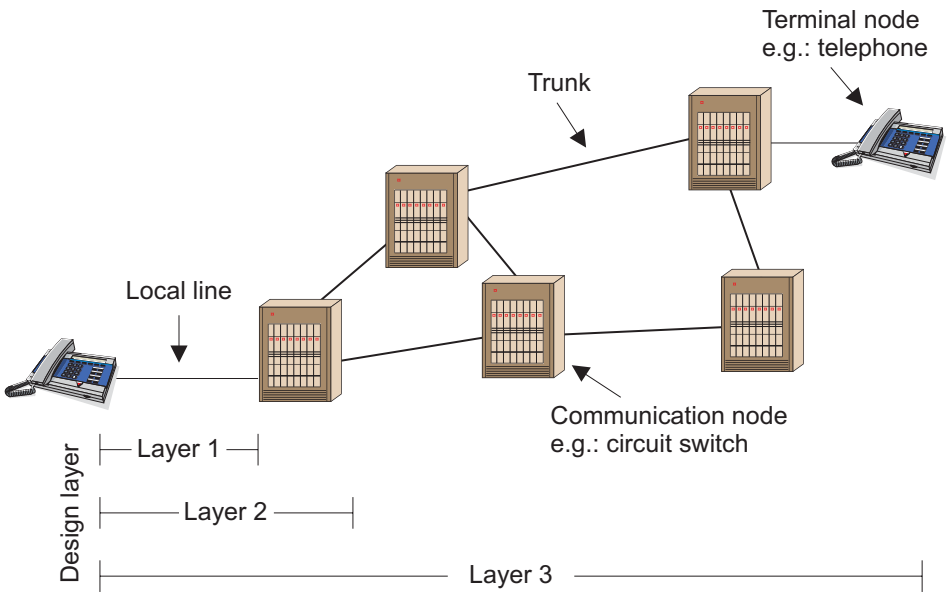


Figure 2.1: *Basic communication network*

trunks which are capable of transmitting information.

A simplified example of the most familiar and ubiquitous communication system, the public telephone network, is depicted in Figure 2.1. The terminal nodes, i.e. the telephones, are connected via local subscriber lines to communication nodes, i.e. circuit switches. The communications nodes are linked by trunks of larger capacity.

The exchange of information in a communication network is facilitated by offering *services*. A service is the provision of a transport mechanism or a function for relaying information. In modern communication systems, complex services are build from basic ones. Every service comprises the execution of distributed service scripts, e.g. protocols, by the entities involved in the communication. The concept of assembling high level services from low level service entities is inducing a *layered structure* for communication networks. A widely accepted layered framework for communication networks, for example, is the *Open System Interconnection reference model (OSI)* proposed by the *International Standard Organization (ISO)*, cf. Day and Zimmermann (1983).

The decomposition of services into smaller entities with specific actions and interaction has three main objectives. First, it permits the *reuse* of service entities, e.g. voice information and data information can use the same low level transport mechanism. Second, it allows the transparent *exchange* of service entities, e.g. for a seamless upgrade to more

efficient technologies. And third, accurately defined services facilitate the *interoperability* across manufactures and, later on, between networks of other *communication service providers*.

An additional feature of service decomposition is that it also permits a layered approach to system engineering and network planning. The design layers are directly related to the functional tasks of the main components of the communication system:

Layer 1: Transmission mechanism design
The design of low level procedures for propagating the information through a physical communication channel, e.g. the local subscriber line or a radio communication link.

Layer 2: Node engineering and dimensioning
The design of the functional features in a single communication node and the appropriate selection of the number of functional components in the node which is required for efficient operations.

Layer 3: Network design
The assessment of the location of communication nodes, their connectivity, and the routing of the information through the network.

Besides the structural description of the network, the functional model also has to consider the teletraffic in the system. The term "teletraffic" denotes the process of events related to demands for the utilization of services or resources in a communication network, cf. ITU-T (1993a). In general, the teletraffic in a network is not constant or homogeneous. Hence, in the context of network design it is necessary to distinguish between *temporal variability* and the *spatial variability* of telecommunication traffic. The temporal variability characterizes the sequence of service requests and the change of the intensity over time, whereas the spatial variability describes the variation of teletraffic intensity due to the location in the network. In particular, the design of large geographically distributed communication networks has to carefully consider the spatial variation of the demand for teletraffic, cf. Chapter 5.

The variability of the teletraffic has a different impact on the three design layers. Transmission mechanisms engineering is only touched in a limited way by time varying traffic. A well designed mechanism has to ensure operability, regardless whether there is high or low traffic intensity. In contrast to this, the influence of temporal variability on node

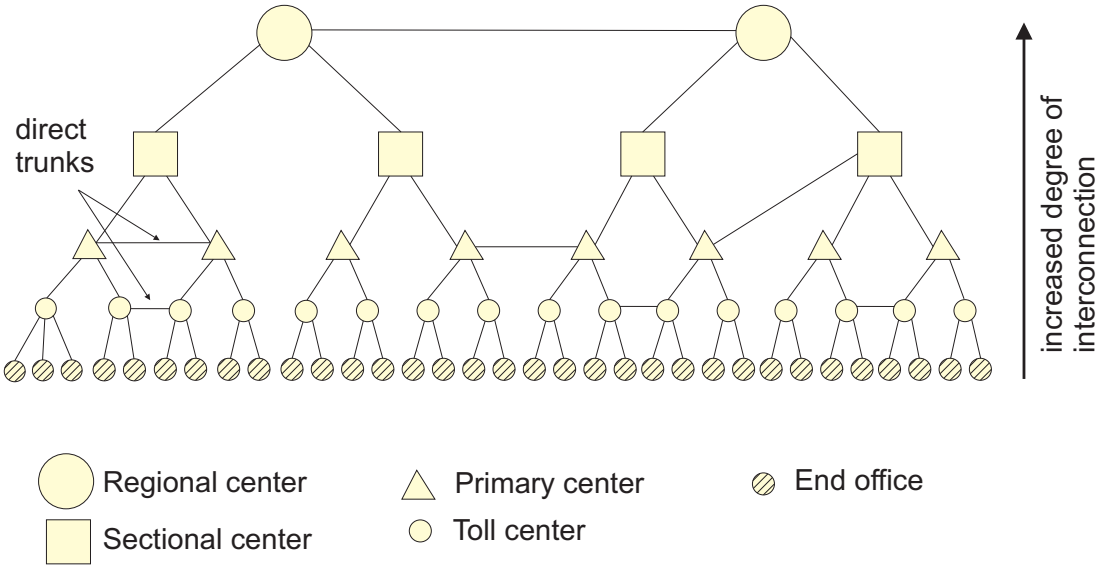


Figure 2.2: *Public wireline telecommunication network*

dimensioning is considerably high. A system can operate sufficiently well for constant traffic, however, its performance can degrade extremely fast for variable traffic, cf. Section 5.4. The impact of the spatial variation of the teletraffic increases with the scope of the design layer. Hence, efficient and demand-oriented design of communication networks has to consider both types of variability in order to obtain good network design solutions.

2.2 Public Wireline Telecommunication Networks

The functional model for public wireline telecommunication systems reflects the basic requirements for these systems. First, these network need to establish the connectivity of users to the system throughout large service areas. And second, they have to provide the communication service, e.g. voice service, to a huge number of customers. Therefore, public wireline communication systems posses a distinct hierarchical structure. In this way, they can efficiently cover large areas and are able to process a huge amount of teletraffic in the system.

Figure 2.2 shows a functional model of large public telephone networks used in North America, cf. Martin (1990). Subscribers are con-

nected on local loops to their nearby telephone exchange, called *local exchange* or *end office*. Local exchanges are the *access facilities* for customers. Interconnecting every end office with every other one would require too many trunks. Therefore, to lessen the number of trunks, further levels of switching are used. Hence, the end offices are connected to *toll centers*, which are further linked to primary centers. The upper part of the switching hierarchy is represented by sectional and regional centers. They establish country-wide and international connectivity. With every switching level, the connectivity between the centers and capacity of the trunks increases. However, the hierarchical structure has not always to be kept. Direct high capacity trunks are able to interconnect centers with high direct traffic. Thus, reducing the number of *hops* required to reach the final destination.

The amount of telecommunication traffic in public wireline networks is characterized by the *teletraffic matrix*. An entry in the matrix describes the traffic from the originating to the terminating node. Due to the tether-bounded access to the network, the spatial traffic distribution in a wireline system only changes only slowly over time. However, the temporal variability of the traffic can be high in these networks.

The planning of public wireline networks comprises three main engineering tasks: *a)* the assessment of the structure, the location and the interconnection of the nodes, *b)* the determination of the routing of the traffic and *c)* the dimensioning of the trunks.

2.3 Public Mobile and Wireless Communication Networks

The main purpose of modern public mobile communication networks is the provision of tetherless communication services at any place and any time. Therefore mobile communication systems consist of two main components: *a)* the *radio network* and *b)* the wireline transport subsystem.

The radio network provides the wireless access to the mobile system using a low power *radio communication link* between a *mobile terminal*, also denoted as a *mobile station*, and a grid of *base stations*. The transport subsystem is responsible for relaying the communication service through a conventional wireline network to its final destination. The endpoint of a connection can either be a customer in a wireline system or another wireless subscriber. The mobile communication system should

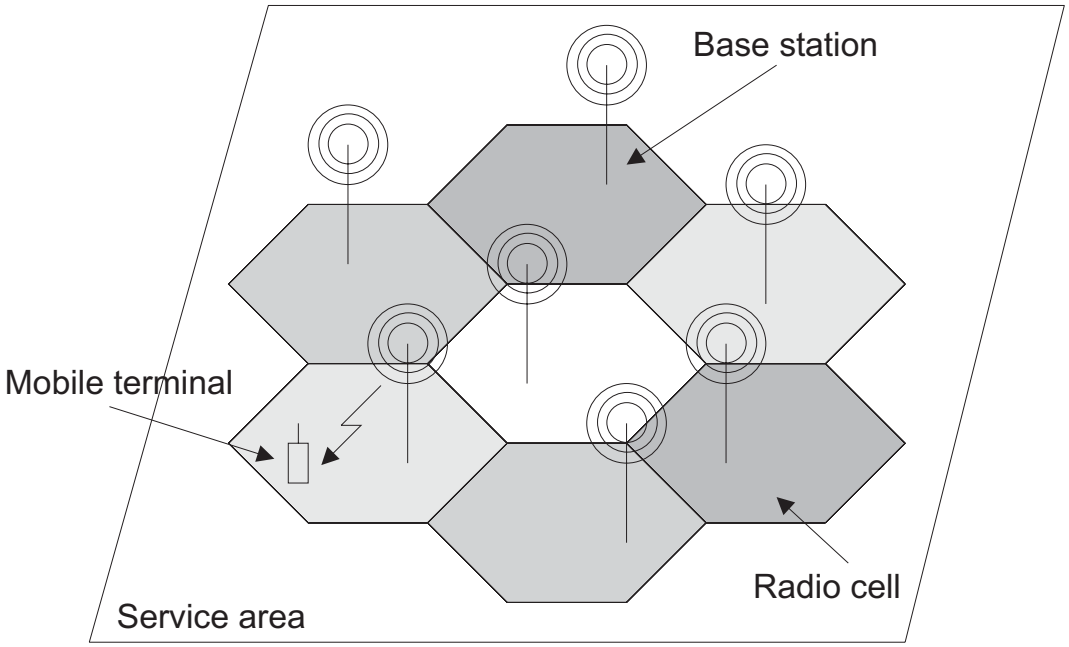


Figure 2.3: *Cellular concept*

permit a *transparent service provision*. A user should not notice whether the service is provided by a wireless or wireline network.

Unlike traditional wireline telephone systems, public mobile communication networks allow for *subscriber mobility*. The customers are permitted to roam around in the service area. Modern mobile networks are capable of autonomously locating and tracking the users.

Due to the limitation of available radio frequencies for public communication purposes, capacity problems occurred early in mobile communication systems. This deficit led to the development of the *cellular concept*, cf. MacDonald (1979). The basic idea of the concept is the partitioning of the service area into radio cells, typically 1 to 20km across, cf. Figure 2.3. A capacity increase is achieved by *reusing* the frequencies in geographically separated cells, indicated in Figure 2.3 by different grey scales. Since radio cells, under ideal conditions, have almost circular shape, they are commonly approximated by hexagons.

Both, the cellular concept and the subscriber mobility increase the complexity of mobile communication networks over the one of their wireline counterparts. Since transparent service provision is mandatory in mobile systems, on-going connections have to be maintained during the crossing of cell boundaries. Therefore, special *handover* mechanisms be-

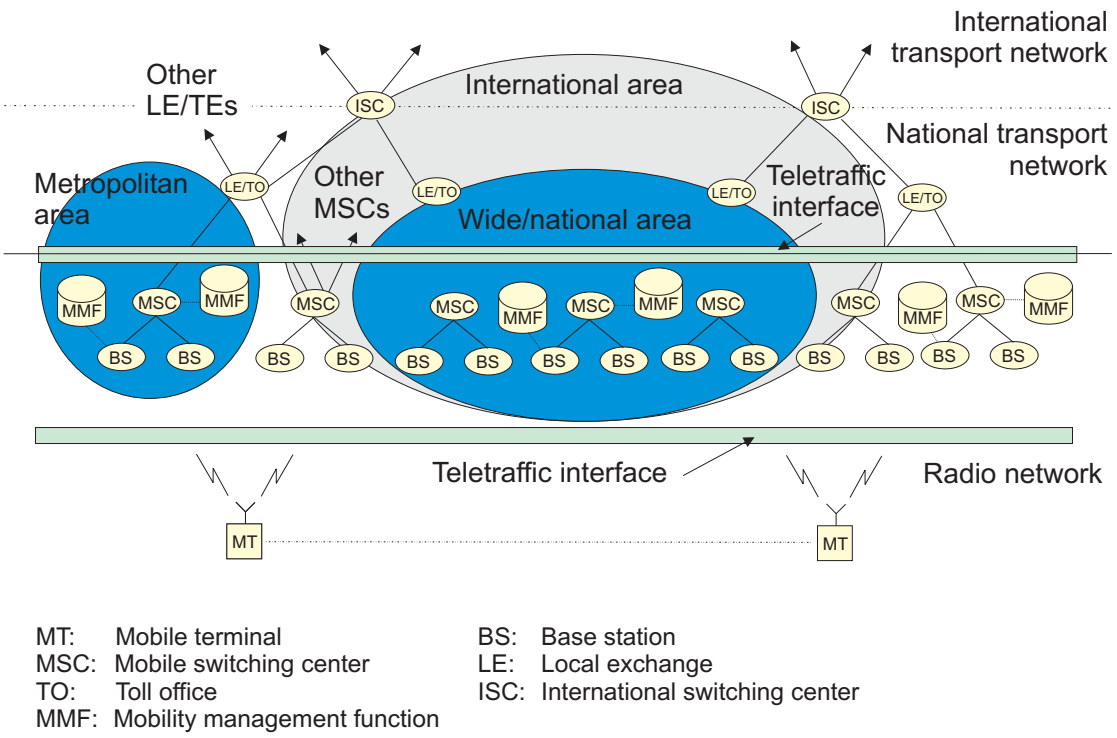


Figure 2.4: *Cellular mobile network, cf. Grillo (1998) et al.*

tween the cells, cf. Chapter 7, and functional entities executing these procedures are needed in the network. Furthermore, designated *mobility management facilities (MMF)* are required for locating and tracking the users in the system.

A basic architectural model of a cellular mobile network is shown in Figure 2.4. The lower part of Figure 2.4 depicts the radio network subsystem. The transport subsystem is shown in the upper part of the figure. In the radio network, the *base stations (BS)* are interconnected with *mobile switching centers (MSC)*. The main task of the MSCs is to relay the connection to the transport network. Additionally, they perform management and control tasks for the base stations. The mobility management facilities are associated either with the base stations or with the mobile switching centers. The MMFs mainly consist of a database which keeps track of the user whereabouts or their last location. Additionally, the MMF maintains information on the users' identity and other information which is vital for accounting and service provision.

The structure of the transport subsystem is similar to the one of common wireline communications systems. The mobile switching centers

are linked by high capacity trunks with local exchange centers or other high level switching centers. Additionally, the local exchange centers can act as *gateways* to other public telephone networks.

A mobile communication network comprises two main teletraffic interfaces, cf. Figure 2.4. The first one is located within the radio network between the mobile terminals and the base station. The second teletraffic interface is located on the boundary between radio network and transport subsystem. Both interfaces have to be cautiously addressed during network dimensioning.

Due to the user mobility, the teletraffic in mobile cellular communication systems reveals both, a high temporal and a large spatial variation. The users are allowed to enter or leave the cells whenever they want. Thus, the offered teletraffic in a cell can rapidly change over time, cf. Chapter 7. Furthermore, since not every part of the service area of a cellular system is equally populated or used in the same way, the spatial teletraffic intensity varies significantly, cf. Chapter 5. The high teletraffic variability is a characteristic which makes mobile communication extremely dependable on careful traffic engineering.

Conventional mobile communication network engineering is separated into two main design tasks: *a)* radio network planning and *b)* transport network engineering. Radio network planning mainly comprises the assessment of both the location and configuration of base stations, the engineering of radio link parameters, the design of the frequency reuse plan and the dimensioning of the teletraffic interface between the mobile terminals and the base stations. Engineering tasks for the transport subsystem are similar to the ones of common wireline networks, cf. Section 2.2. Additionally, special design problems which arise from the distinct features of mobile systems have to be addressed. The most important tasks comprise: design of *location areas*, cf. Madhavapeddy and Basu (1994), interconnection of base stations with mobile switching centers, and dimensioning of the traffic interface between the radio network and transport system.

3 Network Design Objectives and Requirements

The planning of future telecommunication networks faces three new major challenges. First, there is the tremendous increase in the demand for communication services. Two decades ago, only a few telecommunication services, like telephone, fax, or low speed data service, were required and available. In the future a large variety of services are needed to support various demands of new communication applications, like *voice-over-IP*, *electronic commerce*, or *business process management*. Second, the new technologies of upcoming networks require demand based planning methods. For example, in third generation mobile networks, the coverage area of a CDMA transmitter depends on both the radio wave propagation and the user density, cf. Veeravalli et al. (1997). And third, due to the deregulation of the telecommunication market, the competition between the service operators is highly increased. The regulation authorities require the services to be interchangeable. Hence, the customers can switch almost instantaneously to the cheapest provider. As a result of the challenges, new systematic planning methodologies are required which facilitate the engineering of effective, economic and optimal network configurations. To enable a systematic design approach, the network design objectives and requirements have to be precisely stated. Unfortunately, there exists a large number of technical and economical objectives. Moreover these design requirements are often contrary to each other. The purpose of this chapter is to give an overview of the conventional and new design objectives for state-of-the-art wireline and wireless telecommunications networks. In addition, it attempts to classify the objectives.

3.1 General Network Design Objectives

Telecommunication networks are large scale engineering objects. They consist of numerous technical entities and represent high financial investments. Thus, these systems require a purpose-oriented planning using precisely stated design objectives. In general, the design objectives and requirements for telecommunication networks can be separated into three categories: *a)* effectiveness, *b)* efficiency and *c)* responsiveness.

Effectiveness

The effectiveness is traditionally of prime importance for telecommunication network. This category comprises in first place the *technical capability* objective of the system, which describes the technical features of the network and of the service provision; for example: the transmission line specifications, the communication protocols, or the network structure. The technical capability objective guarantees the *connectivity* feature of networks, i.e. the user is able to connect to the network and to use the service.

Another design objective in the effectiveness category is the *compatibility* of the system with networks and hardware of other operators or manufacturers. This requirement is strongly related to the previous one. Usually, a communication network is not an isolated homogenous system. A network may consist of interconnected subsystems operated by different companies. The interoperability guarantees the service provision over network boundaries to a large number of customers. Furthermore, network elements have to be interchangeable and thus allowing an improvement of the system when new technologies are available.

The provision of a *correct service* is another major effectiveness requirement. Telecommunication networks always have to execute the specified actions for a certain service. For example, the signaling protocol for control information in a telephone network might be implemented correctly, however, the system does not perform the proper actions.

From the operators point of view the requirement of *fault tolerance* is also a critical design objective. A network must be able to cope with the failure of communication nodes or lines. Preferably, this has to be accomplished in an autonomous way by the system.

Another crucial design objective is the *quality of service (QoS)*. This design parameter denotes the "collective effect of service performances which determine the degree of satisfaction of a user of the service", cf.

ITU-T (1993b). The quality of service is characterized by the combined aspects of service availability, service reliability, service support performance, service integrity and other factors specific to each service. However, the term "quality of service" is not intended to express a degree of excellence in a comparative sense and should not be used in a quantitative sense for technical evaluations. In these cases a qualifying adjective has to specified.

Another decisive engineering objective, unfortunately often neglected during network design, is the *customer service*. The general customer perception of the helpfulness of the network operator staff and their capability to solve user's problems is important for the commercial success of the system. Thus, the designer should include mechanisms into the network such that the operating staff can trace and fix the problems. Furthermore, customer service accounts for the fair billing of the user for the service.

Efficiency

The second category of design objectives are efficiency requirements. This classes separates into four main engineering aims: *a)* economical efficiency, *b)* resource consumption, *c)* teletraffic capacity and *d)* service cost.

Since a telecommunication network represents a huge monetary asset, the *economical efficiency* of the system is a key criterion for its owner. An adequate financial performance, e.g. sufficient return on capital investment, is crucial for commercial survival.

The *resource consumption* is directly related to the economical efficiency. In order to decrease the cost of a telecommunication system, the use of valuable resources should be reduced. For example, the number of technical objects in the network, like communication nodes, has to be minimized. In addition, rare environmental resources have to be preserved, e.g. the frequency spectrum available for mobile radio communication systems is limited by nature. During network operation, the resources required for network management and maintenance have to be minimized.

An optimal *teletraffic capacity* is another key efficiency objective. In a communication network, the transport of information generates the profit from the system. Hence, the network should be capable to carry all or most of the expected traffic. The cost of operating the system, however, must be covered by the revenue obtained from the service.

The context of network efficiency requires also to consider the *service cost*. From the customer perspective, there must be an appropriate relation between the expense for a service and the benefit of using it.

Responsiveness

The third category of design objectives for telecommunication networks is concerned with their *responsiveness* to customer behaviour and technology progress. Network deployment has to be performed with respect to requirements. Strategies where and when to install equipment, denoted as roll-out plans, have to be developed during system design. In addition, an adequate network design should allow for network evolution. The design should permit system upgrades within a given budget. This step assures that the inventory asset maintains its value over long periods of time. Finally, the network engineering has to prepare the system for *predictive planning*. The extension of the system should be possible with respect to the expected growth of traffic in the system.

Viewpoints

The above mentioned design objectives and requirements can be considered in two different ways. One viewpoint is the perspective of the customer, the other point of view is the context of the network operator. The weight of the design objectives varies with respect to the perspective. A customer likes to obtain as much benefit from the network as possible, whereas the service provider has to consider always about the profit he is receiving from operating the system. A good network design should result in a *win-win situation* for each of the involved parties.

3.2 Wireline Networks

Besides of the general network engineering objectives, the design requirements for wireline telecommunication systems have to address the specific features of these networks, cf. Chapter 2. Hence, the main design objectives for wireline systems comprise four areas: *a)* structural requirements, *b)* performance and reliability objectives, *c)* interoperability requirements, and *d)* network deployment and system life cycle objectives.

The *structural requirements* mainly address the architectural organization of the network and the traffic routing within the system. The

access points to a network should be located in the neighborhood of the potential customers. The interconnections of communication nodes have to be determined such that every object can efficiently exchange information with every other node. A common concept for addressing the large size of public communication networks is to use a hierarchical structure. In this way, the number of interconnections can be minimized while maintaining the connectivity. Furthermore, this concept facilitates the transport of large amount of teletraffic. Local traffic can be kept within the proximity of its origin and only long distance connections need the allocation of valuable resources in the higher layers. After defining the network structure, the traffic routing has to be determined with respect to efficiency and reliability objectives.

The *performance and reliability requirements* comprise mainly service specific parameters, e.g. connection blocking probabilities, bit error rate, packet loss probability, allocated bandwidth, and information transfer delay. But also network-related criteria like down times and fault tolerance have to be addressed in this category. Since wireline networks are expected to be available at every time, very strict requirements are usually imposed on the performance and reliability values.

The *interoperability* requirements are related to the expected inhomogenous structure of future wireline telecommunications system. These objective comprise well defined interfaces among the hardware components and subnetworks as well as comprehensive testing procedures, e.g. the ITU-T's Q.780 series for testing the signalling in ISDN networks, cf. ITU-T (1993c). In addition, the regulation authorities for the telecommunication market require extensive validation procedures in order to guarantee the collaboration of networks.

The *network deployment and life cycle* objective are concerned with the time intervals of installing and operating a network. The deployment of wire-bound transmission links is expensive and time-consuming. Very often, the cables have to be buried under roads or new telephone poles have to be build. Hence, the financial investment into the *passive* part of the network might be larger than the value of the *active* one. The network engineering has to address this feature carefully by designing these passive components for a long life cycle. For example, fiber optic cables should be selected such that they can be reused for other optical transmission techniques. Another possibility is to over-dimension the basic transmission capabilities, e.g. by installing more physical links than necessary.

3.3 New Design Objectives of Mobile and Wireless Networks

Due to the radical changes in technology and usage, the design criteria of next generation cellular networks have altered substantially. Beside the primary RF (Radio Frequency) objective of providing a reliable radio link at every location in the planning region, state-of-the-art network design has to ensure a high quality-of-service as well as considering the aspects of cutting the cost of deploying a cellular radio system. Hence, the new design objectives of mobile networks can be organized in three areas: *a)* RF objectives, *b)* capacity and teletraffic engineering objectives, and *c)* network deployment objectives.

The *RF design objectives* are usually expressed in terms of radio link quality measures. In first place, a good link design has to ensure a sufficient radio signal level throughout the planning region. Additionally, it has to minimize the signal disturbance by co-channel and adjacent-channel interference. Related to that is the provision of a high margin between the signal level and the interference power level in order to support, for example, a high user mobility. Advanced RF performance values, like a low bit error rate (BER), or a low call blocking probability due to insufficient signal strength can be derived from the basic RF performance values.

The *capacity and teletraffic engineering objectives* comprise mainly three criteria: resource requirements, network capacity, and capacity related quality-of-service values seen by the user. Since the available frequency spectrum for mobile radio systems is extremely limited, the network design has to minimize the number of frequencies required in a cell as well as the overall number of carriers used in the network. To increase the total system capacity, the design has to enforce a large frequency reuse factor.

The *network deployment objectives* mainly address the economic aspects of operating and engineering a cellular system. Deploying a complete new network or installing additional hardware in an operating system is highly risky. There are significant costs associated with setting up a new facility. Therefore an efficient network design has to minimize the hardware cost, for example by using as few base stations as possible or by deploying cost-efficient facilities, like low-power transmitters. In addition, the cost-efficiency of the network can be increased if the deployment of new network equipment is based on the analysis of the

demand for the offered service.

4 Engineering Methods for Telecommunication Systems

The engineering and architecting of large telecommunication networks is supposed to be both science and art, cf. Rechtin and Maier (1997). This view is mainly the result of the high complexity of these systems. During the early stages of network engineering, a full assessment of the system design problems is difficult for the network planner. Some of the design challenges only arise due to the collaboration of the network components with each other and can hardly be anticipated. Other engineering problems are postponed to later design stages since their analysis would consume too much time. In particular, efficiency issues are often regarded of minor importance at early network deployment phases, since the teletraffic is usually low in new systems. perception of the quality of a network.

Conventional design methodologies for telecommunication networks address only partially the complexity of systems. In addition, they still require a lot of personal experience and manual interference by the network designer. Therefore, new design procedures are needed for the engineering of future telecommunication networks. The new methods have to focus early and equally on all of the major design objectives, cf. Chapter 3. In particular, they have to put more emphasis on teletraffic issues. Moreover, the methods should be simple in their application and be able to attain efficient network configurations in short time.

The purpose of this chapter is to provide an overview on conventional engineering methods for telecommunication networks. In addition, in this part, a new planning procedure will be presented for fast, efficient and demand-oriented cellular system design. The chapter is organized as follows. In Section 4.1, the conventional reverse engineering design process is reviewed using the example of cellular network design. Section 4.2

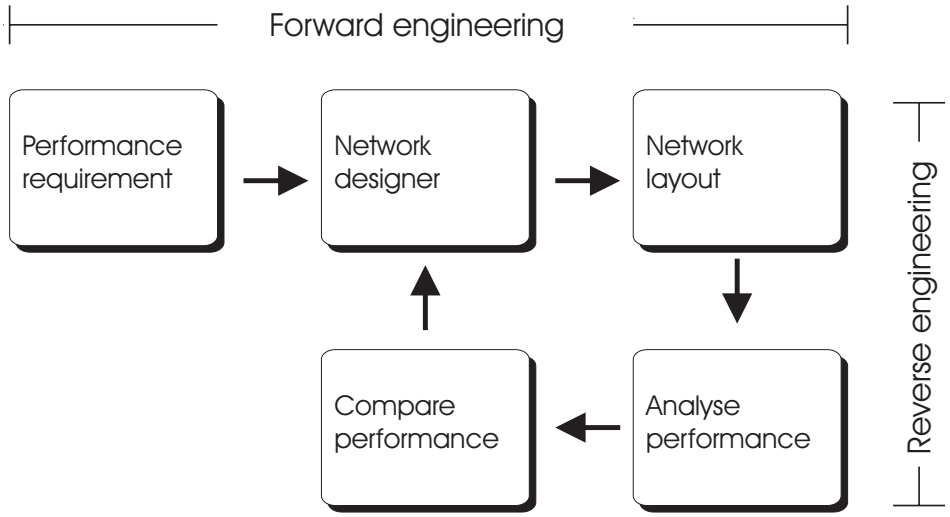


Figure 4.1: *Planning methodology*

introduces the *integrated approach* to cellular network planning. This method is a new forward engineering procedure for cellular design. It is able to overcome some of the drawbacks of the conventional approach. In addition, in Section 4.3 and Section 4.4, two specific network engineering procedures for capacity optimization and system deployment are described. The chapter is summarized by Section 4.5.

4.1 Reverse Engineering

The work of network planners is usually based on two engineering approaches. The majority of the planning decisions can be determined by applying rational and rule-based methods. For some of the planning decisions, however, the designers have to rely on their experience or on well known heuristics. Hence, the conventional design method of telecommunication networks is split into two phases, cf. Figure 4.1. In the first phase, an initial network layout is determined by using simple rules which are mostly based on heuristics. In the second phase, the system is iteratively improved until the specified design objectives and requirements are met. Since the network improvements are derived from the analysis of the current system configuration, this approach is denoted as *reverse engineering*.

In order to make the rough definition of reverse engineering more vivid, an example design procedure using this paradigm is outlined next.

The investigated example is the conventional design procedure for cellular communication networks. The example is particularly interesting since it is widely used in today's cellular system engineering but it is expected to fail in future system design, cf. Cheung et al. (1994).

Conventional Cellular Network Planning

The conventional design procedure for cellular systems is based on the so-called *analytical approach* to cellular network planning, cf. Gamst et al. (1986). This approach is mainly focused on the determination of the transmitter parameters, like transmitter location, antenna type, or transmitting power. It obeys the in Section 3.3 described RF objectives but neglects the capacity and the network design objectives during the engineering process. In addition, the analytical approach is the base for most of today's commercial cellular network planning tools like *PEGASOS*, cf. T-Mobil (1996), or *PLANET*, cf. MSI Plc. (1996).

In principle, the analytical approach consists of four phases, *Radio Network Definition*, *Propagation Analysis*, *Frequency Allocation*, and *Radio Network Analysis*, that are passed in several turns iteratively, cf. Figure 4.2.

During the *Radio Network Definition* phase, a human expert chooses the cell sites. In order to obtain a regular structure, usually the popular concept of distributing the transmitters on a hexagonal grid is used in this step.

Using these transmitter configurations, the *Propagation Analysis* of the area evaluates the radio coverage by field strength prediction methods. Here, stochastic channel models are applied. Usually, several field strength prediction methods are implemented but the planning tools offer little if any support in choosing the appropriate propagation model. If the planning expert decides that the coverage is not sufficient enough, new transmitter locations have to be chosen and the propagation has to be analyzed again.

The radio network capacity issues are addressed in the next step, the *Frequency Allocation*. At first, the teletraffic distribution within the planning region is derived based on rough estimates on the land use of the area. The distribution is then stored in a traffic matrix. In the next step, a hexagonal grid representing the cells, is superimposed on the planning region, cf. Figure 4.3. The different grey values in Figure 4.3 are representing the different entries in the teletraffic matrix. The traffic per hexagonal cell is determined by tallying the entries of the traffic matrix

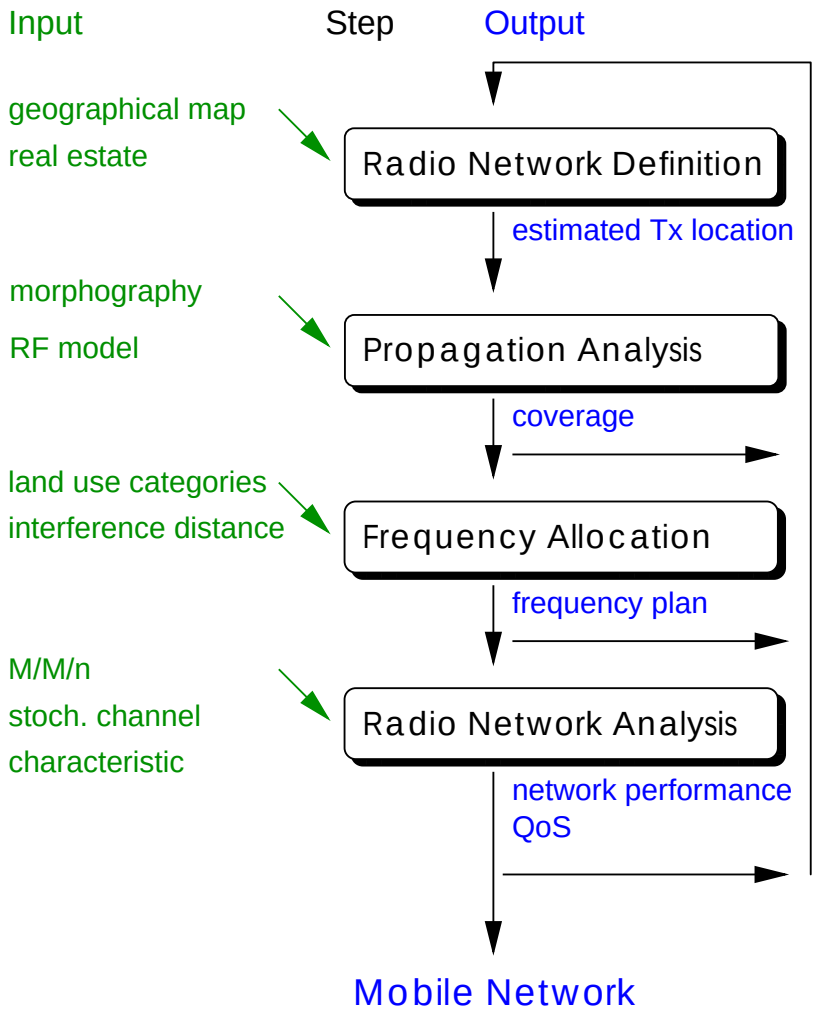


Figure 4.2: *Conventional approach of cellular network planning*

in each cell, cf. Faruque (1996). The required number of traffic channels and frequencies of a cell is computed by using land-line capacity planning techniques like the common Erlang-B-formulae, cf. Mouly and Pautet (1992). If, for a given frequency reuse pattern and for given interference distance constraints, all the cells of the area can be supplied with the required number of channels, the algorithm proceeds to Radio Network Analysis. Otherwise the algorithm starts all over again.

The *Radio Network Analysis* calculates the quality-of-service values of the area with regard to blocking and hand-over dropping probabilities. Again, stochastic channel characteristics as well as user demand estimates from the traffic data-base are used to calculate the network

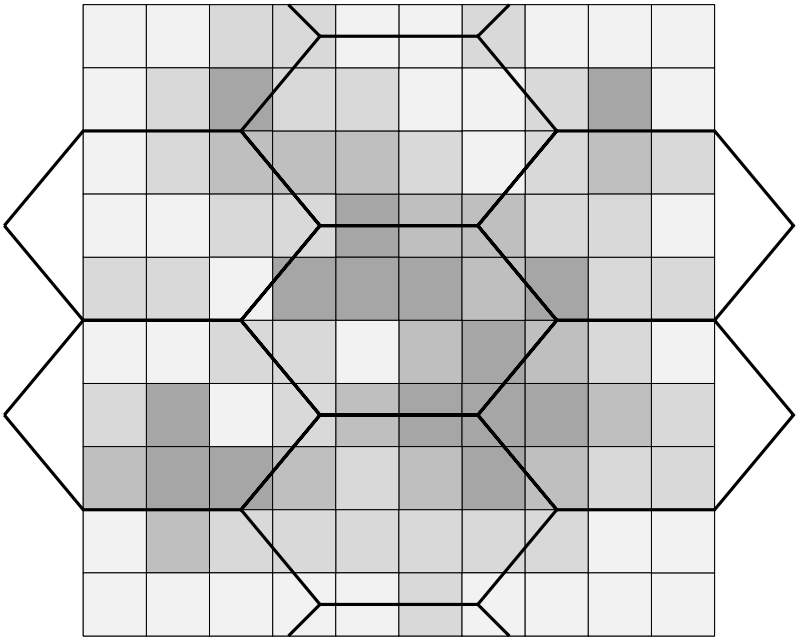


Figure 4.3: *Conventional hexagonal traffic count*

performance. If quality-of-service specifications are met, the task is accomplished, otherwise the algorithm has to be restarted.

The major disadvantage of the analytical approach is its restriction to the RF design objectives. Network and capacity issues are more or less neglected by the approach. In addition, the design steps are treated in isolation and trade-offs between the design objectives are hard to obtain. An overall optimization is not feasible. Moreover, the reverse reasoning technique prohibits the application of algorithmic optimization methods for the generation of synthetic networks. Additionally, it restricts the utilization of automatic network optimization methods due to its slow iterative process.

Resume

The advantage of reverse engineering procedures is their departure from a well-defined network configuration. The network designer is not required to explicitly model every interrelation between the system components. In this way, unknown relationships are very often implicitly included into the design process.

The major drawback of reverse engineering and of the iterative ap-

proach is the time between two improvements. This time can be rather long if field measurements or hardware modification are involved. In addition, an insufficiently selected initial network configuration can significantly slow down the network design process. Moreover, automatic network design and optimization procedures can only partially be employed.

To overcome these drawbacks, new efficient telecommunication network design methods have to be developed. The methods should possess three main features. First, they should eliminate the need of experience or heuristics rules. Second, they must be able to comprehensively address the complexity of the networks and of the design objectives. In particular, they should put more emphasis on efficiency and teletraffic issues. And third, the methods should attain a speed-up in network design by allowing the application of automatic optimization methods.

4.2 Forward Engineering

A promising approach to resolve the disadvantages of reverse engineering methods is the application of *forward engineering* procedures in the design of telecommunication networks. Forward engineering procedures derive a network layout based on high-level requirements specified by the system planner or by other decision makers, like marketing departments or regulation authorities, cf. Figure 4.1.

Without any refinements, forward engineering approaches are also very limited in their application. However, if the design objectives are well-defined, forward engineering becomes particularly efficient. Therefore, this design paradigm is mostly used when system planners have already sufficient experience in network engineering and when they were able to transfer the knowledge into general applicable rules.

In this section, the description of the forward engineering paradigm will also be deepened by using a cellular system engineering example. Therefore the so-called *integrated approach* to cellular network planning will be introduced next. The integrated approach will form the basis for a new demand-oriented design method for cellular systems which is one of the main contributions of the research work presented in this monograph.

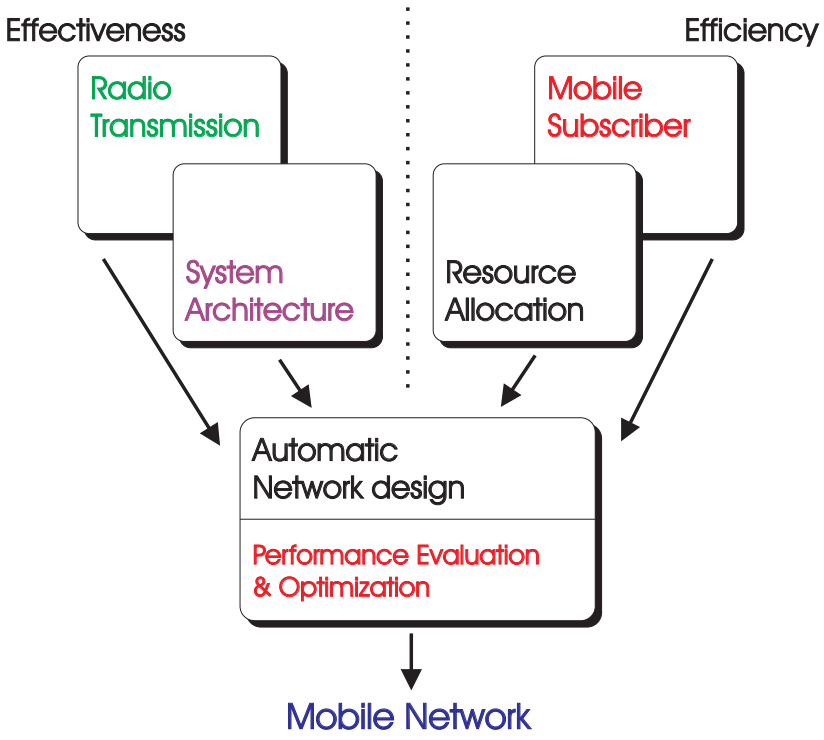


Figure 4.4: *Integrated planning approach*

Integrated Approach to Cellular Network Planning

The *integrated approach* to cellular network planning overcomes the shortcomings of the conventional approach by organizing the cellular design objectives and constraints in four basic modules. The new concept is depicted in Figure 4.4. The four basic design aspects of the integrated approach are: *Radio Transmission*, *Mobile Subscriber*, *System Architecture*, and *Resource Management*. Two of these basic modules are representing effectivesness requirements and two are addressing efficiency objectives, cf. Figure 4.4. The structured set of the input parameters to the modules is used by the concept for the *synthesis* of a cellular configuration. The network configuration is generated by the *Automatic Network Design* module.

Due to the equal and parallel contribution of all the basic modules to the network design, the integrated concept is able to obey the interactions and dependencies between the objectives. In particular, the capacity and efficiency objectives can be addressed early and in an appropriate way. Due to this characteristic, the integrated approach is therefore able

to find a trade-off between contrary objectives and achieves optimized network configurations. Moreover, the integrated approach constitutes a forward-engineering technique. Its methodology facilitates the application of automatic network design algorithms, cf. Chapter 6

Mobile User Characterization

In contrast to the conventional design method, where emphasis is mainly laid on RF issues, the new integrated approach considers the expected teletraffic of the service area as an equally contributing factor to the network planning. Moreover, the integrated method starts its network design sequence with an analysis of the expected teletraffic demand within the considered supplying area. The design sequence is described in detail in Section 6.4.3.

The core technique of the integrated approach is the representation of the spatial distribution of the demand for teletraffic by discrete points, denoted as *demand nodes*, cf. Definition 5.1. Demand nodes are widely used in economics for solving facility location problems, cf. Ghosh and McLafferty (1987). These demand nodes form the common basis of all components of the integrated approach. The application of the demand nodes, denoted as the *demand node concept (DNC)* leads to a discretization of the traffic demand in both space and amount. It constitutes a *static population model* for the description of the mobile subscriber density in the service area. A detailed description of the DNC is presented in Chapter 5. The demand node concept is very useful for the mobile subscriber characterization. Its additional feature is the transformation of the continuous transmitter location problem into an equivalent discrete optimization task. The application of optimization methods is facilitated by a new definition of the term supplying area, cf. Definition 6.1. Due to this definition supplying users with a mobile radio service is equivalent to *covering* demand nodes. An optimization algorithm has to determine the location of the transmitters such that the proportion of demand nodes within the permitted service range is maximized. Hence, the base station locating task is reduced to a *maximal covering location problem (MCLP)*, cf. Section 6.2.

Resource Allocation and Network Design

The demand node concept facilitates not only the mobile user characterization, it simplifies also the resource allocation task. Since the demand

nodes are distributed according to the expected service demand and due to the new definition of the term supplying area, cf. Definition 6.1, the expected traffic in a specific cell can be immediately obtained from the number of nodes in a cell. A potential base station site can be verified before it is selected, whether it obeys the traffic and hardware constraints or not, see Chapter 6. Thus, it is possible to check if a certain cell configuration is able to carry the expected traffic. Otherwise, the configuration is discarded and not considered for optimization. This verification can enforce, for example, the deployment of small and cheap transmitters over using large and heavily loaded macro-cells. Additionally, this leads directly to a higher frequency reuse and can result in a much easier frequency allocation task.

Resume

The disadvantage of the forward engineering design approach is its need of a well-defined system model and of a precise description of the relations between the design objectives. However, this drawback states implicitly the strength of the design paradigm. The advantage of the forward engineering method is its efficiency due to the high-level description of the system. The required simplification of the complexity leads to a concentration on the major design objectives. In addition, the formal description of the relations between the design objectives permits the weighing of the relations according to their importance. In this way, the forward engineering approach enables the application of optimization methods and can obtain a trade-off between contrary design objectives. Moreover, the forward engineering approach facilitates the implementation of network planning tools which are able to synthesis network configurations.

4.3 Capacity Optimization Cycle

For capacity optimization in an existing network it is important to have a good knowledge of the traffic in the system. The size of the traffic streams, their variation with the time of the day, week, and year, and their distribution in the network are important information that are necessary to plan and to optimize the system.

Therefore ITU-T has developed the recommendation E.490 for traffic measurement and evaluation, cf. ITU-T (1992a). An outline of the pro-

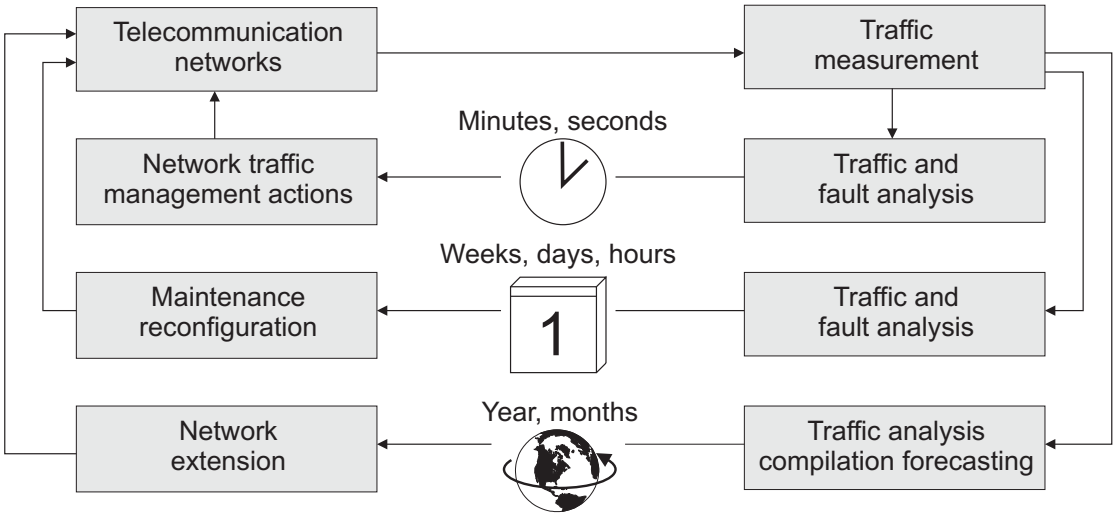


Figure 4.5: *Traffic measurement cycle, cf. ITU-T (1992a)*

posed procedure is depicted in Figure 4.5. The method comprises three interlocked loops. Each loop is defined by its time scale. The inner loop has a time scale of minutes or seconds. The traffic data should continuously be measured and almost in real time be analyzed and reported. Thus, the information can serve as the basis for traffic management actions such as temporary reroutings or temporary lease of additional transport capacity.

The second loop shows a time scale of weeks, days, or hours. There is also a strong suggestion to continuously monitor and analyze this information. In this way, faults disturbing traffic could be discovered and correlated, and appropriate maintenance or engineering actions could be taken.

The outer loop has the longest time perspective of years and months. These traffic measurements can be used to make traffic forecasts and serve as a basis for network extension and long-term network configuration. For these engineering applications a mean traffic intensity value, representing high traffic periods of the day and year, is generally used.

The step "traffic measurement" should be understood to comprise all kinds of recorded traffic parameters necessary for the activities in all loops. However, very often the same parameter could be used for all three loops. It is the reporting interval and analysis which differs.

After the acquisition of the traffic data, the information must be analyzed appropriately. A summary of a capacity optimization cycle suggested for cellular telecommunication networks is depicted Figure 4.6. At

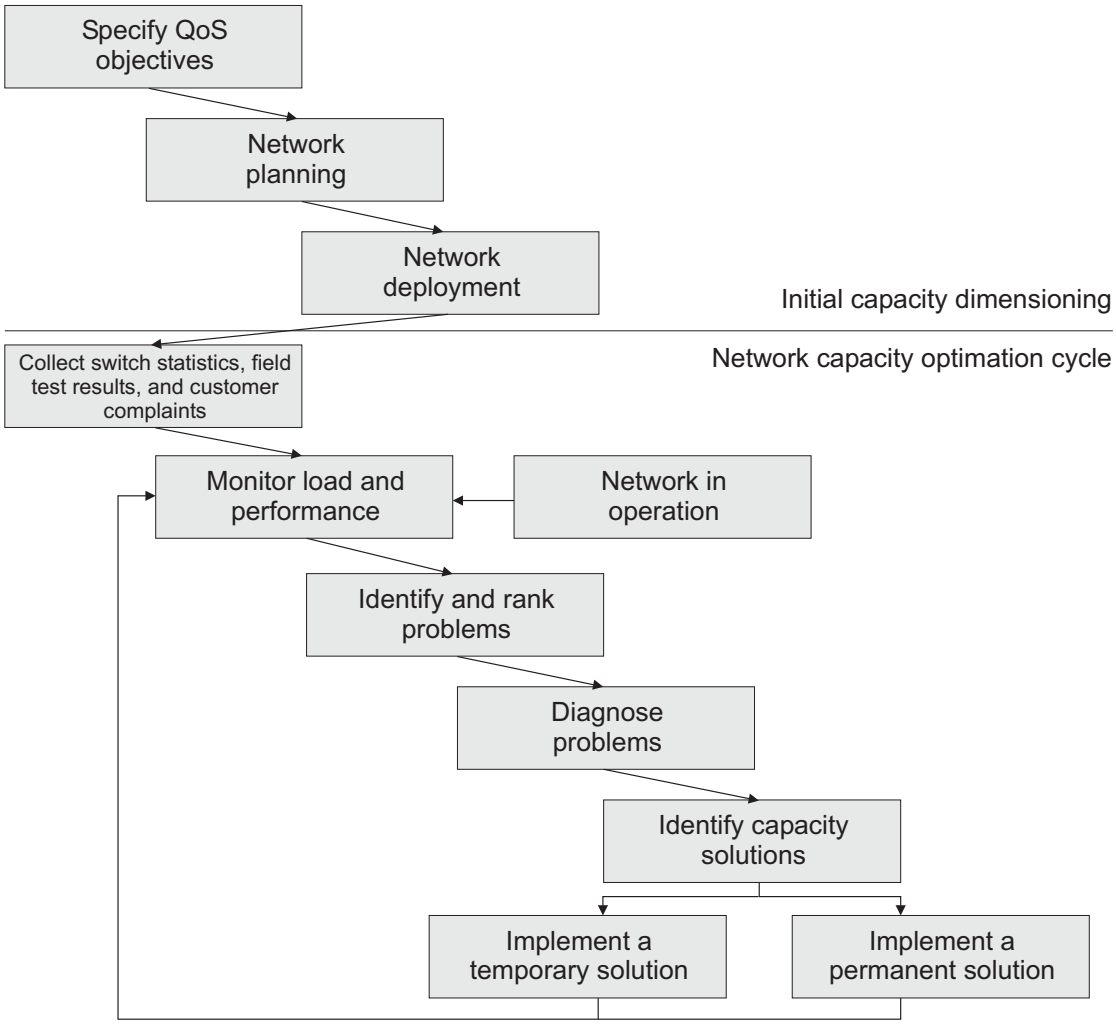


Figure 4.6: *Network capacity optimization cycle, cf. Grillo et.al. (1998)*

network startup, the quality-of-service objectives are defined by the network engineers and the traffic per subscriber is estimated. In the initial phase, this traffic estimation is used for network planning and deployment. During network operation, the operator closely monitors the traffic statistics of the busy hour traffic channel utilization for all the cells. This information can be continuously retrieved from the switch statistics. In case the level of call blocking reaches a predetermined threshold, action has to be taken to increase the traffic capacity, e.g. assigning additional frequencies to a cell. Of course, this action has to be done with regard to spectrum allocation and availability constraints. Once the maximum

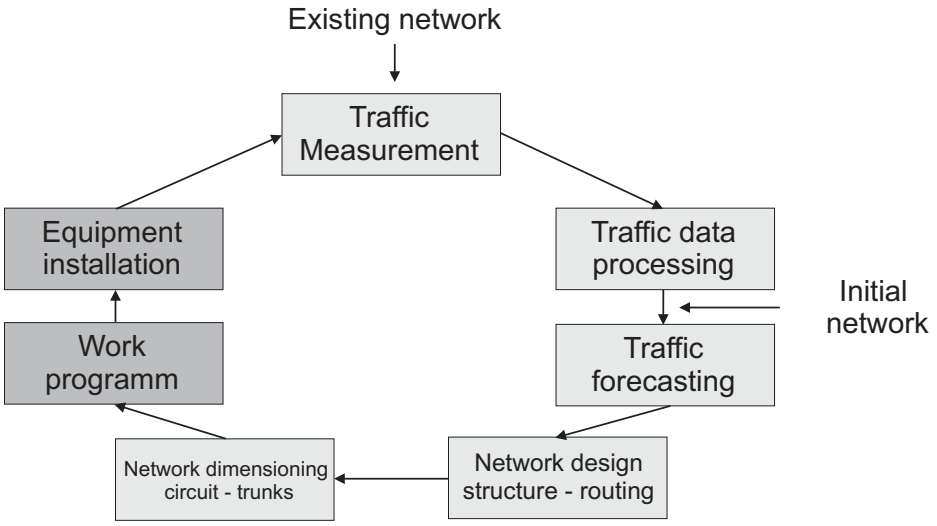


Figure 4.7: *Network installation cycle, cf. Harris (1998)*

number of frequencies for a specific reuse pattern is reached, more complex network modification have to be performed. For example, the installation of new equipment at appropriate base station sites.

The above described capacity optimization cycle also uses the reverse engineering concept. It is focused only on a single optimization criteria and requires experienced network planners in the initial phase. The optimization cycle is useful for improving an existing network. However, its application is limited due to the need of traffic measurements. For green field planning, other capacity design methods are needed.

4.4 Installation Cycle

A combination of the reverse and forward engineering approach was suggested by Harris (1998). The so-called "installation cycle" includes the advantages of the two basic design approaches. In addition, it comprises two extra planning phases: traffic forecasting and equipment installation. The cycle is mainly intended for the design and deployment of wireline networks. It is depicted in Figure 4.7. The cycle has two entry points. For an existing network, it starts with the acquisition of traffic measurements from the system. In the next step, the obtained data is filtered and compressed. The processed information together with the network layout forms the input to the traffic forecasting phase. In this step, the

future teletraffic is estimated using the traffic measurements and the assumptions on the anticipated use. For the estimation, procedures like the Kruithof method are applied, cf. ITU-T (1992b). The forecasting step serves also as an entry point into the cycle if only an initial network configuration without any traffic measurements is available. In this case, the measurements can be replaced by rough estimations. The anticipated traffic is the major input to the network design phase. In this step, the designers determine the structure of the network and the routing of the traffic. The next step in the cycle is the dimensioning of the communication links. After this item, the network layout is completely defined. In the next phase, therefore, the planners define the work program for the network deployment. In the installation step, the network equipment is finally deployed in the service area and brought into operation.

The installation cycle defines a practical and complete procedure network planning and deployment process. It combines the advantages of forward and reverse engineering. The most important feature of the cycle is the inclusion of a traffic forecasting step. In this way, the network can be designed to carry the anticipated traffic.

4.5 Concluding Remarks

The success of future telecommunication networks relies on intelligent network planning to achieve superior service quality, high capacity, and efficient network management. The conventional design methods can not completely accomplish the task since these methods are mostly based on pure reverse engineering design approaches. The conventional methods are either too slow or they are focused on isolated design aspects. The application of forward engineering design procedures can resolve this deficit. Despite their high requirements to system modeling, these methods are able to achieve efficient, demand-oriented, and optimal network configurations. In this way, they eliminate some of the heuristics widely applied in networks design. Of course, the forward engineering design paradigm cannot completely replace the iterative approach. For example the planning of network prototypes will still use the iterative design. However, if the design objectives are well-defined, the forward engineering approach will out-perform the conventional method. This will be demonstrated for case of cellular network planning in Chapter 6.

Part II

Demand-oriented Telecommunication System Design

5 Spatial Customer Traffic Estimation and Characterization

The primary task of telecommunication network planning is to supply communication services to customers which are distributed service area. To provide the service the core task of engineering is to locate and configure transmission facilities, i.e., base stations or switching centers, in the service area of the network and to interconnect these nodes in an optimal way. To achieve an efficient configuration of these spatially extended systems, new teletraffic models are required to evaluate their *spatial performance*, cf. Wirth (1997).

In particular, the design of mobile networks has to be based on the analysis of the *distribution of the expected spatial teletraffic demand* in the complete service area. However, most of the traffic models applied so far for the demand estimation characterize the traffic only in a single cell, e.g. Hong and Rappaport (1986). Other traffic models, like the *highway Poisson-Arrival-Location Model (PALM)* proposed by Leung et al. (1994), give deep theoretical insights, but they are too complex for practical use in mobile system engineering. Hence, the demand-based design of mobile communication systems requires efficient traffic estimation and characterization procedures which are at the same time both accurate and simple to use. An efficient framework which is capable to meet these requirements will be introduced in this chapter.

The chapter is organized as follows. Section 5.1 provides first an overview on *traffic source models* which are used so far in mobile network design. In the second part, a *spatial traffic estimation model* is defined which takes into account the geographical and demographical factors for the expected teletraffic in a service region. Subsequently, the

Demand Node Concept (DNC) is introduced, which is an efficient technique to represent the spatial distribution of the teletraffic using discrete points, called *demand nodes*. Section 5.2 outlines a *traffic characterization procedure* which is capable to derive a demand node distribution from publicly available geographical data. The methods to generate demand nodes distributions are presented in Section 5.3. Here, two *hierarchical clustering algorithms*, a *partitional* clustering method and an *agglomerative* procedure, are discussed in detail and evaluated with regard to their traffic characterization capability. Section 5.4 is devoted to the evaluation of the impact of user clustering on mobile network performance and Section 5.5 outlines some applications of the demand node concept in telecommunication network engineering. The chapter is concluded by Section 5.6, which gives a summary of the results and a short outlook on the future development of the demand node concept.

5.1 Spatial Traffic Estimation

In mobile communication networks the teletraffic originating in the service area of the system can be described mainly by two traffic models which differ by their view of the network. The *traffic source model*, which is also often referred to as the *mobility model*, describes the system as seen by the mobile station. The traffic scenario is represented as a population of individual traffic sources performing a random walk through the service area and randomly generating demand for resources, i.e., the radio channels. An overview of these models is provided in Section 5.1.1.

In contrast, the *network traffic model* of a mobile communication system describes the traffic as observed from the non-moving network elements, e.g. base stations or switches. This model characterizes the *spatial* and *time-dependent* distribution of the teletraffic. The traffic intensity λ is in general measured in call attempts per time unit and space unit ([calls/(sec·km^2)]). Taking additionally the mean call duration $E[B]$ into account, the offered traffic is $a = \lambda \cdot E[B]$ (in [Erlang/km^2]). This measure represents the amount of offered traffic in a defined area.

Both traffic models are used in mobile communication system design. Particularly the latter model is of principal interest when determining the location of the main facilities in a mobile network. These components should be located close to the expected traffic in order to increase the system efficiency. Therefore, we focus in Section 5.1.2 in greater detail on this type of models.

5.1.1 Traffic Source Models

Due to their capability to describe the user behavior in detail, *traffic source models* are usually applied for the characterization of the traffic in an individual cell of a mobile network. Using these models, local performance measures like *fresh call blocking probability* or *handover blocking probability* can be derived from the mobility pattern. Additionally, these models can be used to calculate the subjective Quality-of-Service values for individual users, cf. Section 5.4.

Overview on Traffic Source Models

A widely used single cell model was first introduced by Hong and Rappaport (1986). Their model assumes a uniformly distributed mobile user density and a non-directional uniform velocity distribution of the mobiles. Under this premise, performance values like the *mean channel holding time* and the *average call origination rate* in a cell can be computed.

A more accurate modeling of the calling behavior of users in a single cell was proposed by Tran-Gia and Mandjes (1997). The model considers a base station with a finite customer population and repeated attempts. The appealing characteristic of the model is the assumption of a small, finite user population. However, the model is limited to a single cell and does not consider the spatial variation of the teletraffic within the service area.

El-Dolil et al. (1989) characterized the mobile phone traffic on vehicular highways by assuming a one-dimensional mobility pattern for the customers. They derive the performance values by applying a stationary flow model for the vehicular traffic. An extended one-dimensional highway model with a non-uniform density distribution, denoted as the *highway PALM* model, was investigated by Leung et al. (1994). For the traffic characterization, fluid flow models with time-nonhomogeneous and time-homogeneous traffic have been used, as well as an approximative stochastic traffic model.

A limited directional two-dimensional mobility model was investigated by Foschini et al. (1993). The model assumes a spatially homogeneous distribution of the demand and an isotropic mobility structure. Chlebus (1993) and Chlebus and Ludwin (1995) investigated a mobility model with a homogeneous demand distribution and a non-uniform velocity distribution. The traffic orientation is non-directional and uniformly distributed.

The application of these traffic source models in real network planning cases is strongly limited. Some models, like the highway PALM model give a deep insight on the impact of the terminal mobility on the cellular system performance, however they are rather complex to be applied in real network design. Other models, like the one suggested by Hong and Rappaport (1986), due to their simplification assumptions, can only be applied for the determination of the parameters in an isolated cell.

5.1.2 Traffic Intensity

Since the cellular network planning process requires a comprehensive view of the expected load and since the traffic source models only focus on a single cell, a network teletraffic model has to be specified. Therefore, we define the *traffic intensity* function $\lambda^{(t)}(x, y)$. This function describes the number of call requests seen by the fixed network elements, in a unit area element at location (x, y) during time interval $(t, t + \Delta t)$. The coordinates (x, y) of the area element are integer numbers. Due to the definition given above, the traffic intensity function is a matrix of traffic values representing the demand from area elements in the service region, cf. Figure 5.1(b). The traffic intensity $\lambda^{(t)}(x, y)$ can be derived from the density and the call attempt rate of the mobile stations.

Under the premise that this probability $p_{\mathrm{loc}}^{(t)}(\chi, \psi)$ is known, the average number of mobile units $\#\overline{mob}^{(t)}(x, y)$ in a certain area element at time t is:

$$\#\overline{mob}^{(t)}(x, y) = \int_{x}^{x+\Delta x} \int_{y}^{y+\Delta y} p_{\mathrm{loc}}^{(t)}(\chi, \psi) d\psi d\chi \,. \tag{5.1}$$

Here, $p_{\mathrm{loc}}^{(t)}(\chi, \psi)$ is the density, if the system is viewed from the outside, of mobile stations at location (χ, ψ). The location (χ, ψ) is a coordinate in $\mathbb{R}^2$ and $\Delta x \times \Delta y$ is the size of the unit area element.

Using the assumption that every mobile station has the same *call attempt rate* $r(t)$ at time t, the traffic intensity $\lambda^{(t)}(x, y)$ can be readily obtained:

$$\lambda^{(t)}(x, y) = \#\overline{mob}^{(t)}(x, y)\, r(t) \,. \tag{5.2}$$

Since in real world planning cases it is almost impossible to directly calculate the location probability $p_{\mathrm{loc}}^{(t)}(\chi, \psi)$ from the mobility model, the traffic intensity has to be derived from indirect statistical measures.

5.1.3 Geographic Network Traffic Model

The offered traffic in a region can be estimated by the *geographical* and *demographical* characteristics of the service area. Such a demand model relates factors like *land use*, *population density*, *vehicular traffic*, and *income per capita* with the calling behavior of the mobile units. The model applies statistical assumptions on the relation of traffic and land use type with the estimation of the demand. In the *geographic network traffic model*, the offered traffic $A_{geo}^{(t)}(x, y)$ is the aggregation of the traffic originating from these various factors:

$$A_{geo}^{(t)}(x, y) = \sum_{\text{all factors } i} a_i \cdot \delta_i^{(t)}(x, y), \tag{5.3}$$

where $a_i = \lambda_i \cdot E[B_i]$ is the traffic generated by factor i in an arbitrary area element of unit size, measured in *Erlangs per area unit*, λ_i the average number of call attempts per time unit and space unit initiated by factor i, $E[B_i]$ is the mean call duration of calls of type i, and $\delta_i^{(t)}(x, y)$ is the assertion operator:

$$\delta_i^{(t)}(x, y) = \begin{cases} 0 : \text{traffic factor } i \text{ is not true at location } (x, y) \\ 1 : \text{traffic factor } i \text{ is true at location } (x, y) \end{cases}. \tag{5.4}$$

So far, the planning of public communication systems uses geographic traffic models which have a large granularity. In these cases, a typical *unit area size* is in the order of square kilometers, i.e., in public cellular mobile systems this is the size of *location areas*, cf. Grasso et al. (1996). For the determination of the location of transmission facilities a much smaller value is required. Their locations have to be determined within a spatial resolution of one hundred meters. Thus, a unit area element size in the range of $100m \times 100m$ is here indicated.

Traffic Parameters

The values for a_i, which are the traffic values originating from factor i per area element, can be derived from measurements in an existing

mobile network and by taking advantage of the known causal connection between the traffic and its origin, cf. Grillo et al. (1998). A first approach is to assume a highly non-linear relationship. A general structure to model this behavior is to use a parametric exponential function. In our proposed geographic model, the traffic-factor relationship is defined to be:

$$a_i = c \cdot b^{x_i} \tag{5.5}$$

where c is constant and b is the base of the exponential function.

To reduce the complexity of the parameter determination we introduce the normalization constraint:

$$\frac{A_{\text{total}}}{S_{\text{service area}}/s_{\text{unit element}}} = \sum_{\text{all factors } i} a_i \,, \tag{5.6}$$

where $S_{\text{service area}}$ is the size of the service area, $s_{\text{unit element}}$ is the size of a unit area element, and A_{total} is the total teletraffic in this region. The value of A_{total} can be measured in an operating cellular mobile network. If these measurements are not available, reasonable assumptions for the parameters have to be made, cf. Grillo et al. (1998).

The structure of the geographical traffic model given in Eqn. (5.3) and Eqn. (5.5) appears to be simple. However, due to its structure the model can be adapted to the proper traffic parameters. This capability enables its application for mobile system planning.

Stationary Geographic Traffic Model

The above proposed model $A_{geo}^{(t)}(x, y)$ includes also the temporal variation of the traffic intensity in the service area. Since communication systems must be configured in such a way that they can accommodate the highest expected load, the time index t is usually dropped and the traffic models are reduced to *stationary* models describing the peak traffic. The maximum load is the value of the traffic during the *busy hour*, cf. Mouly and Pautet (1992).

A pitfall for the network designer remains: the busy hour varies over time within the service area. In downtown areas the highest traffic usually occurs during business hours, whereas in suburban regions the busy hour is expected to be in the evening. Therefore, the network engineer has to decide how to weigh the different traffic factors, i.e. how to obey the different shares of various user groups in the traffic model of the network.

5.1.4 Traffic Discretization and Demand Nodes

The core technique of the traffic characterization proposed in this chapter is the representation of the spatial distribution of the demand for teletraffic by discrete points, denoted as *demand nodes*. Demand nodes are widely used in economics for solving facility location problems, cf. Ghosh and McLafferty (1987).

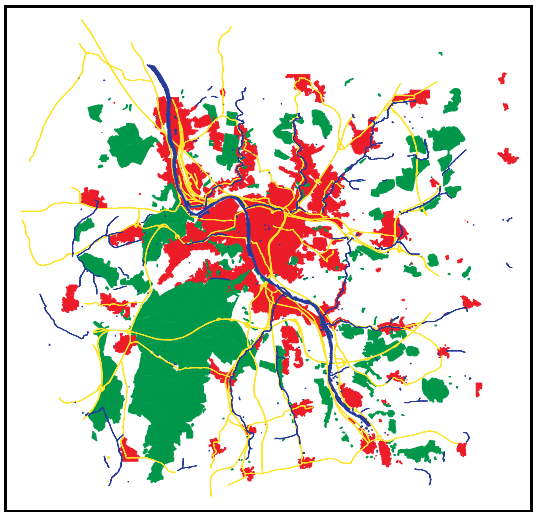
Definition 5.1 : Demand Node

A demand node represents the center of an area that contains a quantum of demand from teletraffic viewpoint, accounted in a fixed number of call requests per time unit.

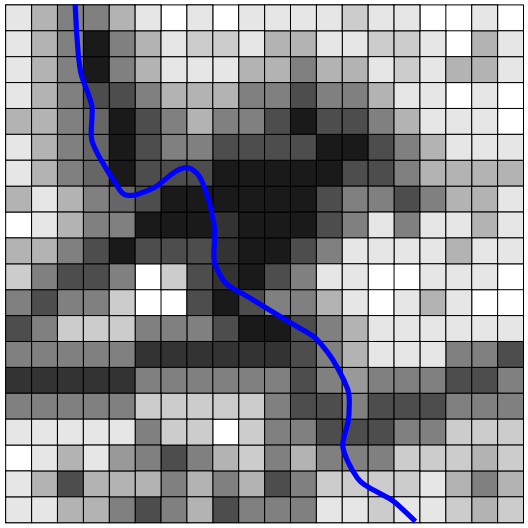
The notion of demand nodes introduces a discretization of the demand in both space and amount. In consequence, the demand nodes are dense in areas of high traffic intensity and sparse in areas of low traffic intensity. Together with the time-independent geographic traffic model, the *demand node concept (DNC)* constitutes, in the context of cellular network design, a *static population model* for the description of the subscriber distribution.

An illustration of the *demand node concept* is given in Figure 5.1: part (a) shows publicly available map data with land use information for the area around the city of Würzburg, Germany. The information was extracted from *ATKIS*, the official topographical cartographical data base of the Bavarian land survey office, cf. ATKIS (1991). The depicted region has the size of $15km \times 15km$. Figure 5.1(b) sketches the traffic intensity distribution in this area, characterized by the traffic matrix: dark squares represent an expected high demand for mobile service, bright values correspond to a low teletraffic intensity. Part (c) of Figure 5.1 shows a tesselation of the service area obtained by the partional clustering algorithm used for generating the demand node distribution, cf. Section 5.3. Figure 5.1(d) depicts a simplified result of the demand discretization. The demand nodes are dense in the city center and on highways, whereas they are sparse in rural areas.

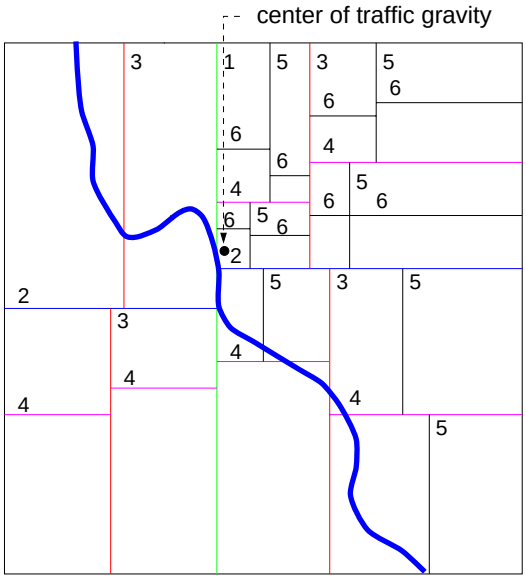
In principle the two-dimensional teletraffic density matrix, cf. Figure 5.1(b), is sufficient to characterize the teletraffic distribution in the service area. However, the application of the demand node representation decreases significantly the computational requirements in network design. Due to the use of a discrete point representation, is not necessary any more in mobile system design to calculate the field strength



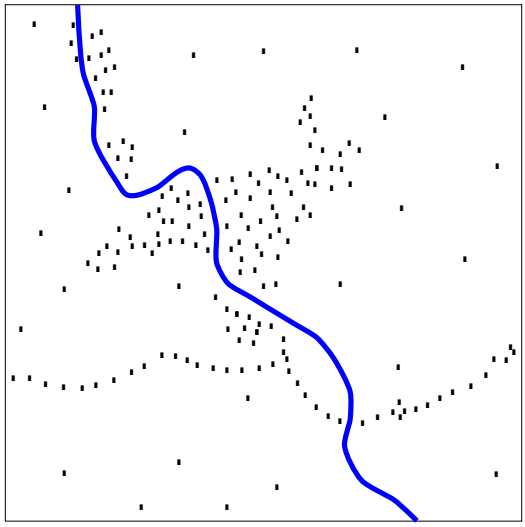
(a) Geographical and demographical data



(b) Traffic matrix



(c) Service area tessellation



(d) Demand node distribution

Figure 5.1: *Demand node concept (DNC)*

at every point in the area. It is sufficient to compute the field strength values only at the location of the demand nodes, cf. Chapter 6. Moreover, the discrete presentation can be used to characterize the clustering effect of users in the service area. The demand node concept enables the evaluation of the impact of this user clumping on network performance, cf. Section 5.4.

5.2 Spatial Traffic Characterization

5.2.1 Traffic Characterization Procedure

Based on the estimation method introduced in the previous section, the traffic characterization procedure has to compute the spatial traffic intensity and its discrete demand node representation from realistic data taken from available databases. In order to handle this type of data, the complete characterization process comprises four sequential steps:

Step 1 **Traffic model definition:**
Identification of traffic factors and determination of the traffic parameters in the geographical traffic model.

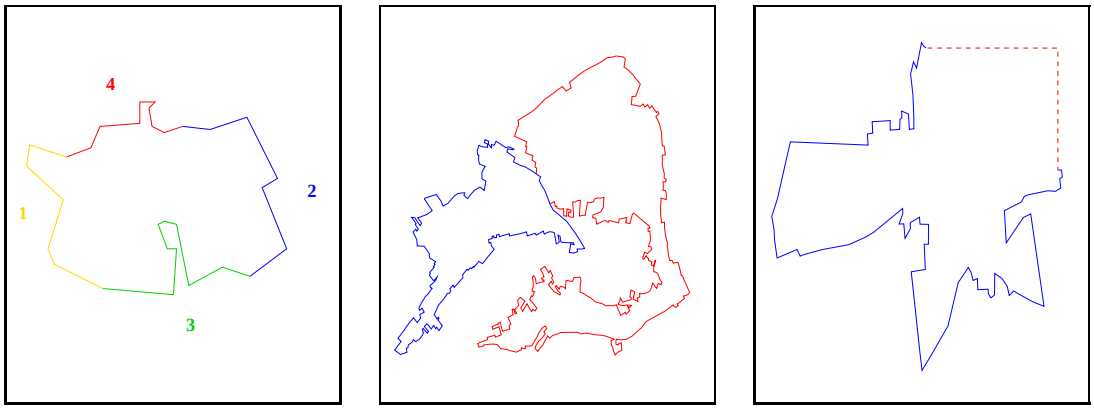
Step 2 **Data preprocessing:**
Preprocessing of the information in the geographical and demographical database.

Step 3 **Traffic estimation:**
Calculation of the spatial traffic intensity matrix of the service region.

Step 4 **Demand node generation:**
Generation of the discrete demand node distribution by the application of clustering methods.

Traffic Model Definition

The definition of the geographical traffic model in *Step 1* of the characterization procedure is based on the arguments given in Section 5.1.3. A simple but accurate spatial geographic traffic model is the foundation for system optimization in the subsequent network design steps.



(a) Unordered lines

(b) Adjacent open and closed polygons

(c) Closing lines

Figure 5.2: *Original map information and data preprocessing*

Data Preprocessing

The data preprocessing in *Step 2* is required since the data in geographical information systems are usually not collected with respect to mobile network planning. For example, ATKIS' main purpose is to maintain map information. It uses a vector format for storing its drawing objects.

To determine the clutter type of a certain location, one has to identify the land type of the area surrounding this point. This requires the detection of the closed polygon describing the shape of this area. Since maps are mostly printed on paper, the order of drawing the lines of a closed shape doesn't matter, see Figure 5.2(a). To identify closed polygons, one has to check if every ending point of a line is a starting point of another one. If a closed polygon has been detected, the open lines are removed from the original database and replaced by its closed representation. Additionally, due to the map nature of the data, two adjacent area objects can be stored by a closed and an open polygon, see Figure 5.2(b). It also can happen that some data is missing, see Figure 5.2(c). In this case, line closing algorithms have to applied. After the preprocessing step only closed area objects remain in the database and the traffic characterization can proceed with the next step: demand estimation.

Traffic Estimation

Step 3 of the traffic characterization process uses the geographical traffic model defined in *Step 1* for the estimation of the teletraffic demand per unit area element. The computed traffic values are stored in the *traffic matrix*. To obtain the traffic value on a certain unit area element, the procedure first determines the traffic factors which are valid for this element and then computes the matrix entry by applying Eqn. (5.3).

5.3 Demand Node Generation

The generation of the demand nodes in *Step 4* of the traffic characterization process is performed by a *clustering method*. Clustering algorithms are in general distinguished into two classes, cf. Jain and Dubes (1988): *a)* the *Hierarchical Clustering* procedures and *b)* the *Partitional Clustering* methods. The hierarchical procedures construct a nested sequence of partitions between the properties of the data points, whereas general partitional methods try to find suitable taxonomies in the data space. Thus, the class of hierarchical procedures is a special subclass of the partitional methods.

In the context of mobile network planning, the design procedures for these systems should be fast, efficient, and flexible. Therefore, general partitional clustering methods are not considered in this work due to their high complexity and limited flexibility. However, it should be mentioned that significant research has been carried out in this field. In particullar, the procedures based on Voronoi diagrams are well investigated. A comprehensive presentation of these algorithms can be found in Okabe et al. (1992).

In the following section two instances of hierarchical clustering algorithms are investigated in detail: a divisive procedure and an agglomerative method. The divisive procedure starts with considering the whole service area and its traffic. After that, it recursively subdivides the area into smaller pieces. Due to its partitioning nature, this procedure will be denoted as the "partitional clustering algorithm". The agglomerative method reverses this process by initializing every traffic matrix element with a cluster object. After that, it gradually merges these atomic clusters into larger and larger clusters until all clusters obey the predefined constraints.

Since both algorithms correspond to a different choice of constructing the nested sequence rather than to a different kind of clustering, the

Algorithm 5.1 (Generate Demand Nodes by partitioning)

variables:

dnode_set	*global variable for the set of generated demand nodes*
orient	*orientation of partitioning line*
θ	*traffic quantization value*

algorithm:

```
1  proc gen_dnodes(area, T_max, orient = 0) ≡
2    begin
3      if (traffic(area) < T_max)
4        then dnode_set ← center_traffic(area);
5             return;
6        else orient ← (orient + 90°) mod 180°;  /* turn partitioning line */
7             a_l = left_area(area, orient);
8             a_r = right_area(area, orient);
9             gen_dnodes(a_l, θ, orient);          /* do the recursion */
10            gen_dnodes(a_r, θ, orient);
11     fi
12   end
```

Algorithm 5.1: Demand node generation by partitional clustering

results obtained by both methods are expected to be in general comparable. The differences and similarities between of these two algorithms is the topic of the following sections.

5.3.1 Partitional Clustering

The first algorithm proposed for the demand generation is a recursive partitional clustering method. It is based on the idea to divide the service area until the teletraffic of every tessellation piece is below a quantization value $T_{\max}$. Thus, the algorithm constructs a sequence of bisections of the service area. The demand node location is the center of gravity of the traffic weight of the tessellation pieces. The main procedure of the partitional demand node generation algorithm is shown in Algorithm 5.1.

The function `left_area()` divides the *area* into two rectangles with the same teletraffic and returns the left part of the bisection. The function `right_area()` returns the right piece. In every recursion step, the orientation of the partitioning line is rotated by 90°. The recursion stops, if every rectangle represents a traffic amount less than the maximal tele-

traffic quantization value $T_{\max}$. The function `traffic()` evaluates the amount of expected teletraffic demand in the area.

Due to its recursive and partitional nature, the algorithm creates demand node distributions with 2^N nodes, where the parameter N denotes the recursion depth of the algorithm. Therefore, the maximum teletraffic criterion "`traffic`$(area) < T_{\max}$" in line 3 of Algorithm 5.1 can be replaced by using the maximum recursion depth $r_{\max}$. The corresponding value of $r_{\max}$ is given by:

$$r_{\max} = \lceil \mathrm{ld} \frac{T_{\text{total}}}{T_{\max}} \rceil, \tag{5.7}$$

where T_{total} is the total teletraffic emerging from the studied area and the function $\lceil x \rceil$ denotes the smallest integer q such that $q \geq x$.

An example for the bisection sequence of the partitional clustering algorithm is shown in Figure 5.1(c). The numbers next to the partitioning lines indicate the recursion depth. To make the example more vivid, not every partition line is depicted in the example. The upper left quadrant of the Figure 5.1(c) only shows the lines until the recursion depth 3, the lower left part only the lines until depth 4, the lower right quarter the lines until depth 5 and the upper right quadrant lines until depth 6.

Two results of the partitional clustering method are depicted in Figure 5.3 and Figure 5.4. They were obtained in two real-world planning case studies. Figure 5.3 shows the two-dimensional demand node distribution for the Würzburg area; a description of this scenario was already given in Chapter 5.1.4. The algorithm received as input the traffic matrix of this region. The matrix was obtained by applying the traffic estimation procedure presented in Chapter 5.1. As expected the demand nodes are dense in urban areas and on highways, cf. Figure 5.1(a), and they are sparse in open areas and forestall regions. An interesting, but not desired feature, is the formation of regular structured patterns in the node distribution. In the city center of Würzburg, the demand nodes are placed in a grid-like manner.

Figure 5.4 depicts the two-dimensional demand node distribution of a $160km \times 160km$ area around the Dallas/Fort Worth metroplex in Texas, USA. The teletraffic matrix used in this experiment was obtained from real traffic measurements and was provided by Nortel Wireless Networks, Richardson, Tx. As can be clearly seen, the demand node distribution is of strong clustered nature. The regions around the downtown areas of Dallas and Fort Worth show a much higher node density than the

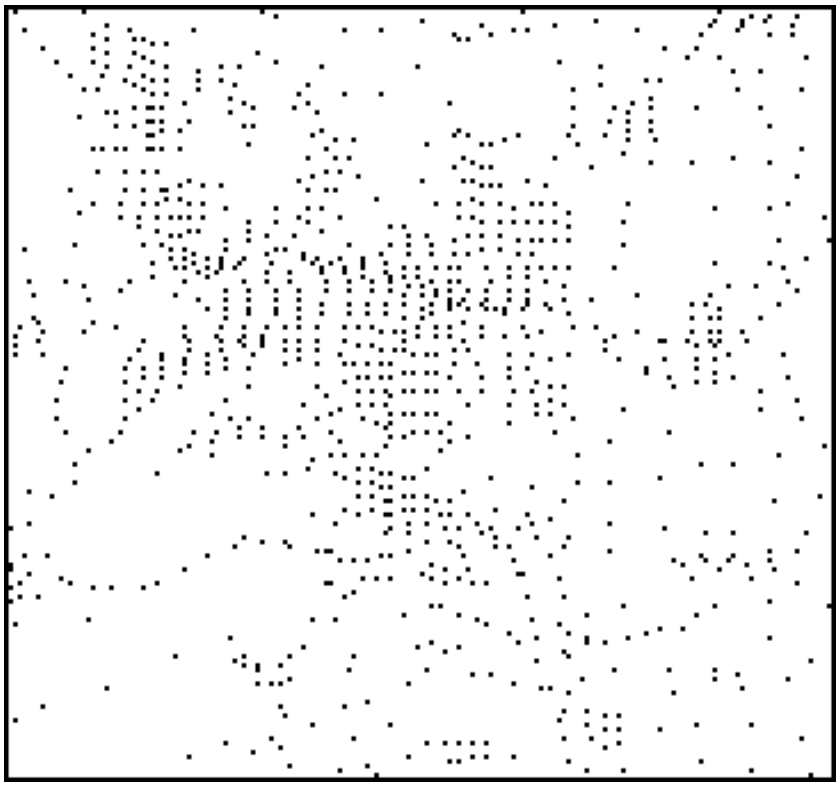


Figure 5.3: *Two-dimensional demand node distribution of the Würzburg area generated by partitional clustering*

regions of the surrounding smaller cities and suburbs. The node distribution reveals, as in the Würzburg experiment, large areas with regular structured demand node patterns.

Characteristics of the Algorithm

The partitional clustering procedure of Algorithm 5.1 is a fast but simple clustering method. However, its accuracy depends strongly on the quantization value $T_{\max}$. This value gives only an upper bound for the traffic represented by a single demand node. A more elaborated evaluation of the spatial statistic of demand node distributions generated by the partitional method is presented in Chapter 5.3.3.

Moreover, since the partitional algorithm constructs a sequence of right-angled bisections, the shape of the tessellation pieces is always rectangular. Hence, the two-dimensional demand node distribution tends to form regular structured patterns in areas with almost constant teletraffic.

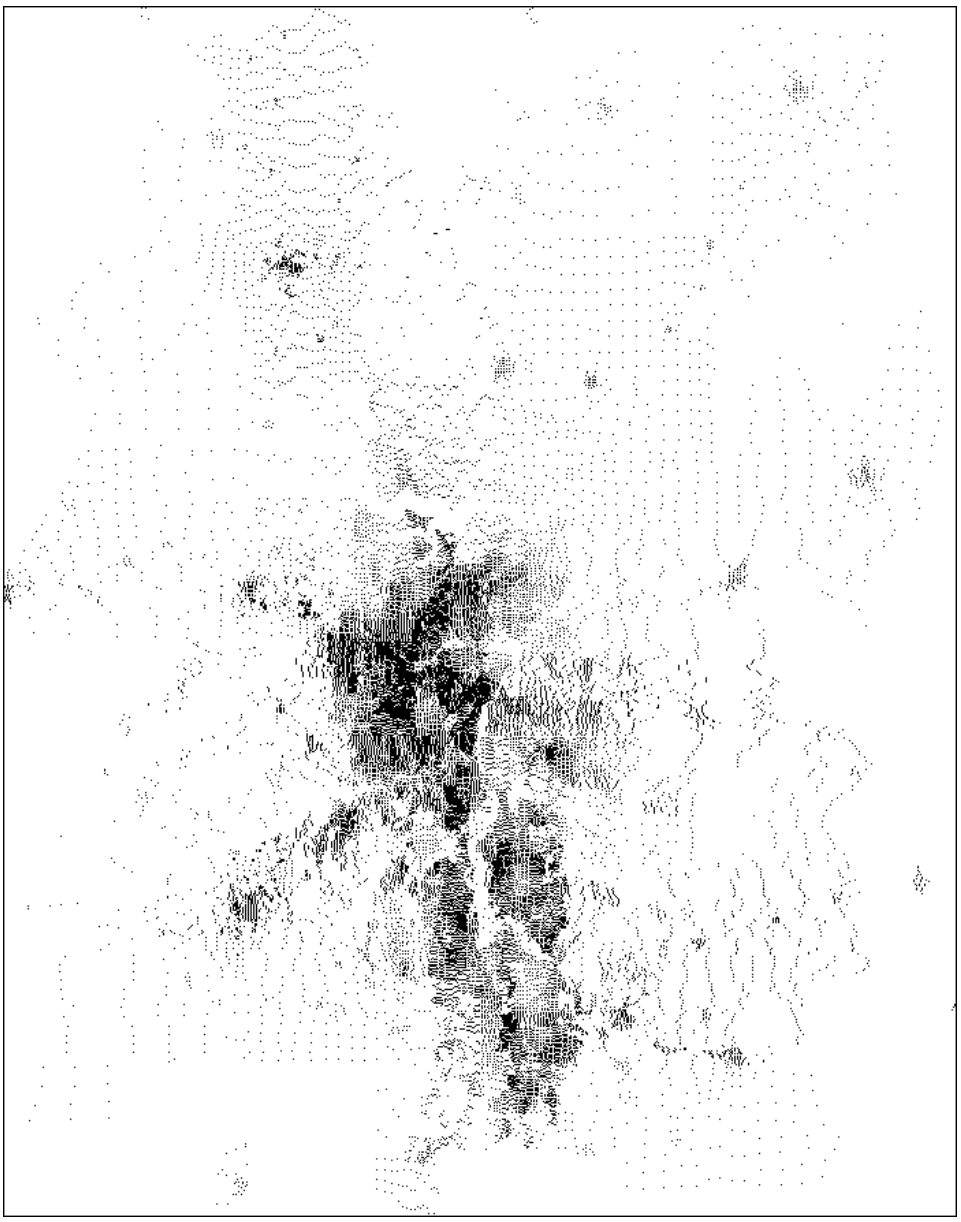


Figure 5.4: *Two-dimensional demand node distribution for the Dallas/Fort Worth area generated by partitional clustering*

Another not desired feature is that no size restriction can be imposed on the area represented by a demand node. This feature is, however, necessary for the application of demand nodes in the design of radio networks of mobile communication systems. In this case, the demand node are considered as sensors for the field strength. They are representing their area portion. If the represented area is too large then the values is not representative any more.

5.3.2 Agglomerative Clustering

The second method proposed for clustering is a hierarchical agglomerative algorithm. The basic idea of this method is to initialize every element of the teletraffic matrix with a cluster object and to merge these atomic objects successively into larger cluster objects while obeying the maximum teletraffic and size constraints. Therefore the algorithm requires as input the teletraffic matrix T, the quantization value for the maximum traffic $T_{\max}$, and the threshold value for the maximum area $A_{\max}$ which can be represented by a demand node. For improving the teletraffic approximation characteristic, an additional parameter for the minimum traffic of a demand node $T_{\min}$ is needed. The main procedure of the agglomerative demand node generation algorithm is depicted in Algorithm 5.2.

The agglomerative algorithm is subdivided into four stages. In the first stage, a cluster object is created for every entry in the teletraffic matrix, cf. Algorithm 5.3.

In the second stage, the algorithm performs the main clustering. A random cluster k is selected from set of all clusters and a suitable neighbor w of k is identified for merging, cf. Figure 5.5(a). The procedure for identifying the suitable neighbor is depicted in Function 5.1. Two cluster objects can only be merged if the resulting object obeys the maximum teletraffic constraint and the maximum size constraint. Hereby, the operator "+" denotes the arithmetical summation and the operator "$\oplus$" the join operation of two area elements. The merging of the cluster objects is repeated until no clusters can be joined anymore. In order to prohibit the algorithm from creating pathological and regular tessellation pieces, the join orientation of the cluster objects is alternated after every step, cf. Figure 5.5(c). In this way, the algorithm is capable of creating a two-dimensional demand node distribution with maximum randomness in the node pattern. The function `neighbors_right_and_below()` identifies the suitable neighbor clusters right next and below to the investigated ob-

Algorithm 5.2 (Generate Demand Nodes by agglomeration)

variables:

Traffic(x,y)	*traffic matrix,* $x \in [1, X]$, $y \in [1, Y]$
$C(i)$	*vector of clusters,* $i \in [1, \|C\|]$
$dnode\_set$	*set of generated demand nodes*
w	*index of neighbor cluster*

algorithm:

```
 1  proc gen_demand_nodes(Traffic, dnode_set, T_max, A_max, T_min) ≡
 2
 3    /* stage 1 */
 4    init_clusters(Traffic, C);
 5
 6    /* stage 2 */
 7    while any clusters left for agglomeration
 8          k := rnd([1, |C|]);        /* randomly pick a cluster */
 9          w := find_neighbor(C, k, T_max, A_max);
10          if w ≠ 0 then              /* merge cluster i and w */
11            C(k).T := C(k).T + C(w).T;
12            C(k).A := C(k).A ⊕ C(w).A;
13          fi
14    end
15
16    /* stage 3 */
17    while any clusters can be optimized
18          k := rnd([1, |C|]);        /* randomly pick a cluster */
19          if C(k).T < T_min then
20            split_cluster(C, k, T_max, A_max)
21          fi
22    end
23
24    /* stage 4 */
25    dnode_set ← center_traffic(C);
26  end
```

Algorithm 5.2: Generate demand nodes by agglomerative clustering

ject. The function `neighbors_left_and_above()` identifies the neighbor clusters left and above.

According to Defintion 5.1, every demand node has to represent the same teletraffic amount. However, due to the randomness in the selec-

Algorithm 5.3 (Initial clustering)

variables:

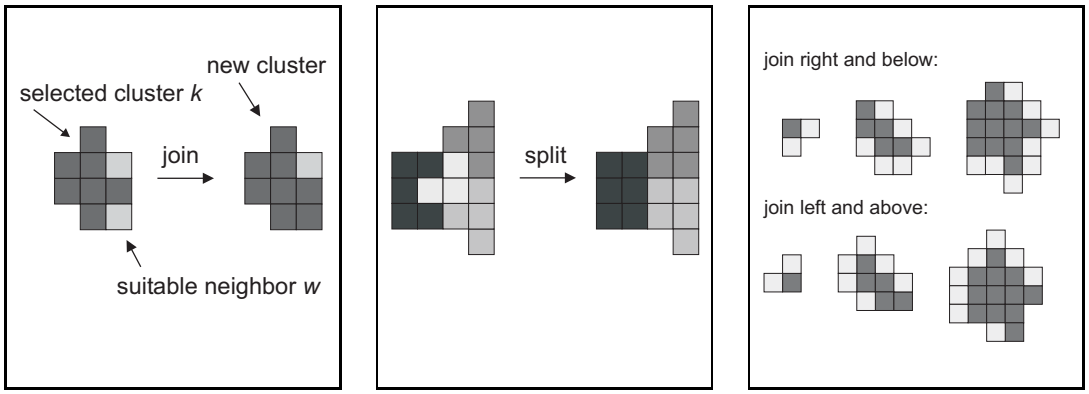
$Traffic(x, y)$	*traffic matrix,* $x \in [1, X]$, $y \in [1, Y]$
$C(i)$	*vector of clusters,* $i \in [1, \|C\|]$

algorithm:

```
1  proc init_clusters(Traffic, C) ≡
2    /* for every matrix element create a cluster */
3    i := 0;
4    for x := 1 to X do
5        for y := 1 to Y do
6            C(i).T = Traffic(x, y);
7            C(i).A = (x, y);
8            i := i + 1;
9        end
10   end
11 end
```

Algorithm 5.3: Initial clustering

tion of joining clusters and the additional size constraint, the main stage of agglomerative algorithm is not able to assure this requirement anymore. Hence, to obtain an applicable and efficient clustering algorithm, an optimization phase is executed after the main clustering stage. The optimization stage reduces the impact of this characteristic on the traffic



(a) Merge cluster (b) Split cluster (c) Change join orientation

Figure 5.5: *Basic operations of the agglomerative clustering algorithm*

Function 5.1 (Search for suitable neighbor for cluster k)

variables:

$C(i)$	*vector of clusters,* $i \in [1, \|C\|]$
$N(i)$	*vector of neighbor cluster,* $i \in [1, \|N\|]$
k	*specifies the cluster under investigation*

algorithm:

```
 1  funct find_neighbor(C, k, T_max, A_max) ≡
 2
 3    /* search for neighbors and change join orientation */
 4    if C(k).orient ≠ 0
 5      then N = neighbors_right_and_below(C, k);
 6           C(k).orient = 1;
 7       else N = neighbors_left_and_above(C, k);
 8           C(k).orient = 0;
 9    fi
10
11    /* start with the first neighbor */
12    n_min = N(1);
13
14    /* search suitable neighbor with smallest traffic*/
15    foreach n ∈ N do
16      if C(n).T + C(k).T < T_max      ∧
17        C(n).|A| + C(k).|A| ≤ A_max   ∧
18        C(n).T < C(n_min).T
19        then n_min = n;
20      fi
21    end
22
23  return n_min   /* return index of neighbor */
```

Function 5.1: Search for a suitable neighbor for merging

representation capability.

In the optimization stage, all cluster objects with a teletrafic value less then the minimum teletraffic constraint $T_{\min}$ are split up and their area elements are distributed among the neighbors, cf. Figure 5.5(b). The split procedure is depicted in Algorithm 5.4. Due to this splitting operation, the algorithm obtains a leveling of the teletraffic values of the cluster objects. The requirement of an equal teletraffic amount for every demand node is still not sharply obeyed. However, the probability that

Algorithm 5.4 (Split a cluster k by distributing its elements among the neighbors)

variables:

$C(i)$	*vector of clusters,* $i \in [1, \|C\|]$
$N(i)$	*vector of nieghbor clusters,* $i \in [1, \|N\|]$
k	*specifies the cluster to be split*
e	*atomic traffic matrix element*

algorithm:

```
1  proc split_cluster(C, k, T_max, A_max);  ≡
2    foreach e ∈ C(k).A do
3
4      /* check for neighbors */
5      N = neighbors(C, C(k).A(e));
6
7      /* start with the first neighbor */
8      n_min = N(1);
9      found = FALSE;
10
11     /* search suitable neighbor with smallest traffic */
12     foreach n ∈ N do
13       if C(n).T + C(k).A(e).T < T_max   ∧
14          C(n).|A| + C(k).|A(t)| ≤  A_max  ∧
15          C(n).T < C(n_min).T
16          then n_min = n;
17               found = TRUE;
18       fi
19     end
20
21     /* move element e to cluster n_min */
22     if found ≠ TRUE then
23       C(n_min).T := C(n_min).T + C(k).T(e);
24       C(n_min).A := C(n_min).A ⊕ C(k).A(e);
25       C(k).T := C(k).T − C(k).A(e).T;
26       C(k).A := C(k).A ⊖ C(k).A(e);
27     fi
28   end
29 end
```

Algorithm 5.4: Split cluster

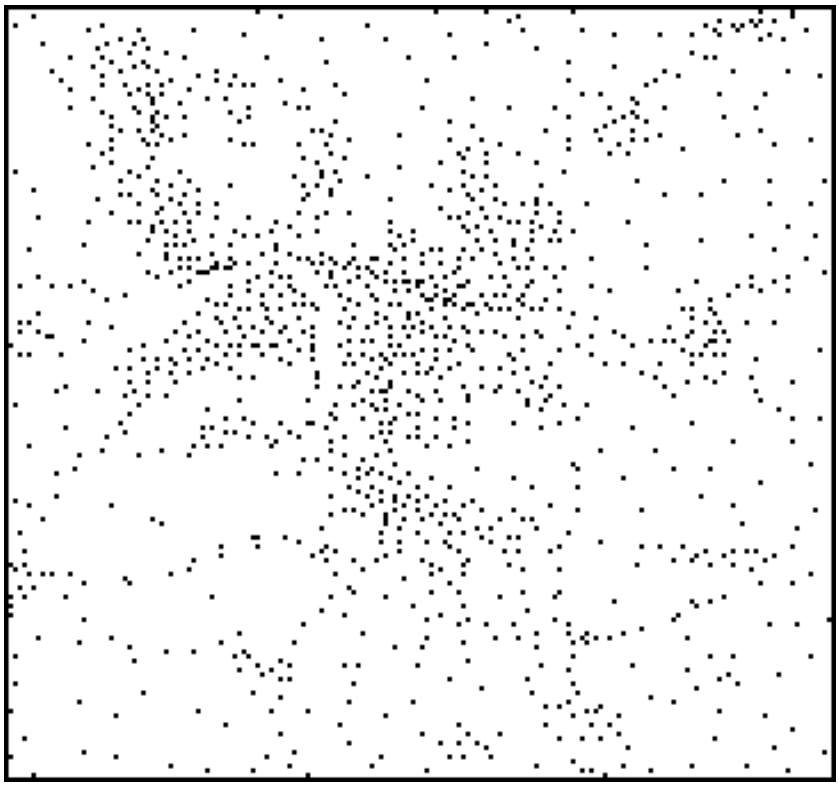


Figure 5.6: *Two-dimensional demand node distribution for the Würzburg area generated by agglomerative clustering without optimization*

a demand node represents a traffic value less than $T_{\min}$ is significantly reduced. An elaborated analysis of this behaviour is presented in Chapter 5.3.3. The function `neighbor()` identifies for a given element of the traffic matrix the suitable neighbor cluster for merging. The operator "$-$" denotes the arithmetical subtraction and the operator "$\ominus$" the split operation of two area elements.

In the last stage of the agglomerative algorithm, a demand node is created for every cluster object. This is performed by the function `center_traffic()`. The demand node location is the center of gravity of the traffic weight of the area represented by the cluster object.

During numerous experiments it turned out that the best performance of the optimization stage was obtained when the ratio of traffic quantization parameters $T_{\min}/T_{\max}$ is set to 0.99. For this value almost every cluster is checked whether it can be split or not.

Three results of the agglomerative clustering method are shown in

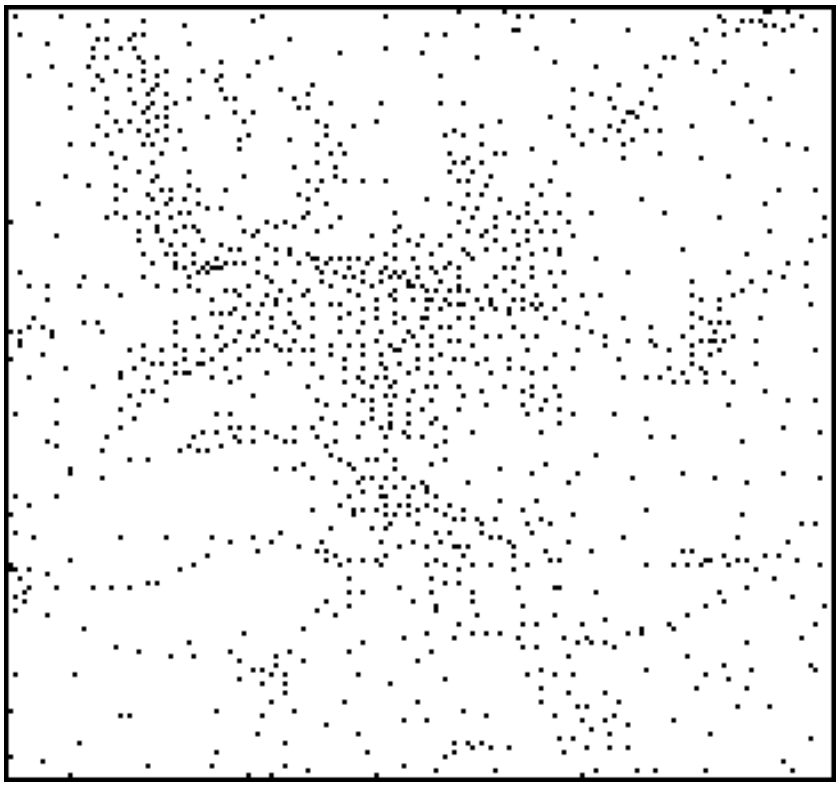


Figure 5.7: *Two-dimensional demand node distribution for the Würzburg area generated by agglomerative clustering with optimization*

Figure 5.6, Figure 5.7, and Figure 5.8. They were obtained in the same planning experiments as described in the previous section.

Figure 5.6 depicts the two-dimensional demand node distribution for the Würzburg area. In this experiment, the optimization stage was not performed. The distribution shows an area of a high node density in the region of the city center, whereas the demand nodes are sparse in rural regions. In contrast to the partitional algorithm, cf. Figure 5.3, it is not possible to identify regular structures in the demand node pattern. The demand nodes are clearly clustered but randomly distributed.

Figure 5.7 depicts the result of the agglomerative algorithm for the Würzburg area obtained by the optimization phase. By visual inspection, no notable difference can be recognized between this distribution and the one generated without optimization. However, both distributions differ significantly in their traffic quantization capability, cf. Chapter 5.3.3.

Figure 5.8 shows the result of the agglomerative method for the Dal-

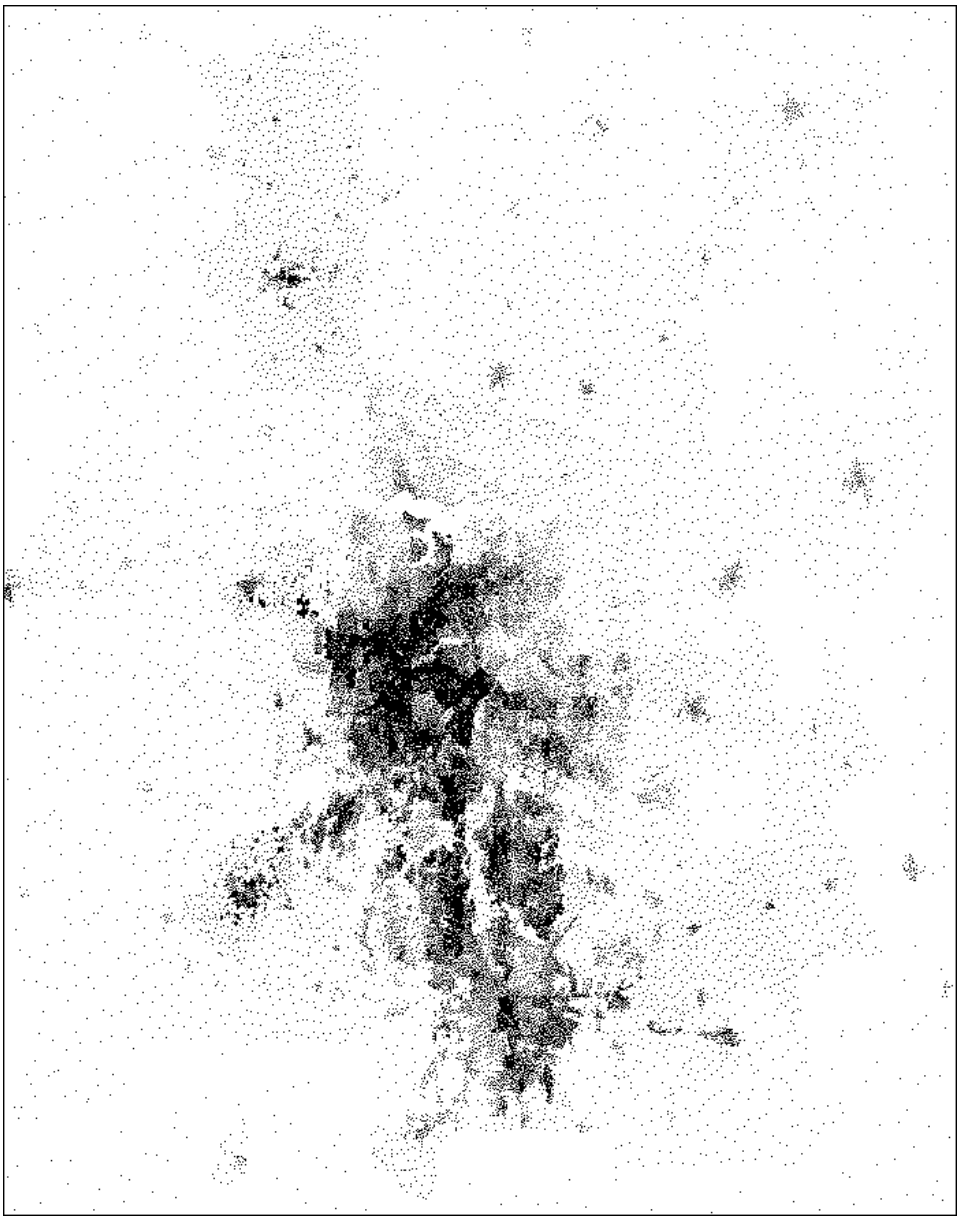


Figure 5.8: *Two-dimensional demand node distribution for the Dallas/Fort Worth area generated by agglomerative clustering with optimization*

las/Fort Worth area. The distribution shows a strong clustering of the demand nodes. However, there are no areas with regular structure visible in the distribution. This example demonstrates that the agglomerative method is able to handle large scale areas for clustering.

Characteristics of the Algorithm

The agglomerative clustering procedure of Algorithm 5.2 is more complex than the partitional method. It requires a significantly higher computational effort. The bottom-up characteristic of the agglomerative approach and the randomness in selecting the cluster objects for merging weakens the traffic description concept of the demand node notion, cf. Defintion 5.1. However, the application of an optimization stage reduces the impact of this property on the teletraffic approximation capability. The accuracy of the algorithm depends on the proper selection of the teletraffic threshold and quantization values $T_{\min}$ and $T_{\max}$. For a suitable choice of these parameters, the agglomerative algorithm behaves stable.

Furthermore, the hierarchical agglomerative clustering method is capable to obtain spatial tessellation pieces of arbitrary shape and of predefined size and traffic value. Thus, the agglomerative algorithm is able to overcome the drawbacks of the partitional approach.

5.3.3 Spatial Evaluation

The applicability of the clustering algorithms depends on the accuracy of the methods to approximate the spatial properties of the teletraffic in the study area. The generated demand node distributions have to show the same spatial variation as the actual teletraffic in this area.

Since the demand node concept introduces a discretization of the teletraffic demand in both amount and space, a demand node representation has to fulfill two criteria: *a)* every demand node has to represent the same quantum of teletraffic and *b)* the spatial statistics of the two-dimensional distributions should be equal to the one in the studied area. In the following section the clustering methods will be validated according to these properties.

Traffic Quantization

The partitional algorithm inherently guarantees the quantization criterion. The recursive separation of the area in two pieces with the same

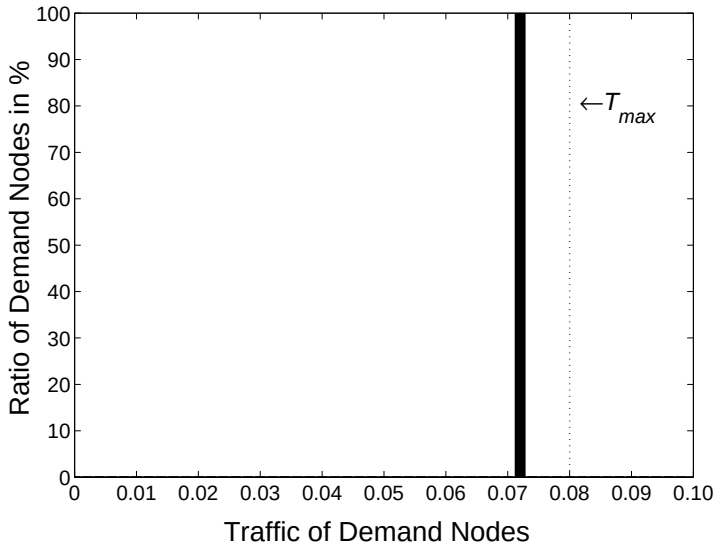
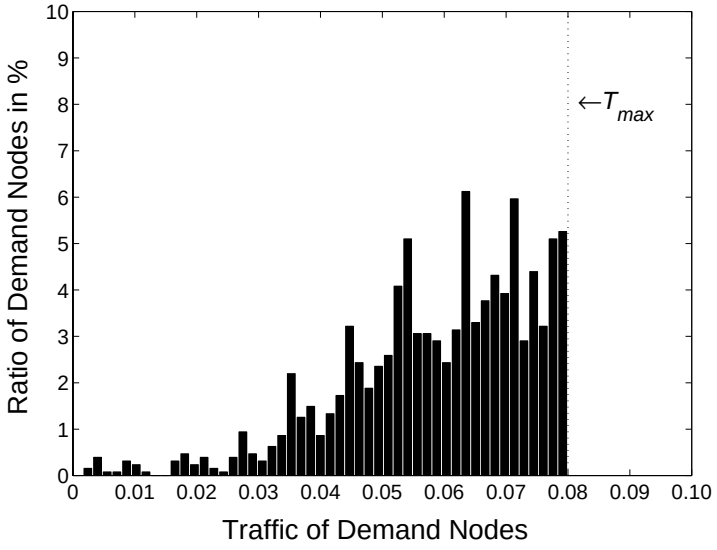


Figure 5.9: *Histogram of the traffic values for the demand node distribution of the Würzburg area created by partitional clustering*

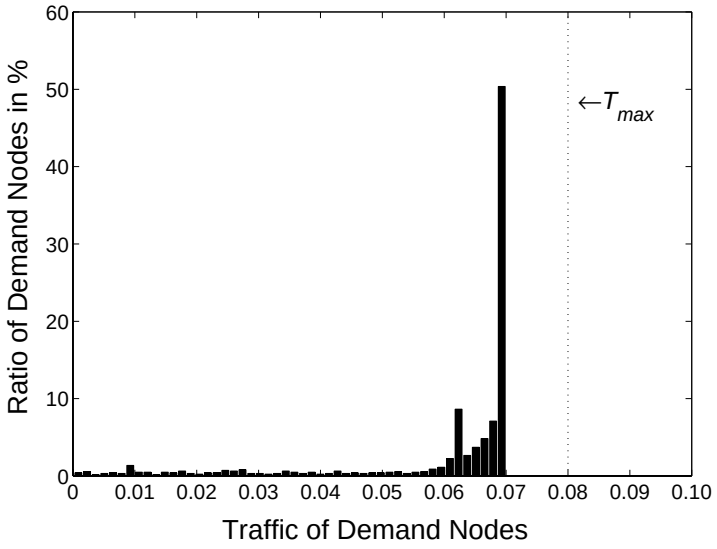
teletraffic constructs a binary tree of decisions. In each recursion level, the decision of performing the recursion is the same. Thus, the traffic weight of the branches of the decision tree is equally balanced and every leaf, i.e., demand node, represents the same teletraffic quantum. Hence, the histogram of the teletraffic values of a demand node distribution generated by the partitional clustering algorithm depicts only a peak, cf. Figure 5.9. The value of $T_{\max}$ is not sharply obeyed due to the granularity of the recursion, cf. Eqn.(5.7). The teletraffic values shown in this histogram and the subsequent ones are relative teletraffic values.

In contrast to the partitional method, the agglomerative clustering algorithm can not assure the same teletraffic value for every demand node. However, due to the optimization stage, the proportion of demand nodes representing a traffic value less than the given minimum teletraffic $T_{\min}$ is minimized. Figure 5.10 shows the impact of the optimization phase on the histogram of the traffic values for the Würzburg planning experiment. Figure 5.10(a) depicts the histogram without the optimization. Here, a large number of nodes have traffic values significantly less than $T_{\max}$. Figure 5.10(b) shows the histogram of the traffic values after the optimization. The proportion of nodes with a teletraffic value close to $T_{\max}$ is highly increased.

Figure 5.11 depicts the histograms of the demand node distributions for the Dallas/Fort Worth planning experiment. The distribution is sig-

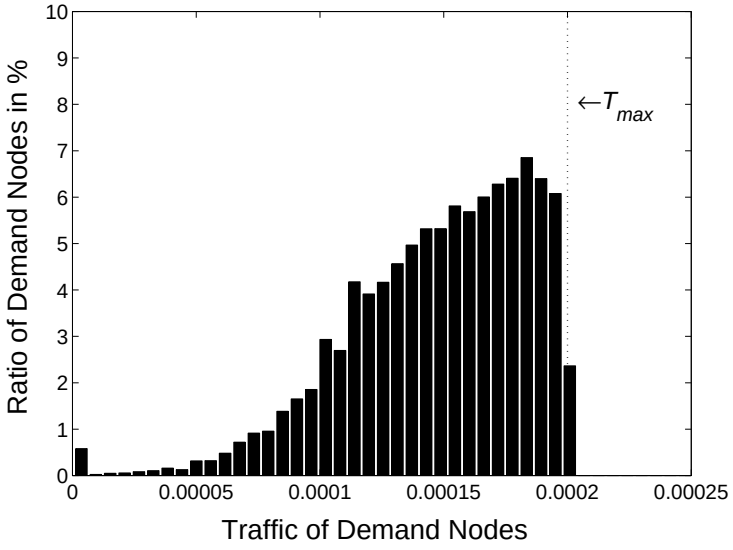


(a) without optimization

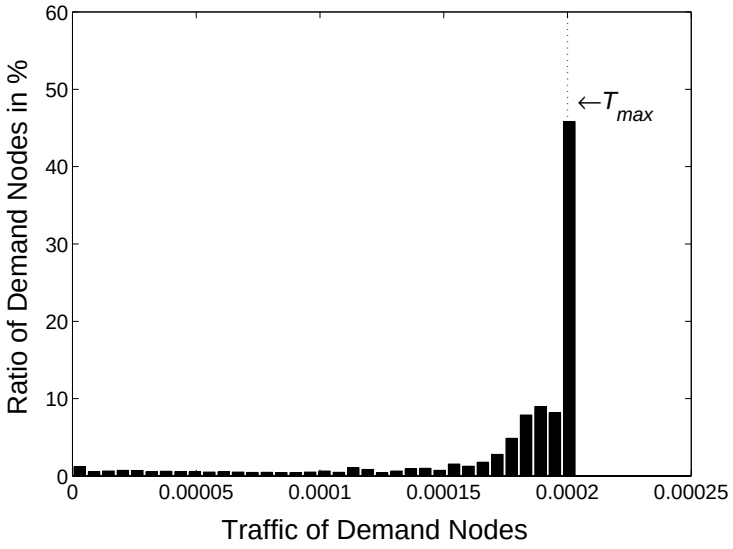


(b) with optimization

Figure 5.10: *Histogram of the traffic values for the demand node distribution of the Würzburg area by agglomerative clustering*



(a) without optimization



(b) with optimization

Figure 5.11: *Histogram of the traffic values for the demand node distribution of the Dallas/Fort Worth area by agglomerative clustering*

nificantly improved by the optimization. Moreover, the value of $T_{\max}$ is reached better since the granularity of elements of the traffic matrix is smaller.

Accuracy

The second criterion for the clustering algorithms is approximation accuracy of the spatial variation of the teletraffic in the studied area. The spatial statistics of the demand node distribution have to show the same properties as the actual teletraffic of the service area represented by the traffic matrix.

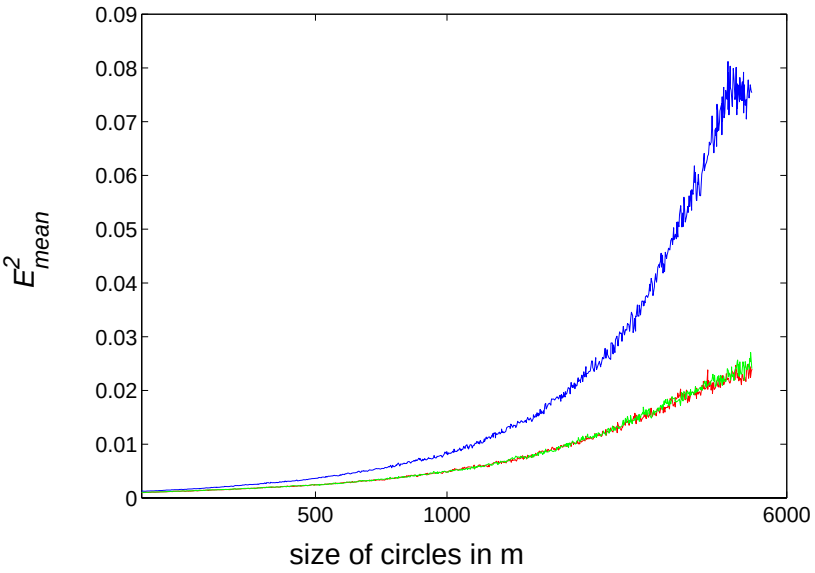
A common procedure to validate the spatial statistic of a discrete point pattern is the *quadrat sampling method*, cf. Cressie (1991). Quaddrat sampling involves collecting the traffic count in subsets of the studied area. Traditionally, the subsets are rectangular, therefore the name quadrat. Any other contour of the test sets is also possible. Since in cellular mobile communication systems the shape of a cell is more likely a circle, circular test squares are used for the validation of the traffic.

To evaluate the spatial accuracy of demand node distributions, a modification of the quadrat sampling method has been used. Circular test quadrats of random diameter have been placed randomly at the same location in the demand node distribution and in the study area, cf. Figure 5.13. The teletraffic value of the circular test quadrat in demand node distribution T_{nodes} was sampled by counting the teletraffic of the nodes. The traffic amount of the studied area T_{matrix} was obtained from counting the traffic from the elements of the teletraffic matrix. The matrices were obtained as described in Section 5.3.1. The accuracy of the demand node distribution was judged by the mean squared error, i.e. the difference of these teletraffic values for circular test quadrats with the same diameter:

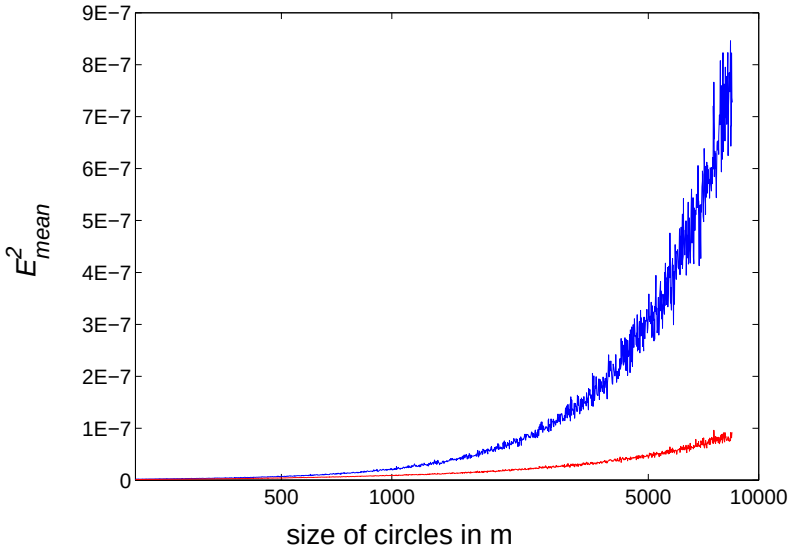
$$E^2_{mean} = \frac{1}{N} \sum_{i=1}^{N} (T^{(i)}_{nodes} - T^{(i)}_{matrix})^2, \tag{5.8}$$

here the parameter N is the number of circular test quadrats with the same diameter.

Figure 5.12 depicts the traffic approximation capabilities of the partitional and the agglomerative clustering method. Part (a) of Figure 5.12 shows the result for the Würzburg planning experiment. The blue line is the mean squared error for the partitional clustering algorithm. The



(a) Würzburg area



(b) Dallas/Fort Worth area

Figure 5.12: *Traffic approximation accuracy of the partitional and the agglomerative clustering method*

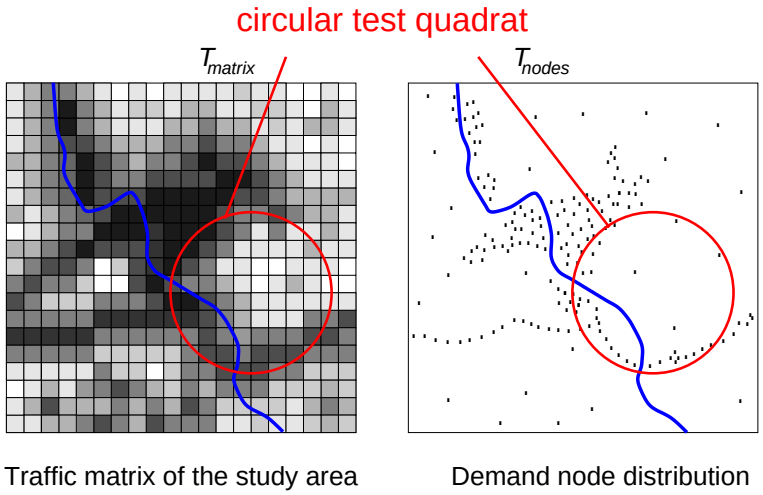


Figure 5.13: *Spatial validation*

green curve denotes the error for the agglomerative method without the optimization and the red line denotes the error of agglomerative method with optimization. The error of the agglomerative algorithm is always below the error of the partitional method. The optimization stage does not reduce the spatial accuracy. Similar results have been obtained in the Dallas/Fort Worth planning experiment, cf. Figure 5.12(b). The blue curve is the error of the partitional method and the red line the error of the agglomerative algorithm. Due to the smaller teletraffic granularity, the agglomerative method outperforms the partitional algorithm even stronger.

5.4 Impact of User Clustering on Cell Performance

In the previous section it was demonstrated that the user distribution in the service area of a mobile communication network is strongly clustered. In this section, it will be shown that the clustering effect has a significant impact on the performance of cell and has to be carefully considered during network design.

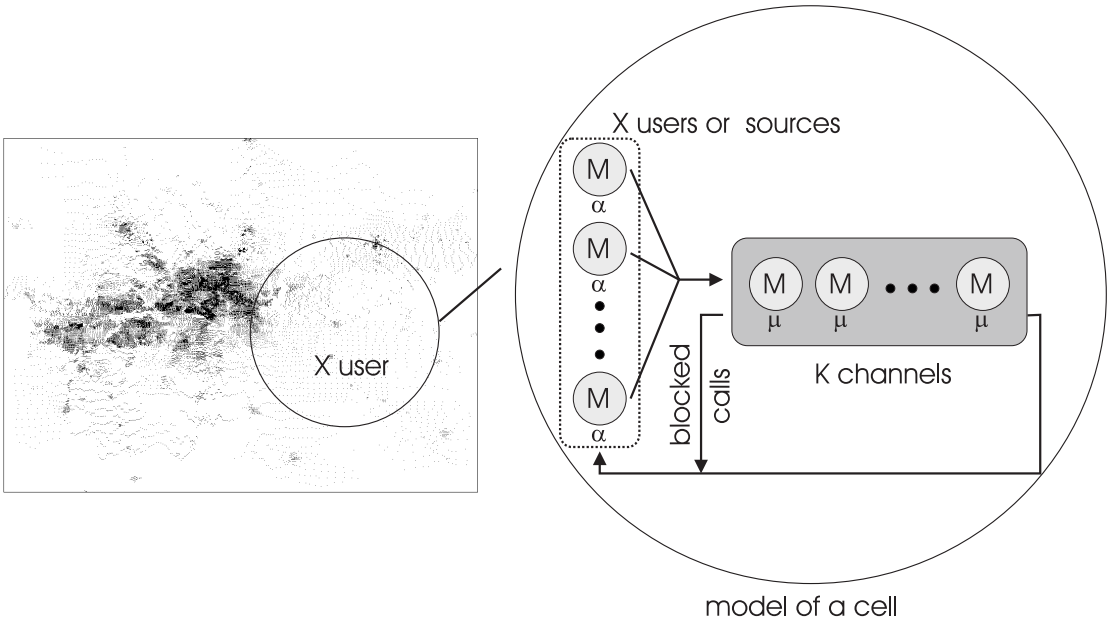


Figure 5.14: *Model of a cell in clustered user environment; enhanced version of a figure taken with permission from Tran-Gia and Gerlich (1996)*

5.4.1 Model Description

To evaluate the effect, a basic queuing system, a finite population m-server loss system, is investigated with respect to clustering. The model is often used in mobile and telecommunication network design, cf. Tran-Gia and Gerlich (1996). The examined performance value is the Quality-of-Service experienced subjectively by the test user. During the evaluation it was assumed that the clustering was neglected in system design.

The model considers a test user in a cellular network with a clustered structure described by a demand node distribution. The model is depicted in Figure 5.14. A circular-shaped omni-antenna is located in the studied area and forms a circular cell. A constant number of K radio channels are allocated to the cell. Due to the clustered structure, the number of users in the cell depends on the cell's location. This number is described by the random variable X.

The user traffic process in the time domain is modeled by a finite source model with two states. A user can either be in *Active* or *Idle* state. The Active state presents the use of a radio channel. The time intervals the user is in one of the two states are assumed to be exponentially

distributed. The random variable I with mean $1/\alpha$ denotes the sojourn time in the Idle state and the random variable B with mean $1/\mu$ the time interval in the Active state. After finishing a call, a user will stay in the Idle state until he generates the next call. If a call request is rejected, the user remains in the Idle state. If a call request is accepted, the user transfers to the Active state. He remains in this state during the connection holding time. At the end of a connection, the user moves back to the Idle state.

The offered traffic intensity of a user is $a^* = \alpha/\mu$ and equal for every customer. This parameter is often used as input in mobile system engineering and is measured in Erlangs. Additionally, the user activity factor, i.e. the probability that a user is active when the blocking effect would be neglected, is given by:

$$\rho_M = \frac{a^*}{1 + a^*} = \frac{\alpha/\mu}{1 + \alpha/\mu}. \tag{5.9}$$

5.4.2 Subjective Quality-of-Service in a Clustered Environment

The subjective Quality-of-Service in a clustered environment is defined as the call blocking probability seen by an arbitrary test user in the studied area.

The probability of the test user to be in a cell of $X = i$ users is given by, cf. Tran-Gia (1988):

$$x^*(i) = \frac{ix(i)}{E[x]}, \tag{5.10}$$

where $x(i) = P(X = i)$ is the probability of having a number of $X = i$ users in the cell and the expected value of X.

The conditional blocking probability $p_B(i)$ of a test user being in a cell with $X = i$ users is obtained using the well-known Engset-formula, cf. Kleinrock (1975):

$$\begin{aligned} p_B(i) &= P(\{\text{test user rejected} \mid X = i\}) \\ &= \frac{\binom{i-1}{K}(\frac{\alpha}{\mu})^K}{\sum_{k=0}^{K}\binom{i-1}{k}(\frac{\alpha}{\mu})^K} \end{aligned} \tag{5.11}$$

Taking Eqn.(5.10) and Eqn.(5.11) together, the blocking probability p_B of an arbitrary test user, which is the subjective QoS described above, is given by:

$$p_B = \sum_{i=K+1}^{\infty} p_B(i) x^*(i) \tag{5.12}$$

5.4.3 Performance under User Clustering

For the evaluation of the clustering effect, the above described model is considered in the context of the Dallas/Fort Worth planning scenario, cf. Section 5.3.1.

To conduct the calculation of the subjective Quality-of-Service, the empirical probability $x(i)$ is sampled from the demand node distribution of the Dallas region. A number N of circular-shaped test cells are placed in the study area. The number of users in a test cell X_j is obtained by counting its traffic T_j, represented by the demand nodes, and by considering the assumed traffic intensity per user:

$$X_j = \frac{T_j}{a^*} \quad j \in \{1, \ldots, N\}. \tag{5.13}$$

Hence, the empirical probability of having i users in a cell is given by:

$$p(i) = P(X = i) = M(i)/N \tag{5.14}$$

where $M(i) = \mid \{1 \leq j \leq N \mid i - 1 < X_j \leq i\} \mid$ is the number of test cells containing i users.

The radii of the test cells are selected such that the expected mean value of the number of users is fixed at $E[X] = 30$ for a given user activity value.

The system behavior without clustering is evaluated assuming that $p(i)$ is either deterministic or Poisson distributed. The deterministic case can be described with the users being uniformly located in the study area. The Poisson case is equivalent with the users being randomly distributed. In both cases, the user population is regarded to be not clustered, however the spatial statistics of the point pattern are different, cf. Stoyan and Stoyan (1992).

Figure 5.15 depicts the subjective Quality-of-Service for a test cell with $K = 7$ channels. This number is typical for a GSM system. The gen-

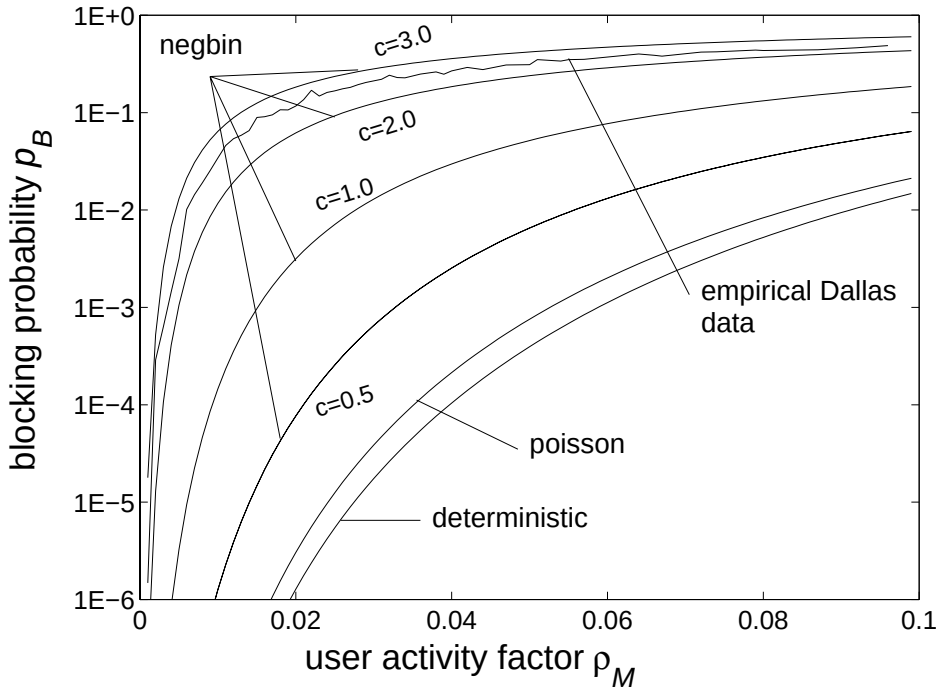


Figure 5.15: *Impact of user clustering on subjective Quality-of-Service* ($E[X] = 30, K = 7$)*; enhanced version of a figure taken with permission from Tran-Gia and Gerlich (1996)*

eral behavior of the system is similar for all types of users distributions. However, it can be clearly seen that the subjective Quality-of-Service in the clustered Dallas environment degrades much faster than without the clustering assumption. The blocking probability is two orders of magnitude higher as in the Poisson case. The difference between the deterministic case and the Poisson case is quite small. Thus, this example demonstrates that the user clustering effect significantly reduces system performance. The effect has to be considered during network design and traffic engineering.

The performance of the model in the Dallas scenario resembles a system where the number of users in a test cell follows a negative-binomial distribution. However, it is not possible to conclude that the system behavior can be fit by a model with two moments since the behavior in the Poisson case with the same mean and coefficient of variation differs significantly. The type of the distribution has to be taken into account.

5.5 Application of Demand Node Concept

Due to the efficient capability to model the spatial teletraffic variation in the service area, the framework of the demand node concept and the demand node distributions facilitate a large variety of planning and traffic engineering tasks in telecommunication networks. the demand node concept enables a *revenue-oriented* system design.

The main area of application of the concept is the demand-oriented locating of transmission facilities, e.g. base stations or switching centers. In wireless systems, providing a mobile service to a customer can be regarded as supplying a demand node with a sufficient radio signal. Hence, the base stations have to be located such that the number of nodes in the service range is maximized. Under this view, the transmitter locating task becomes equivalent to the *set cover* problem which is well known in facility location science, cf. Drezner (1995). A comprehensive description of automatic radio network design and optimization using the demand node concept is presented in Chapter 6. An extension of the approach was suggested by Mathar and Niessen (1997). The proposed method uses a demand node distribution to optimize the locations of the base stations as well as for the allocation of the frequencies to the cells in an FDMA cellular system.

Various studies have shown that the multiplexing and access technologies applied in "third generation" mobile systems such as UMTS, cf. Grillo et al. (1995) and Schwarz da Silva et al. (1997), require to consider the customer population carefully during network planning. Especially the performance of CDMA (code division multiple access) systems and of SDMA (space division multiple access) systems is highly sensitive on the spatial user distribution. An analysis of the impact of clustering on the performance CDMA networks can be found in Tran-Gia et al. (1998) and Leibnitz et al. (1998). A detailed evaluation of SDMA systems considering the clustering is presented in Gerlich (1997) and Remiche (1998).

However, most of the above mentioned studies utilize theoretical planar point processes to approximate the clustering. Whereas in practical network design real world customer distributions should be considered. The application of the demand node concept enables the matching of practice and theory, cf. Grillo et al. (1998). As a result, a first method for locating of transmitters in a CDMA network using a set cover approach was presented in Yu et al. (1998).

Future applications of the demand node concept are not limited

solely to mobile communication systems. The concept can also be used in fixed network planning scenarios where transmission facilities have to be deployed in a service area, such as in wireline data networks. In ADSL (Asymmetric Digital Subscriber Line) access networks, the DSLAM (Digital Subscriber Line Access Multiplexer) has to be located within a certain distance from a customer, cf. Maxwell (1996). The potential customer locations can be approximated using the demand node concept.

Other applications of discrete point patterns in wireline system engineering encompassing the estimation of cable geometry and average length in rural and urban areas, cf. Grahovac and LeBourges (1996) and the design of hierarchical long-distance switched networks, cf. Baccelli et al. (1997).

5.6 Concluding Remarks

This chapter presented a new and complete framework for the estimation and characterization of the expected spatial demand distribution in mobile communication systems. The proposed methods considers the teletraffic from the viewpoint of the network. Its traffic estimation is based on the *geographic traffic model*, which obeys the geographical and demographical factors in the service area for the teletraffic demand estimation. The characterization of the spatial distribution is facilitated by the application of discrete points, denoted as *demand nodes*. Additionally, it was demonstrated how the demand node pattern can be derived from the information in publicly available geographical data bases. For the generation of the demand node distributions, two clustering algorithms have been introduced, a *recursive partitional clustering* algorithm and an *agglomerative clustering* method, and their traffic description capability was examined. Furthermore, it was shown that, even for a simple queuing model, user clumping leads to a decreased system performance. Customer clustering has to be considered during the design of a spatial extended communication network.

However, since in practical cellular system design often only rather insufficient geographical and demographical information is available, it would be of great interest to generate demand node patterns artificially by the application of spatial point processes, like the processes described by Baum (1998), Latouche and Ramaswami (1997), Stoyan and Stoyan (1992), and Cressie (1991). Furthermore, these reference processes could

be used to evaluate different planning scenarios.

The proposed framework of the demand node concept enables the matching of practice and theory in mobile network planning. The methods suggested in this chapter have been submitted to the ITU-T's focus group on traffic engineering for personal communications (FG-TEPC). FG-TEPC is considering how to harmonize the Demand Node framework with the procedures described in ITU-T E.750 series draft recommendations on terminal mobility.

6 Demand-oriented Radio Network Synthesis

The core task of cellular system design is to set up an optimal radio network, which provides the best feasible *coverage* of the investigated planning region. During the design process, this aim is achieved by an optimal selection of base station sites and the perfect determination of the basic RF parameters, like the maximal base station transmitting power, the antenna height, or the number and the orientation of the sectors.

In the past, the major design criterion of cellular networks was the *area coverage* like in the so-called "analytical approach", cf. Chapter 4. The design mainly focused on providing the best possible radio signal at *every* location of the planning region. Capacity aspects were almost neglected or addressed only later stages of the planning process.

In contrast to this, in economic network design, the cost of providing the service has gained greater and greater importance. *Demand coverage* can be viewed as *revenue coverage*. Thus, this criterion has to be considered as a key engineering constraint in cellular mobile system engineering.

The disadvantages of the conventional approach, of being mainly focused on RF issues, can be overcome by the application of the demand-adaptive *integrated approach* to cellular system design, cf. Chapter 4. The application of the new approach is mainly facilitated by the use of the *demand node concept (DNC)*, cf. Chapter 5. Due to this concept, it is possible to include efficiently the demand distribution in radio network engineering. The transmitter site selection design task can be formulated as a problem of the class of *set covering problems (SCP)*, which are well-known in economics to model and solve *facility location problems*, cf. Ghosh and McLafferty (1987). Thus, the *integrated approach* enables

a demand-based cellular planning methodology and is capable to address the optimization problems emerging in cellular radio network design.

This chapter is organized as follows. In Section 6.1 numerous other approaches to automatic transmitter locating are reviewed. Section 6.2 is devoted to the formulation of the transmitter locating task as a discrete set covering problem. Furthermore, the computational complexity of set covering models is investigated. Unfortunately, set covering problems belong to the class of *NP-hard* optimization tasks, which are well-known as notoriously intractable. Therefore, Section 6.3 introduces efficient approximation algorithms which are capable of obtaining a feasible solution for covering models in reasonable time. In Section 6.4, the planning tool demonstrator ICEPT, which implements the proposed methods is presented. In Section 6.5 a second important constraint in radio network design is investigated: the interference minimization. The set covering model is extended to optimize both the demand coverage as well as the minimization of the co-channel interference. Section 6.5 summarizes this chapter and gives an outlook to further extensions of the proposed method.

6.1 Automatic Transmitter Locating Algorithms

Since the design complexity of mobile cellular networks, in the past, has drastically increased, automatic network design algorithms came into the focus of various research groups. In the following section, the major approaches to this problem are reviewed.

6.1.1 Adaptive Base Station Posititioning Algorithm

The *Adaptive Base station Positioning Algorithm (ABPA)* was first presented by Fritsch and Hanshans (1993) and extented by Fritsch et al. (1995). It already uses an early version of the demand node concept, cf. Chapter 5. This was one of the first methods that considers, parallel to the RF objectives, the expected traffic as a direct constraint for the location of the cell site. ABPA is based on the idea of *competing base stations* which try to cover as many demand nodes as possible. The algorithm determines two parameters of transmitters, the position and

the transmitting power. To locate the base stations, ABPA shifts the transmitters around in the virtual scenario. The movement of the base station is conducted by the attraction of base stations to not covered demand nodes, cf. Figure 6.1(a) and by the repellation of base stations from multiply supplied nodes. cf. Figure 6.1(b). In a similar way, the algorithm adapts the power level of the transmitters. To prevent ABPA from getting stuck in locally optimal configurations, the algorithm performs transitions to a worse configuration with a predefined probability as in Simulated Annealing, cf. Aarts and Korst (1990). The major drawback of ABPA is its speed. The shifting of the transmitters requires the coverage to be recalculated after every transition.

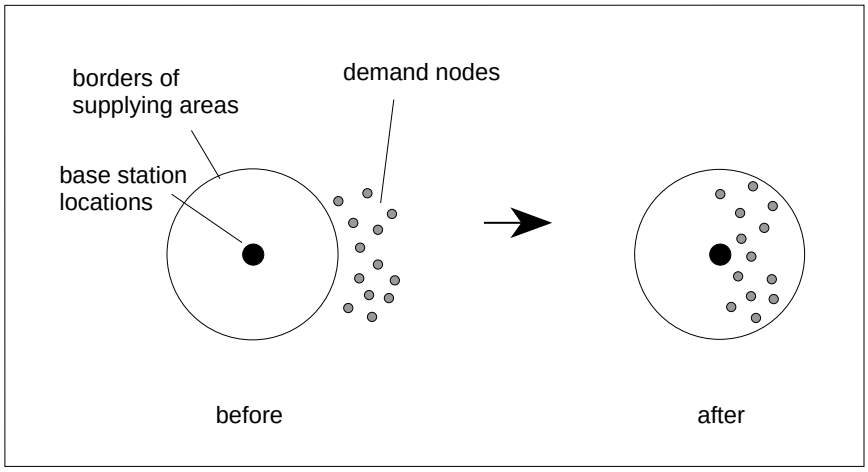
6.1.2 Other Approaches

A promising approach to automatic network design was presented by Chamaret et al. (1997). The radio network design task is modeled as a *maximum independent set* search problem. However, the approach only addresses RF design aspects. It uses the *area coverage per base station* as the objective function of the optimization. Other design constraints are not considered. A similar approach, using a genetic algorithm, was investigated by Calégari et al. (1997).

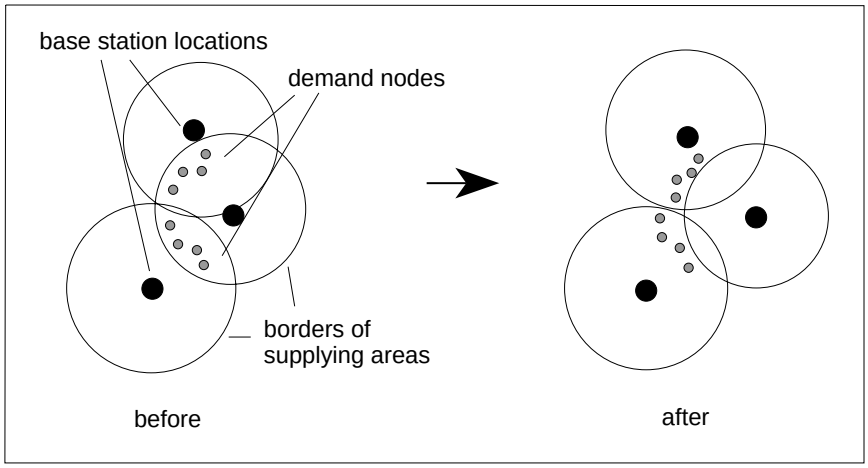
In contrast to this, an algorithm which considers only the traffic distribution as a constraint for cell site locations was proposed by Ibbetson and Lopes (1997). The algorithm uses an computational geometry approach and constructs a tessellation of the planning region with a k-D-tree-like search algorithm.

Two methods have been proposed for the design of indoor wireless systems, cf. Fortune et al. (1995) and Wright (1998), and micro-cellular radio communication systems, cf. Sherali et al. (1996). Both approaches focus on RF constraints, since in an indoor and micro-cellular environment, so far, network capacity is not of major importance due to the small number of users.

A set covering optimization model, similar to the method that will be introduced in Section 6.2, was independently investigated by Floriani and Mateus (1997). The set covering model was solved using *integer programming* and *mixed integer programming*. Two objective functions have been investigated by the authors, the total area coverage and the maximal spatial channel utilization. The second objective was introduced in order to address the co-channel interference problem, see also Section 6.5. However, no assumption on the teletraffic distribution has



(a) Attraction of transmitters



(b) Repellation from multiple covered nodes

Figure 6.1: *Adaptive base station positioning algorithm (ABPA), cf. Fritsch et al. (1995)*

been made. The application of Integer Programming imposed a strong limitation on the size of the investigated service area.

6.2 Coverage Models in Cellular Mobile Network Planning

The assumption underlying all coverage models for the facility location problem is that the accessibility to a service network within some critical range is an essential criterion of the service use. Customers beyond the access range are not adequately served by the service facilities and are therefore not likely to utilize the offered service. In planning a network of *service centers*, a crucial objective is to maximize the proportion of the demand within the specified range of the service facilities, cf. Ghosh and McLafferty (1987). In contrast to usual facility location problems, the term *range* implies, in the context of wireless network planning, not the physical distance between the demand node and the transmitter. Here, the term *range* denotes the path loss L of the radio signal strength between the base station and the receiving node, cf. Parsons (1992).

This covering approach taken together with the demand node concept which is used to characterize the demand distribution in the service area, cf. Chapter 5, enables a new and demand-oriented definition of the term *coverage area* of a given transmitter:

Definition 6.1 : Coverage Area

The *coverage area* of a transmitter is the set of demand nodes which are provided with a usable radio link, i.e. the path loss from the transmitter to the demand node is less than a certain threshold.

Of course, this definition is aimed at an economic approach to cellular radio network engineering, since the demand distribution is a core ingredient of the new notion. Nevertheless, the conventional design objective in cellular engineering of providing a radio signal at any time at any location is not completely disregarded. In Section 5.3 it was outlined that it is feasible to limit the area size represented by a demand node without loosing the ability to accurately characterize the traffic distribution in the service area. Hence, demand nodes can be distributed in the service in such a way that a significant characterization of the path loss in the service area is possible. Experiments have shown, see Section 6.4.4, that the application of Definition 6.1 together with the demand node concept is a valid *approximative model* for the transmitter locating task.

6.2.1 Minimum Set Covering Model

The objective of the set covering model in cellular engineering is to determine the number of required service centers, i.e. base stations, and their location such that *all* users of the wireless network are served with an adequate service level, i.e. field strength level. This model ensures that the service level at each demand node is greater than a specified level if a solutions exists.

In mathematical terms, the *minimum set covering problem (SCP)* is defined as follows, cf. Ghosh and McLafferty (1987):

$$\text{Minimize} \quad Z = \sum_{j \in J} x_j \tag{6.1}$$

subject to:

$$\sum_{j \in N_i} x_j \geq 1 \quad \forall j \in J, \forall i \in I \tag{6.2}$$

where

Z number of facilities;

J set of potential facility sites (indexed by j);

I set of demand nodes (indexed by i);

$x_j = \begin{cases} 1 & \text{if facility at } j \\ 0 & \text{otherwise} \end{cases}$;

N_i $\{j \mid f_{ij} \leq L\}$; the set of base stations j located within the standard range L of demand node i;

f_{ij} the path loss between demand node i and potential base station location j; the dimension of $F = (f_{ij})$ is $\mid I \mid \times \mid J \mid$;

L the maximal eligible path loss for the mobile service utilization

The objective in Eqn. (6.1) is to minimize the number of required facilities and the constraint in Eqn. (6.2) enforces that, in a valid solution, each demand node is covered at least once within the standard range

L. In the considered set covering model, potential facility locations are equivalent to the location of demand nodes. In mobile network planning this is not necessarily true. Transmitters can be located only at positions where real estate can be purchased or leased. However, it can easily be shown that this constraint can be obeyed by the SCP. Furthermore, as introduced in Section 3.3, the objective function should not only comprise the demand coverage as the sole objective. The cost-effectiveness has to be addressed also in an appropriate way. Therefore, a more detailed discussion on the objective cost function is presented further below.

6.2.2 Maximal Covering Location Problem

A valid solution of the SCP requires coverage of all demand nodes, no matter how cost-effective the solution is. However, for an economic design of wireless communication networks a trade off between the cost of coverage and the benefit resulting from covering this area is desired. This objective leads to the definition of the transmitter location problem as a location problem that does not require the coverage of *all* demand nodes. Church and ReVelle (1974) define this problem as the *maximal coverage location problem (MCLP)*. The MCLP assumes a limited budget and includes this point as a constraint on the number of facilities to be located. Thus, the optimization tries to place a fixed number of base stations p so that the proportion of demand nodes covered by the cells within the permitted range is maximized. The mathematical definition of the MLCP is:

$$\text{Maximize} \quad Y = \sum_{i \in I} a_i y_i \tag{6.3}$$

subject to:

$$\sum_{j \in N_i} x_j \geq y_i \;\; \forall i \in I \;\; \wedge \;\; \sum_{j \in J} x_j = p \;\; \forall i \in I, \forall j \in J. \tag{6.4}$$

As additional notation for the MLCP we use:

$$
\begin{array}{ll}
Y & \text{the weighted coverage;} \\
y_i & = \begin{cases} 1 & \text{if demand node } i \text{ is covered} \\ 0 & \text{otherwise} \end{cases}, \\
p & \text{the number of base stations to be deployed, and} \\
a_i & \text{the population at demand node } i.
\end{array}
$$

All other variables and parameters are the same as defined in the SCP. The objective of Eqn. (6.3) is to maximize the sum of covered demand nodes. The first constraint in Eqn. (6.4) states that demand node i cannot be covered unless at least one server is located within the standard range L. The second constraint of Eqn. (6.4) defines the budget constraint by forcing the number of placed base stations to be exactly p.

Due to the factors a_i, the MCLP maximizes the weighted objective coverage function, cf. Eqn. (6.3). In the demand node concept the weights a_i , which represent the traffic value, are equal for all nodes i, cf. Definition 5.1. Using different values for the weights a_i, however, would introduce a prioritization of certain nodes. This can be used to favor the coverage of important areas within in the planning region, for example airports or train stations.

6.2.3 Complexity of Covering Problems

Set covering models belong to a class of very intractable problems, denoted as *NP-hard*. So far, no *efficient* algorithms have been found for solving them. In the next section, the impact of this feature on finding a feasible solution for the transmitter location task is outlined. Before presenting the results, some basic terms and results from complexity theory are introduced first, cf. Horowitz and Sahni (1978).

The Classes NP-Hard and NP-Complete

In theoretical computer science, the complexity of an algorithm is characterized by the order of the magnitude of its computing time complexity, respectively using its asymptotic notation. The hardness of a problem is expressed by a function $f(n)$ restricting the number of steps needed to solve the problem. The number of steps is usually measured by the length n of the input. Hence, $T(n) = O(f(n))$ denotes that the computing time of the algorithm is bounded above by the function $f(n)$.

Another important idea in complexity theory is the distinction between problems whose solution can be obtained by a polynomial time algorithm and problems for which no polynomial time algorithm is known. An algorithm is said to be a *polynomial time algorithm* if there is a monomial $f(n) = Cn^k$ such that its run time is bounded above by $f(n)$. These algorithms are denoted as *efficient* algorithms.

Especially the class of non-deterministic polynomial time algorithms is important. Algorithms whose computing time is greater than polynomial, require a vast amount of time to obtain the solution and even moderate sized problems cannot be solved in reasonable time.

General proofs of the classification of optimization problem are very difficult to obtain. Therefore, the complexity theory instead uses the concept of *decision problems* with "yes" or "no" answers. At first a glance, the idea of a decision problem appears very restrictive. However many optimization problems can be easily recast into decision problems with the property that the decision problem can be solved in polynomial time iff the corresponding optimization problem can. Using the concepts outlined above, it is possible to define the most basic classes of problems:

Definition 6.2 : The class P

P is the set of all decision problems solvable by a deterministic algorithm in polynomial time.

Definition 6.3 : The class NP

NP is the set of all decision problems solvable by a non-deterministic algorithm in polynomial time.

For a definition of the term *non-deterministic algorithm*, the reader is conferred to Garey and Johnson (1979). Using the concept of **P** and **NP**, it is possible to define the *NP-hard* and *NP-complete* classes of problems. However first, it is necessary to define the notion of reducibility.

Definition 6.4 : L_1 reduces to L_2 ($L_1 \propto L_2$)

Let L_1 and L_2 be decision problems. L_1 **reduces to** L_2 ($L_1 \propto L_2$) iff there exists a polynomial time computable function f such that for any input $x, x \in L_1 \Leftrightarrow f(x) \in L_2$ holds.

This definition implies that if there is a polynomial time algorithm for L_2 then L_1 can be solved in polynomial time.

Definition 6.5 : NP-complete

A problem L_2 is defined to be **NP-complete** if $L_2 \in \mathbf{NP}$ and for all problems $L_1 \in \mathbf{NP}$, $L_1 \propto L_2$.

Using the definitions introduced above it is easy to show, that if any single NP-complete problem can be solved in polynomial time, then *all* problems in NP can be solved. If any NP-complete problem L is intractable, then so are all NP-complete problems, cf. Garey and Johnson (1979)

Definition 6.6 : NP-hard

A problem L_1 is defined to be **NP-hard** if for all other problems $L_2 \in \mathbf{NP}$, $L_1 \propto L_2$.

It is easy to see that there are NP-hard problems that are not NP-complete. Only a decision problem can be NP-complete. However, an optimization problem may be NP-hard. If L_1 is a decision problem and L_2 an optimization problem, it is quite possible that $L_1 \propto L_2$.

Classification of the Set Cover Problem

The main interest of this section is to classify the complexity of the set cover problem. In complexity theory, the common procedure to categorize an algorithm, is to exploit the concept of polynomial time reducibility. Thus, the classification procedure for problem L consists of three steps: *a)* selecting a known NP-complete problem L', *b)* constructing a transformation function f from L' to L, and *c)* proving that f is a polynomial transformation.

Since the complete sequence of proofs for the classification of the set cover problem is rather long and well documented in literature, cf. Garey and Johnson (1979), Papadimitriou (1994), and Hochbaum (1996), this section only outlines the classification procedure and presents the most important results. Thus, a more practically interested reader should be able to follow the results and judge their impact on the base station location task without being bothered by the details. A more theoretically-oriented person should use this summary as a guideline to the actual proofs, which can be found in the above mentioned literature.

The road map for the classification of the set cover problem is depicted in Figure 6.2. The initial transformation will be from the SATISFIABILITY problem to the 3SAT problem. In the following step 3SAT

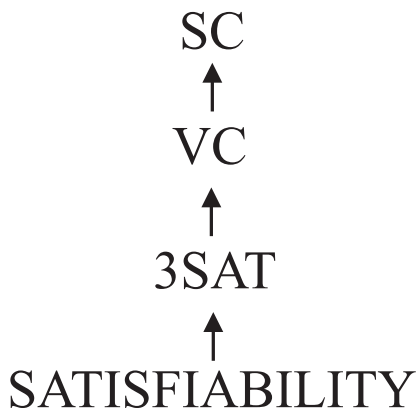


Figure 6.2: *Diagram of the sequence used to prove that the Minimum Set cover decision problem is NP-complete*

will be transformed to the Vertex Cover (VC) problem. The Set Cover (SC) problem is a special case of the VC problem.

The most basic NP-complete problem is the decision problem from Boolean logic, which is usually referred to as the SATISFIABILITY (SAT) problem.

Definition 6.7 : **SATISFIABILITY (SAT) problem**

Let $U = \{u_1, u_2, \ldots, u_n\}$ denote boolean variables. Let $\bar{u_i}$ denote the negation of u_i. A literal is either a variable or its negation. A boolean formula C is represented as $\bigwedge_{i=1}^{k} c_i$. The c_i are clauses each represented by $\bigvee l_{ij}$. The l_{ij} are literals. The **SATISFIABILITY (SAT)** problem is defined as the decision problem whether there is an assignment for U, such that all clauses in C are simultaneously satisfied.

Theorem 6.1 : **Cook's Theorem**

SATISFIABILITY is **NP**-complete.

Proof: see Garey and Johnson (1979), pp. 38. □

The 3-SATISFIABILITY problem is just a restricted version of the SATISFIABILITY in which all clauses consists of exactly three literals. Its simple structure makes it one the most widely used problems for proving other NP-completeness results.

Definition 6.8 : 3-SATISFIABILITY (3SAT) problem

Let $U = \{u_1, u_2, \ldots, u_n\}$ denote boolean variables. Let $C = \bigwedge_{i=1}^{k} c_i$ be a boolean formula, where c_i are clauses of literals. Let $c_i = l_{i_1} \vee l_{i_2} \vee l_{i_3}$ for $1 \leq i \leq k$. The **3-SATISFIABILITY (3SAT)** problem is defined as the decision problem whether there is an assignment for U, such that all clauses in C are simultaneously satisfied.

Theorem 6.2 :

3-SATISFIABILITY is **NP**-complete.

Sketch of the Proof: Since 3-SATISFIABILITY is a special case of SATISFIABILITY, it is in **NP**. To show NP-completeness of 3-SATISFIABILITY, it is required to demonstrate that there is a polynomial time transformation from SATISFIABILITY to 3-SATISFIABILITY. Therefore it is necessary to construct a new boolean formula C' with three literals per clause, such that C' is satisfiable if and only if C is. This can be obtained by examining the clauses of C one after each other and by replacing each c_i by an equivalent set of clauses, each with three literals. The detailed construction of the new clauses can be found in Papadimitriou and Steiglitz (1982), p. 359.

The construction of C' can obviously be carried out in polynomial time. Thus, a polynomial time transformation from SATISFIABILITY to 3-SATISFIABILITY has been found. Hence, 3-SATISFIABILITY is NP-complete. □

The next step in showing that the set cover problem is NP-hard, is to find a transformation of the 3SAT problem to the Vertex Cover problem:

Definition 6.9 : VERTEX COVER (VC) problem

Let $G = (V, E)$ be a graph, where V is a set of vertices and E is a set of edges. Let K be a positive integer with $K \leq \mid V \mid$. The **VERTEX COVER (VC)** problem is defined as the decision problem whether there is a vertex cover of size K or less for G, that is, a subset $V' \subseteq V$ such that $\mid V' \mid \leq K$ and, for each edge $\{u, v\} \in E$, at least one of u and v belongs to V'.

Theorem 6.3 :

VERTEX COVER is **NP**-complete.

Sketch of the Proof: It is obvious that VC $\in$ **NP**, since a non-deterministic algorithm can find a subset of vertices and check in polynomial time whether that subset contains at least one end point of every edge and has size less equal then K. Now let $U = \{u_1, u_2, \ldots, u_n\}$ and C an instance of 3-SATISFIABILITY. 3SAT can be transformed to VERTEX COVER by constructing a graph $G = (V, E)$ and a positive integer $K \leq \mid V \mid$ such that G has a vertex cover of size less equal K if and only if C is satisfiable. This can be obtained by regarding a truth-setting component $T_i = (V_i, E_i)$ for each variable $u_i \in U$ and a satisfaction testing component $S_j = (V_j', E_j')$ for each clause $c_j \in C$. These components are augmented by some additional communication edges to form the graph. For details of the construction, the reader is referred to Garey and Johnson (1979), pp. 54.

Obviously, the construction of the graph can be achieved in polynomial time. All that remains to be shown is that C is satisfiable if and only if G has a vertex cover of size less equal K.□

The VERTEX COVER problem is a special case of the SET COVER problem. The vertices of G have to be covered by exactly one of two existing sets. These two sets correspond to the endpoints of an edge in a bipartite graph. This relationship is illustrated in Figure 6.3. Part (a) of Figure 6.3 depicts four base stations and their corresponding supplying areas. Part (b) of Figure 6.3 shows an equivalent bipartite graph representing the transmitters and their supplying areas.

Hence, it is time to define the SET COVER decision problem.

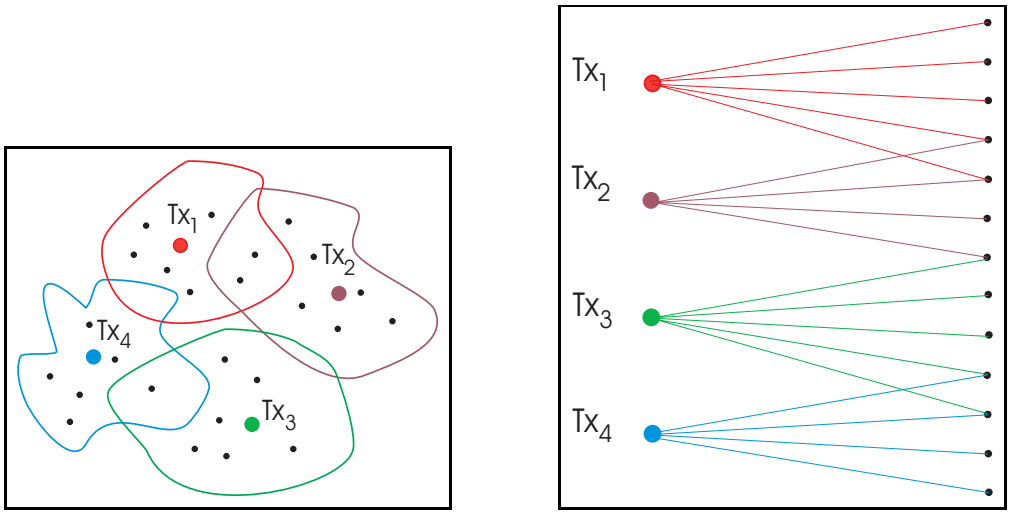
Definition 6.10 : SET COVER (SC) problem

Let C be a collection of subsets of a finite set S and $K \leq \mid C \mid$. The **SET COVER (SC)** problem is defined as the decision problem whether C contains a cover for S of size less equal K. That means, does there exist a subset $C' \subseteq C$ with $\mid C' \mid \leq K$ such that every element of S belongs to at least one member of C'.

Theorem 6.4 :

SET COVER is **NP**-complete.

Sketch of the Proof: It is obvious that SC $\in$ **NP**, since a non-deterministic algorithm only needs to guess a subset of C and check in polynomial time whether that subset contains at least every element of S and has



(a) (b)

Figure 6.3: *Modeling the set cover as a bipartite graph.*

size less equal then K. Now let $G = (V, E)$ be a bipartite graph and V a vertex cover. The transformation from VERTEX COVER to SET COVER can be obtained by constructing sets for every vertex in V to every endpoint of the edges originating from the vertex.

It is easy to see that this transformation can be performed in polynomial time.□

Since the SET COVER decision problem cannot be solved in polynomial time, unless $\mathbf{P} = \mathbf{NP}$, the set cover optimization problem (SCP) cannot be either. Thus, the SCP is *NP-hard*.

The Maximum Coverage Location Problem

The maximum coverage location problem (MCLP) is a special case of the SCP. Hence, MCLP is clearly NP-hard, as set cover is reducible to it. A polynomial time algorithm which would solve the maximal covering location problem could easily solve the set cover problem, cf. Hochbaum (1996).

Conclusion

set covering models are powerful tools to approach the site selection task in the cellular radio network engineering. The intuitive view of a base station supporting a finite number of users in the coverage area efficiently facilitates the modeling of the problem as a well known combinatorial optimization task. The mathematically precise definition of the problem enables the exact determination of an optimal configuration. However, it was shown that the class of set covering problems is *NP-hard*. No efficient, i.e. polynomial time, algorithm has yet been found for solving them and it is very unlikely that such an algorithm will be found in the next future. Thus, selecting the optimal solution out of a large number of potential candidate configurations still requires a vast amount of computing time. A promising possibility of resolving this dilemma is the application of an *approximation algorithm* and will be discussed in the next section.

6.3 Approximation Algorithms for Covering Problems

From a practical viewpoint, if the optimal solution is unattainable then it is reasonable to sacrifice optimality and settle for a *good* feasible solution that can be computed efficiently. Of course, it is obvious that optimality should only be restricted as little as possible, while gaining as much as possible in efficiency. Trading-off optimality in favor of tractability is the paradigm of *approximation algorithms*.

Approximation algorithms are programs which generate feasible solutions that are *close* to the value of an optimal one. Especially, algorithms running in polynomial time are of important interest. In this section, three approximation algorithms for the transmitter location optimization task are introduced: *a)* two Greedy-based heuristics (GRH) and *b)* a Simulated Annealing heuristic (SAH).

6.3.1 Greedy Heuristic Solutions

In oder to motivate a Greedy-like heuristic algorithm for solving the set covering problems in radio network engineering, a more detailed set theory definition of the SCP according to Chvatal (1979) will first be given.

In the SCP, the data consists of finite sets $P_1, \ldots, P_n$ and positive numbers $c_1, \ldots, c_n$ which denote the cost of using the sets P_j. In the set covering approach, these sets correspond to the supplying areas of transmitters, cf. Definition 6.1. For the SCP, the union of all P_j is $I = \bigcup(P_j : 1 \leq j \leq n)$ and $I = \{1, \ldots, m\}$ and $J = \{1, \ldots, n\}$. A subset J^* is called a *cover* if:

$$\bigcup_{j \in J^*} P_j = I; \tag{6.5}$$

The cost of this cover is $c(J^*) = \sum_{j \in J^*} c_j$. The problem is to find a cover J^* with minimum cost.

Greedy Heuristic for the Set Covering Problem

The original Greedy heuristic algorithm assumes that the desirability of using the set j in an optimal cover increases with ratio $\mid P_j \mid /c_j$, i.e. the better the ratio of covered demand nodes per cost unit for a set j, the higher the probability that the set is in the optimal cover J^*. This expectation suggests a recursive procedure for finding near-optimal covers. The Greedy algorithm selects, one after the other, the sets which provide the best additional coverage per cost ratio. Therefore, the term *Greedy* is used for the heuristic.

The complete Greedy algorithm for the SCP is depiced in Algorithm 6.1. It starts with an initialization of the solution variable in line *3*. Hereby, J^* denotes the set of selected transmitter configurations.

The main philosophy of the Greedy heuristic is located in line *6*. Here, the algorithm selects the transmitter configuration k with the best coverage per cost-unit ratio. Lines *9* and *10* are an adjustment step. The selected base station k is inserted in the solution J^*. In line *9*, the remaining sets P_j, with $j \neq k$, are updated. The demand nodes already covered by the selected base station k are removed from the sets that have been not selected yet. In this way, the number of multiple covered nodes is minimized and the overall number of covered nodes is increased. These steps are repeated until all nodes are covered. Hence, the algorithm stops, returning the solution J^*.

It is obvious that Algorithm 6.1 can be computed in polynomial time. For every selected configuration the Greedy heuristic requires only one sweep through the set of all candidate configurations.

Algorithm 6.1 (Set Cover Problem)

variables:

J	*set of indices of all transmitter configurations*
J^*	*set of indices of transmitter configuration in cover*
P_j	*cover of transmitter configuration* j
c_j	*cost of transmitter configuration* j

algorithm:

```
 1  proc set_cover() ≡
 2    begin
 3      J* = ∅
 4      if ∀j ∈ J : P_j = ∅
 5        then stop        /* J* is a cover */
 6         else find k with P_k/c_k is max.
 7              goto 10
 8      fi
 9      ∀j ∈ J : P_j = P_j − P_k
10      J* = J* + k
11      goto 5
12    end
```

Algorithm 6.1: Greedy heuristic for solving the set cover problem

The set theory definition of the SCP reveals directly the major shortcoming of this approach. The union of all the sets P_j, with $\forall j \in J^*$, has to comprise every demand node in the service area. However, in a wireless network, it is almost impossible to supply every demand node with a sufficient service level. Thus, a different problem definition is required, the *maximal coverage location problem.*

Greedy Heuristic for the Maximal Coverage Location Problem

An efficient greedy heuristic for the maximal coverage location problem (MCLP) was proposed by Vohra and Hall (1993). In contrast to the heuristic for the SCP, the MCLP algorithm drops the constraint of covering every demand node. The proposed algorithm terminates as soon as the solution consists of exactly p transmitter configurations. Furthermore, the MCLP heuristic instead uses the coverage-per-cost-unit ratio, i.e., the sum of the weights of the covered nodes $w(P_j) = \sum_{i \in P_j} w_i$ as its objective function. The Greedy heuristic for the MCLP is depicted in

Algorithm 6.2 (Maximal Cover Location Problem)

variables:

J	*set of indices of all transmitter configurations* j
J^*	*set of indices of transmitter configuration in cover*
p	*maximum number of transmitters*
r	*number of transmitter configurations in cover*
P_j	*cover of transmitter configuration* j
w_i	*weight of a demand node* i

algorithm:

```
1  proc maximum_cover_location() ≡
2    begin
3      J* = ∅, r = 0;                          /* initialize the variables */
4      while (r < p) do
5              find k ∈ J : max(w(P_k) = Σ(w_i : i ∈ P_k))
6              ∀j ∈ J : P_j = P_j − P_k;
7              J* = J* + k;                    /* add configuration k to cover */
8              r = r + 1;
9      od
10     stop                                    /* J* is a cover */
11   end
```

Algorithm 6.2: Greedy heuristic for solving the maximal cover location problem

Algorithm 6.2. It is very similar to the one of the SCP. The variable w_i in line 4 is denoting the weight of demand node i.

6.3.2 Approximation Capability of Greedy Heuristics

Greedy heuristics are a very intuitive and simple approach for attaining a feasible solution for set covering problems. The algorithms can be efficiently implemented and possess a polynomial computing time. Hence, it is of major interest to ask how much optimality has been traded in favor of exploiting the computing efficiency of the Greedy algorithm. Therefore, the approximation capability of the heuristics with respect to the size of the optimal solution is outlined in the following subsection.

a) Set Cover Problem

Although the Greedy algorithm for the set covering problem can easily be stated, its analysis is far from trivial. Here only the import results are presented and the reader is conferred to the referenced literature for complete proofs.

An upper bound for the approximation ratio of the Greedy algorithm for the weighted set cover problem, as defined in Section 6.3.1, has been proved by Chvatal (1979):

Theorem 6.5 :

The cost of the cover returned by the Greedy heuristic is at most $H(d) = \sum_{i=1}^{d} \frac{1}{i}$ times the cost of an optimal cover.

Proof: see Chvatal (1979). □

Here, d denotes the size of the largest set. The *harmonic function* $H(d)$ is bounded by $1 + \log d$.

A lower threshold for which Set Cover can not be efficiently approximated was shown by Feige (1996):

Theorem 6.6 :

Set Cover can not be approximated efficiently within $(1-o(1)) \ln d$ unless NP has slightly superpolynomial algorithms.

Proof: see Feige (1996). □

Thus, the Greedy heuristic is an $O(\log d)$-approximation algorithm for any set cover. However, better results can obtained for special cases, for example *high coverage* instances, cf. Hochbaum (1996) p. 102.

b) Maximal Cover Location Problem

The results for the approximation ratio of the MCLP can be obtained when investigating the optimization of a submodular function f, cf. Hochbaum (1996).Now, let p the limit for the number of sets in the solution, and w(GREEDY) and w(OPT) the total weight of the elements of the solution of the Greedy heuristic and the optimal solution. Hence, the following theorem can be stated:

Theorem 6.7 :

$$\begin{aligned} w(\text{GREEDY}) &\geq \left[1 - (1 - \frac{1}{p})^p\right] w(\text{OPT}) \\ &> (1 - \frac{1}{e}) w(\text{OPT}) \end{aligned}$$

Proof: see Hochbaum (1996), pp. 136. □

As a result, the Greedy heuristic is a $(1 - \frac{1}{e})$-approximation algorithm for the Maximum Coverage Location problem. Hochbaum (1996)backs up this theory by providing an example which illustrates that the bound is tight.

An interesting extension of the results of Theorem 6.7 was presented by Vohra and Hall (1993):

Theorem 6.8 :

$$w(\text{GREEDY}) \geq \max\left\{\frac{p}{n}, 1 - (1 - \frac{1}{p})^p\right\} w(\text{OPT})$$

Proof: see Vohra and Hall (1993). □

The difference between the two lower bounds for the MCLP Greedy algorithm, is the factor $\frac{p}{n}$. Hereby, p denotes the limit of sets in a solution and n is the number of candidate sets, i.e. the number of potential transmitter configurations from which p ones are to be chosen.

If the Greedy has to choose more $(1 - \frac{1}{e}) \cdot n \approx 0.632 \cdot n$, i.e. the limit of $1 - (1 - \frac{1}{k})^k$ for $k \to \infty$, out of n potential transmitters, then the guaranteed bound is lifted up, which indicates a great deficiency. However, the impact of this feature on radio network engineering is not as significant as one might imagine. Usually a much small number of transmitter configuration has to be selected as potential candidates are provided.

6.3.3 Objective Cost Function

The proposed Greedy heuristics assumes that the desirability of having transmitter configuration j, in an optimal cover increases with ratio $\mid P_j \mid$

$/c_j$, with P_j and c_j as defined in Section 6.3.1. Hence, the algorithms are maximizing the coverage per cost unit.

From an economical viewpoint, the cost of using a specific transmitter configuration at location with index j, i.e. having set j in the solution, usually depends on three factors: *a)* the specific hardware costs, e.g. high-power transmitters are usually more expensive than low-power ones, *b)* the costs of leasing or purchasing real estate, and *c)* the infra-structural costs of deploying and connecting the cell site to the fixed part of the mobile network, e.g. providing electrical power or data connections to the transport network, is much more expensive in remote regions than in developed areas.

Apart from the financial costs of a transmitter configuration, it is possible to define other virtual cost factors of base stations. In this way, complex radio network design constraints which usually require an extensive computational amount can easily be included into the objective function of the optimization. Of course, this is only an approximation of the actual design constraint. However, in this way the computing time can be reduced, as in the case of *radio signal interference*. A high-power transmitter which is located on a hill top deteriorates the interference situation in the whole network scenario and decreases the cellular capacity, see also Section 6.5. Therefore, using this transmitter in a solution would be very "expensive". Thus, assigning a high value to c_k, for a certain transmitter configuration k which is expected to produce high radio interference, will reduce the probability of having this configuration in the final solution.

Nevertheless, an accurate way of including the interference constraint is presented in Section 6.5. The interference of an investigated base station j is evaluated by calculating the average interference level of an intermediate solution J extended by transmitter j at every demand node locations. However, the accurate inclusion of the interference will increase the run time complexity of the transmitter locating task, cf. Section 6.5.

6.3.4 SCBPA Algorithm

Based on the proposed Greedy heuristic for the MCLP presented above, it possible to define the complete *set cover base station positioning al-*

gorithm (SCBPA). The algorithm comprises four phases:

Phase 0	Calculate all possible coverage sets P_j
Phase 1	Verify the traffic and hardware constraints for all P_j
Phase 2	Compute a first coverage by using the greedy heuristic for the MCLP
Phase 3	Optimize the coverage by changing transmitters (optional)

In *Phase* 0, the algorithm computes for all valid cell site locations, all transmitter power levels, and all possible antenna types the supplying areas P_j of these configurations. *Phase* 1 verifies whether the transmitters obey the traffic and hardware constraints or not, cf. Section 6.4.3. In *Phase* 2, a *first coverage* of the planning region is computed using the Greedy algorithm for the MCLP. Due to the heuristic nature of this procedure, the obtained solution might only be suboptimal. An improvement can be found by the application of an optional *Phase* 3, where the result is improved by exchanging transmitter using a *local search* algorithm; see also Section 6.3.5 for an Simulated Annealing approach and Papadimitriou and Steiglitz (1982) for other Local Search methods.

6.3.5 Simulated Annealing

The second option presented here for solving the maximal cover location problem in the context of the transmitter locating task a *Simulated Annealing*. Simulated Annealing is a competing approximate method to solve large scale combinatorial optimization problems. In contrast to the above introduced Greedy heuristics, which are based on experience and expectation rules, Simulated Annealing is a more general method for solving optimization problems. It belongs to the class of *local search* algorithm. In fact, it is a *randomization* algorithm and it can be asymptotically viewed as an optimization algorithm.

In this section the major components of the Simulated Annealing method are outlined. A detailed presentation of the theory and applications can be found in Aarts and Korst (1990).

First, some basic notions required to outline Simulated Annealing are introduced. An instance of a combinatorial optimization problem is given by an ordered pair $(\mathcal{V}, f)$, where $\mathcal{V}$ is a finite set and $f : \mathcal{V} \to \mathbb{R}$

Algorithm 6.3 (Simulated Annealing)

variables:

$\mathcal{C}$	*current state*
$\mathcal{C}'$	*subsequent state*
$\mathcal{C}_{start}$	*initial state*
c_k	*control parameter , i.e. the temperature, at iteration* k
c_{start}	*start temperature*
k	*iteration counter*

algorithm:

```
1  proc simulated_annealing() ≡
2    begin
3      C ← C_start
4      k ← 0;
5      c_0 ← c_start;
6      repeat
7            C' ← generate();
8            if f(C') ≤ f(C) then C ← C'
9                          else if exp((f(C)−f(C'))/c_k) > random[0,1)
10                                then C ← C' fi
11           fi
12           k ← k + 1;
13           c_k ← new_temperature(); /* adapt c_k to cooling schedule */
14     until stopcriterion
15   end
```

Algorithm 6.3: Simulated Annealing

is a cost function. The aim is to find an element in $\mathcal{V}$ minimizing or maximizing f, respectively.

The basic Simulated Annealing algorithm is shown in Algorithm 6.3. Its core components are a subroutine, commonly called `generate()`, and an appropriate cooling schedule for a control parameter c_k, with $c_k \in \mathbb{R}^+$, denoted as the *temperature*. The procedure `generate()` creates from the current state $\mathcal{C} \in \mathcal{V}$ at random with probability $p_{\mathcal{C}\mathcal{C}'}$ a new state $\mathcal{C}' \in \mathcal{V}$. The set $\{\mathcal{C}' \in \mathcal{V} \mid p_{\mathcal{C}\mathcal{C}'} > 0\}$ is called the neighborhood of $\mathcal{C} \in \mathcal{V}$. The new state $\mathcal{C}'$ is always accepted if the cost function is improved, cf. line *8*. A deterioration of the solution is also accepted to a limited extent. This is implemented by comparing the value of $\exp((f(\mathcal{C}) - f(\mathcal{C}'))/c_k)$ with a random number generated from a uniform distribution on the interval

$[0,1)$, see line *9*. In every iteration, the control parameter c_k is adapted to the cooling schedule by the function `new_temperture()`, cf. line *13*.

Starting with large values of c_{start}, Simulated Annealing will accept large deteriorations. As c_k decreases, the algorithm will accept only small deteriorations, and finally, as $c_k \to 0$, no deteriorations will be accepted at all. Due to this feature, the Simulated Annealing is able to escape from local minima while it still exhibits the favorable features of Local Search algorithms, i.e. simplicity and general applicability.

Simulated Annealing can easily be be adapted to solve the MCLP problem. The state space can be described by $\mathcal{V} = S^p$, where S is the set of all potential transmitter configurations and p is the number of transmitters to be deployed. Therefore, $\mathcal{C} = (c_1, \ldots, c_p) \in \mathcal{V}$ corresponds to building base station configurations $c_1, \ldots, c_p$, with neglecting duplicates, i.e. $c_i \neq c_j$ for $i \neq j$. Hence, at most p base stations can be selected in a solution. The cost function f measures the number of covered demand nodes, i.e. the amount of served traffic.

A careful choice of neighborhoods and transition probabilities is essential for the performance of Simulated Annealing. Roughly speaking, neighboring states should not differ too much so that updated states are improved slowly step by step. To meet this requirement, for each configuration i a set of configurations $\mathcal{C}(i)$ is defined. Given a current state $\mathcal{C} = (c_1, \ldots, c_p)$, the procedure `generate()` first chooses with probability $1/p$ an index $j \in \{1, \ldots, p\}$. Then, the component c_j of $\mathcal{C}$ is replaced by $c'_j \in \mathcal{C}(c_j)$, where c'_j is chosen with probability $1/\mid \mathcal{C}(c_j) \mid$. The output of `generate()` is the state $\mathcal{C}' = (c_1, \ldots, c_{j-1}, c'_j, c_{j+1}, \ldots, c_p)$.

The procedure `generate()` is thus completely described by the sets $\mathcal{C}(i)$. Simulated Annealing performed best on the investigated instances, when each $\mathcal{C}(i)$ contains M, $50 \leq M \leq 150$, different configurations serving a set of demand nodes close to that of i. Closeness is measured by the number of commonly served demand nodes.

Due to its generality, the Simulated Annealing approach was investigated in this monograph mainly for the interference minimizing transmitter locating task. An elaborated description of the performance and the results obtained by the Simulated Annealing approach are presented in Section 6.5.

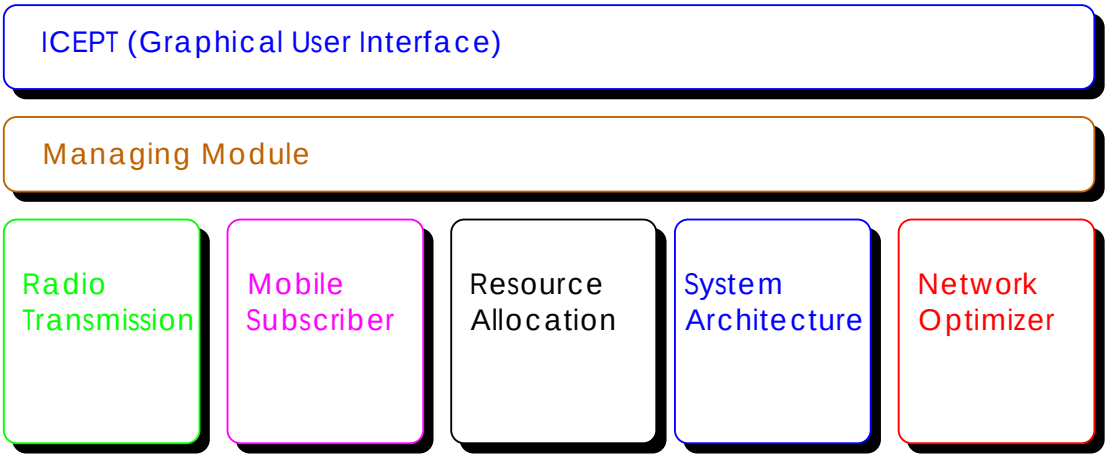


Figure 6.4: *The modules of ICEPT*

6.4 The ICEPT Planning Tool Demonstrator

To prove the capability of the new demand-oriented design approach for cellular communications systems, the radio network optimization methods presented above were implemented in the planning tool demonstrator ICEPT. The abbreviation ICEPT stands for Integrated CEllular network Planning Tool. ICEPT is able to synthesize a complete radio network configuration for a given realistic planning case. The tool is based on the *Integrated Design* approach, cf. Chapter 4. It addresses all the five major design aspects of the new method.

6.4.1 Tool Prototype

Despite being a technology demonstrator, ICEPT already includes the following features of a real planning tool: the use of real–world data, a fast network design algorithm, scalable accuracy and speed, a modular and reconfigurabel program structure, and an easy–to–use graphical user interface.

The current structure of ICEPT is depicted in Figure 6.4. The demonstrator is organized into three layers. The lowest layer consists of five core modules, paralleling the five aspects of the *Integrated Design* approach. The five core modules are: *a)* the Mobile Subscriber component for characterizing the traffic distribution in the planning region using the

demand node concept; *b)* the Radio Transmission module for estimating the radio wave propagation in the service area; *c)* the Resource Allocation component for performing the frequency allocation task; *d)* the System Architecture module for adapting the specific system characteristics to general applicability in the optimization stage, and *e)* the Network Optimizer, which implements the radio network optimization algorithms. The integration of these modules is done by the middle layer, the Management Module. The management module conducts the process of the network synthesis. It defines the *Network Design Sequence*, cf. Section 6.4.3. On top of the managing module resides an easy-to-use graphical user interface (GUI). ICEPT is completely object-oriented and written in C++. It can be adapted to different planning scenarios and system architectures via a the GUI and a configuration file.

The Radio Transmission module comprises the two commonly used two-dimensional outdoor radio wave propagation prediction methods Hata and COST231. The required topographical and morphographical input data of the module is depicted in Figure 6.5. Part (a) of Figure 6.5 shows the bird's-eye view of the three-dimensional digital terrain model of the Würzburg planning scenario. Part (b) of Figure 6.5 depicts the land use data of this region. A more detailed presentation of the internals of the Radio Transmission module is provided in Section 6.4.2

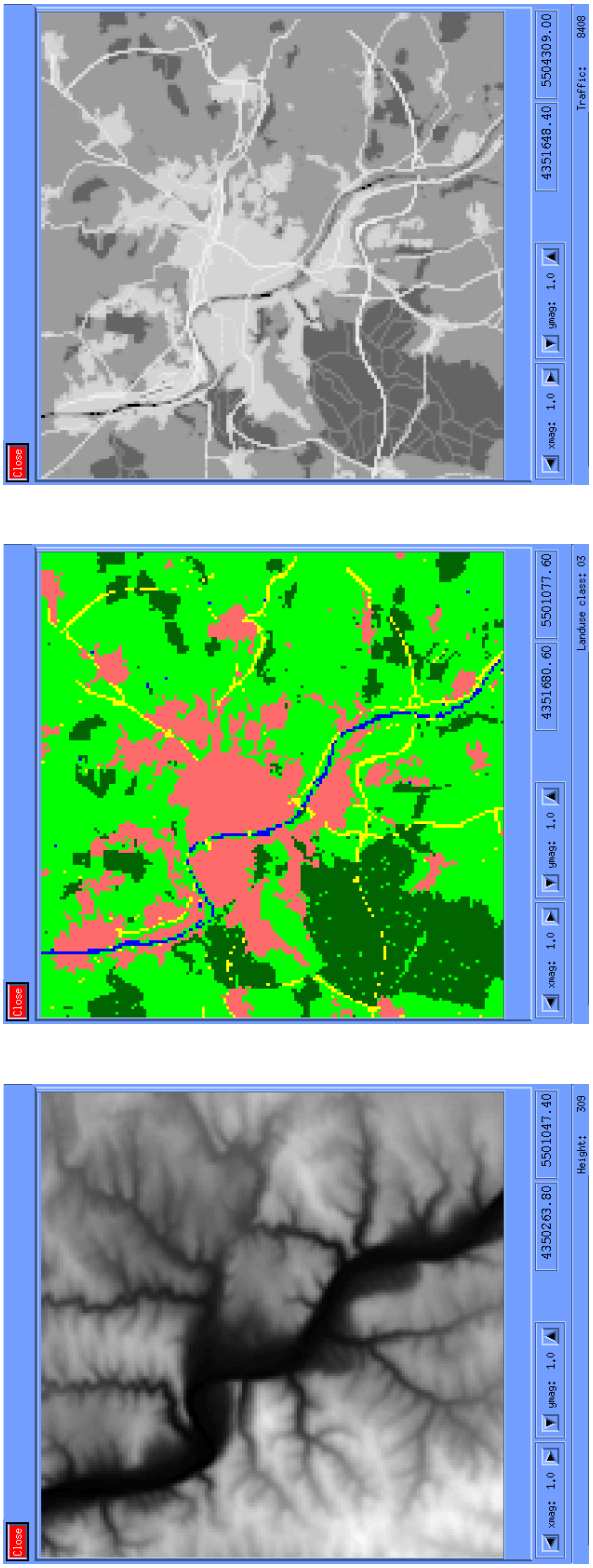
The Mobile Subscriber component implements the partitional clustering algorithm as well as the agglomerative method, cf. Chapter 5. Figure 6.5(c) shows the traffic matrix which is required as input to the clustering algorithms.

The main task of the Resource Allocation module is to compute the frequency allocation plan for the network configuration. In particular, ICEPT uses the *Generalized Sequential Packing (GSP)* algorithm for frequency assignment, cf. Sung and Wong (1995).

The System Architecture module adapts the specific system parameter to general use, e.g. it provides a mapping from logical channel number to the physical carrier frequency and looks up technical parameters like receiver sensitivity.

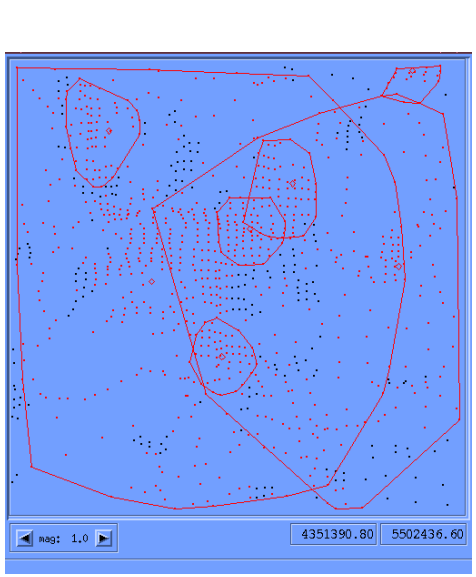
The Network Optimizer module implements the optimization methods. ICEPT uses the SCBPA for optimizing the transmitter locations.

The graphical output of ICEPT is shown in Figure 6.6. Here, seven transmitters have been deployed in the Würzburg planning scenario. Part (a) of Figure 6.6 depicts the transmitter locations and the corresponding sets of covered demand nodes indicated by the convex hull around the sets. This description is better suited in the context of the set covering

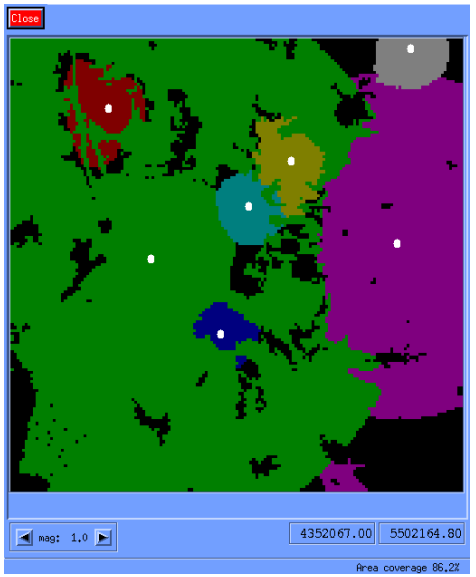


(a) Digital terrain model (b) Land use (c) Traffic matrix

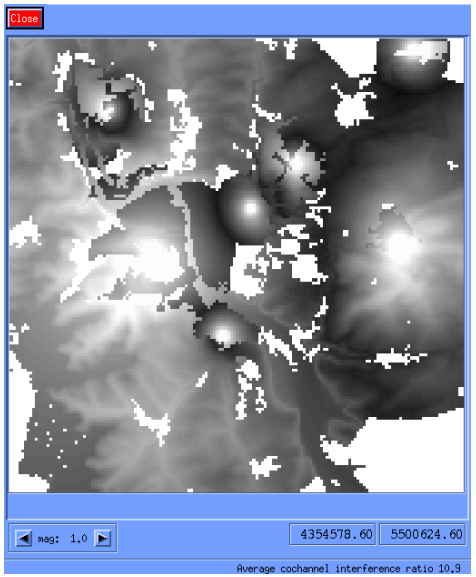
Figure 6.5: *Input data to ICEPT*



(a)
Convex hulls



(b)
Best server



(c)
Co-channel interference

Figure 6.6: *Graphical output of ICEPT*

approach to the transmitter locating task. Figure 6.6(b) shows the *best server* plot for the planning region, which is commonly used in radio network engineering. Part (c) of Figure 6.6 depicts the graphical co-channel interference evaluation of the service area. Bright values represent a low interference, dark values indicate a high interference level.

6.4.2 Radio Wave Propagation

The present version of the Radio Transmission module in ICEPT comprises the outdoor radio wave prediction methods of Hata and COST231, cf. Hata (1980) and Stüber (1996).

The Hata model is based on the empirical data collected by Okumura et al. (1968). It is widely used in cellular radio engineering since its frequency range covers most of the spectrum used by second generation cellular systems, like the GSM900 cellular system, cf. ETSI (1995a). The model is valid for the following range of parameters: carrier frequency $450MHz \leq f_c \leq 1000MHz$, base station antenna height $30m \leq h_b \leq 200m$, mobile station antenna height $1m \leq h_m \leq 10m$, and the distance $1km \leq d \leq 20km$ between the transmitter and the mobile station. In the Hata model, the path loss L (in dB) at distance d from the transmitter is

$$\begin{aligned} L_{(dB)}(d) &= 69.55 + 26.16 \cdot \log_{10}(f_c) - 13.82 \cdot \log_{10}(h_b) \\ &\quad -a_x + (44.9 - 6.55 \cdot \log_{10}(h_b)) \cdot \log_{10}(d) \end{aligned} \tag{6.6}$$

with a_x denoting the *land use type* component:

$$\begin{aligned} &\text{DenseUrban} \quad a_{du} = 3.2 \cdot (\log_{10}(11.75 \cdot h_m))^2 - 4.97, \\ &\text{Urban} \quad a_u = (1.1 \cdot \log_{10}(f_c) - 0.7) \cdot h_m \\ &\qquad\qquad -(1.56 \cdot \log_{10}(f_c) - 0.8), \\ &\text{Suburban} \quad a_{su} = a_u + 5.4 + 2 \cdot (\log_{10}(\tfrac{f_c}{28}))^2, \\ &\text{Rural/Open} \quad a_r = a_u + 40.94 + 4.78 \cdot (\log_{10}(f_c))^2 \\ &\text{Outdoor} \qquad\qquad -18.33 \cdot \log_{10}(f_c). \end{aligned}$$

The COST231 model is an extension of the Hata method for a higher carrier frequency range of $1500MHz \leq f_c \leq 2000MHz$, which is commonly used in the european DCS1800 system, cf. ETSI (1995b), and north-american PCS1900 cellular systems, cf. ANSI (1997). The other

parameters in the COST231 model are limited to the same constraints as in the Hata model. In the COST231 model, the path loss L (in dB) is

$$\begin{aligned} L_{(dB)}(d) &= 46.3 + 33.9 \cdot \log_{10}(f_c) - 13.82 \cdot \log_{10}(h_b) \\ &\quad -a_x + (44.9 - 6.55 \cdot \log_{10}(h_b)) \cdot \log_{10}(d) \end{aligned} \tag{6.7}$$

with a_x :

DenseUrban	a_{du}	$=$	$3.2 \cdot (\log_{10}(11.75 \cdot h_m))^2 - 7.97,$
Urban	a_u	$=$	$(1.1 \cdot \log_{10}(f_c) - 0.7) \cdot h_m$ $-(1.56 \cdot \log_{10}(f_c) - 0.8),$
Suburban	a_{su}	$=$	$a_u + 3 + 12.11,$
Rural/Open Outdoor	a_r	$=$	$a_u + 3 + 27.23.$

The Hata model as well as the COST231 model are limited to path lengths above $1km$. It should not be used for smaller ranges where the path loss becomes highly dependent upon the local topography.

Land Use Decision

For an accurate approximation of the radio wave propagation, the selection of the proper land use component is very important. In the present implementation of the *Radio Transmission* module the land use type of the transmitter is defined as the land use at the location of the transmitter. However, a demand node represents a whole area, and therefore its land use class is defined as the land use that occurs most in the area represented by the node. The worst case of these two land use types is taken into account for selecting the land use component a_x. Thus, the radio transmission module of ICEPT guarantees a worst case approximation of the radio wave propagation in the planning region.

Received Signal Power Level

The received signal power level $\Omega_{(dB)}$ at the distance d from a transmitter is, cf. Faruque (1996) for further details:

$$\Omega_{(dB)}(d) = P^{ER}_{(dB)} - L_{(db)}(d). \tag{6.8}$$

Hereby $P^{ER}_{(dB)}$ denotes the effective radiated power of the transmitter in the direction of the mobile (in dB).

Link Budget

The Radio Transmission module of ICEPT uses the concept of the *link budget* for deciding whether a demand node receives a sufficient signal strength or not. Hence, besides the hypothetical path loss, the other physical parameters such as the radiated power, the cable loss, or the antenna gain, are taken into account for the coverage decision. The *forward link budget* (B^F) (in dB) from the base station to a mobile station is defined as

$$B^F_{(dB)} = p_{BTS} + g_{BTS} - l_{ccd} + s_{MS} + g_{MS} - l_{\Delta}; \tag{6.9}$$

and the *reverse link budget* (B^R) (in dB), i.e. in the direction from the mobile station to the base station, is

$$B^R_{(dB)} = p_{MS} + g_{MS} - l_{ccd} + s_{BTS} + g_{BTS} - l_{\Delta} \tag{6.10}$$

with:

p_{BTS}	the nominal transmitting power of the *base station*,
p_{MS}	the nominal transmitting power of the *mobile station*,
g_{BTS}	the signal improvement at the *base station* due the use of specific antenna type and its directivity,
g_{MS}	the signal improvement at the *mobile station* due the use of specific antenna type and its directivity,
l_{ccd}	the signal deterioration in cables and connectors to and from the antenna,
l_{Δ}	the loss parameter that accumulates all other factors which lead to signal deterioration,
s_{BTS}	the receiver sensitivity of the *base station*, and
s_{MS}	the receiver sensitivity of the *mobile station*.

All parameters are given in dB. Since the radio signal processing is asymmetric on the forward link and on the reverse link, it is necessary to use a worst case approximation. Hence, a demand node receives a sufficient signal if:

$$\min\{B^F_{(dB)}, B^R_{(dB)}\} > L_{(dB)}. \tag{6.11}$$

Programming Interface

Due to the modular and object-oriented design of ICEPT, additional radio wave propagation prediction methods can easily be added or the existing one can be replaced for a better suitability in certain environments, e.g. for planning micro cells in dense urban areas.

6.4.3 Network Design Sequence

The *network design sequence* of ICEPT is depicted in Figure 6.7. In contrast to the conventional cellular design method, the demand-based approach of ICEPT starts with the traffic characterization. Therefore, the tool generates at first the demand node distribution of the planning region. Afterwards, the program computes the coverage areas for all possible transmitter configurations, see *Phase* 0 of the SCBPA algorithm, cf. Section 6.3.4. The potential configurations are defined by an iteration over the parameters "location", e.g. overlaying a grid of candidate locations on the planning region, "transmitting power level", e.g. power level steps as defined in the system specifications, see ETSI (1995a), and "base station antenna height". The modules' names to the left and to the right of the sequence indicate the core components which perform current phase of the design process.

In the next step, ICEPT checks whether the traffic and hardware constraints are obeyed at these configurations or not, see also *Phase* 1 of SCBPA. Invalid configurations are removed and not considered during optimization. This step contributed strongly to the great flexibility of ICEPTs' demand-oriented and economic network design approach. For example, a network operator defines the constraint to deploy only transmitter configurations which have at most two radio modems, since the smaller configurations are cheaper than the bigger ones. However, a specific transmitter configuration at a certain location would supply more traffic than can be supported by the two modems. Thus, the configuration is invalid, deleted, and not regarded during optimization.

After completing the traffic and hardware verification, the optimizer computes the optimal transmitter locations using *Phase* 2 of the SCBPA algorithm. Subsequently, the tool computes the carrier separation constraints and constructs a frequency allocation plan. If the tool is unable to calculate a valid frequency plan, it has to split certain cells.

If the frequency allocation plan is valid, ICEPT verifies the carrier-to-interference (C/I) values of the configuration. In case the C/I constraints

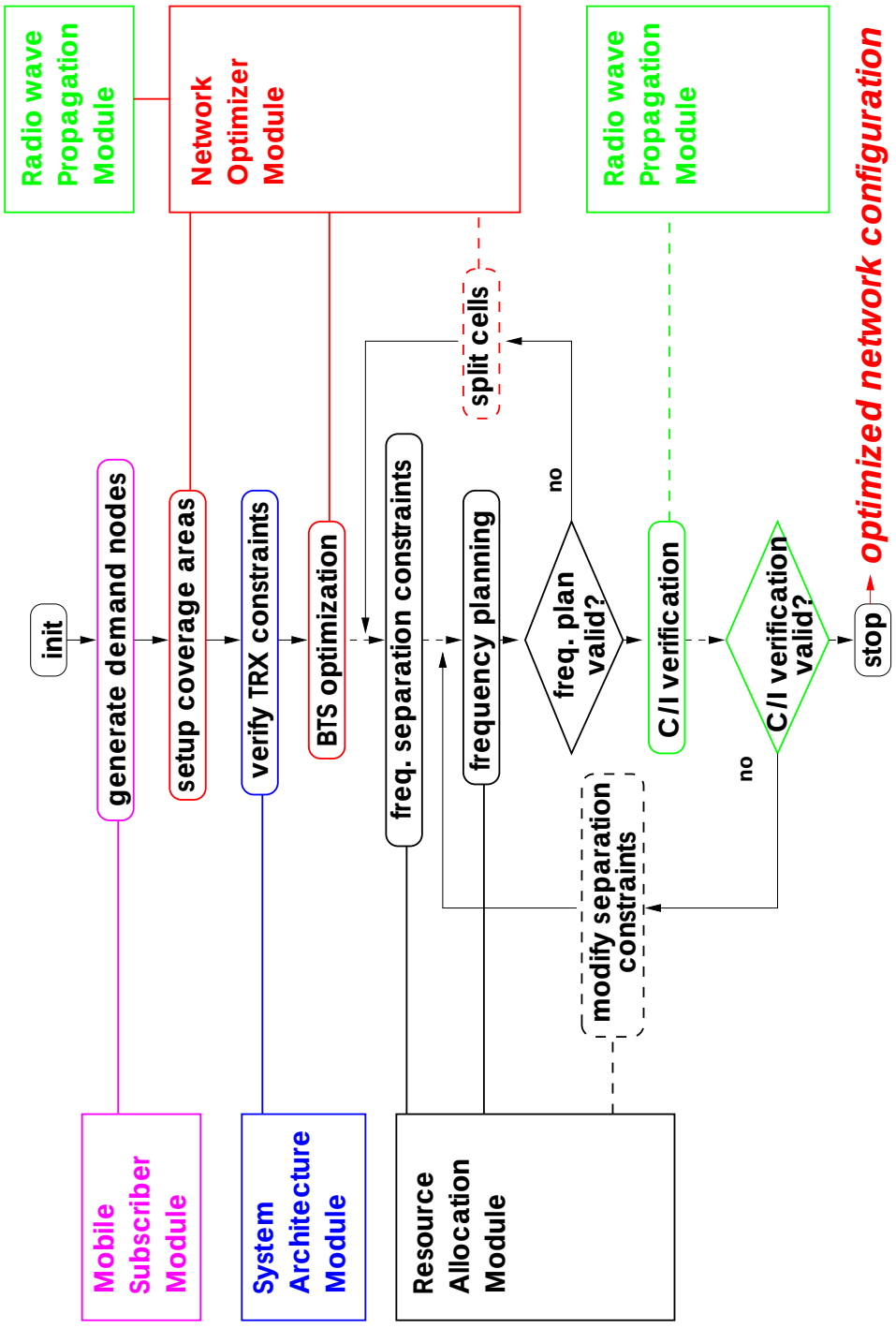


Figure 6.7: *ICEPT's network design sequence*

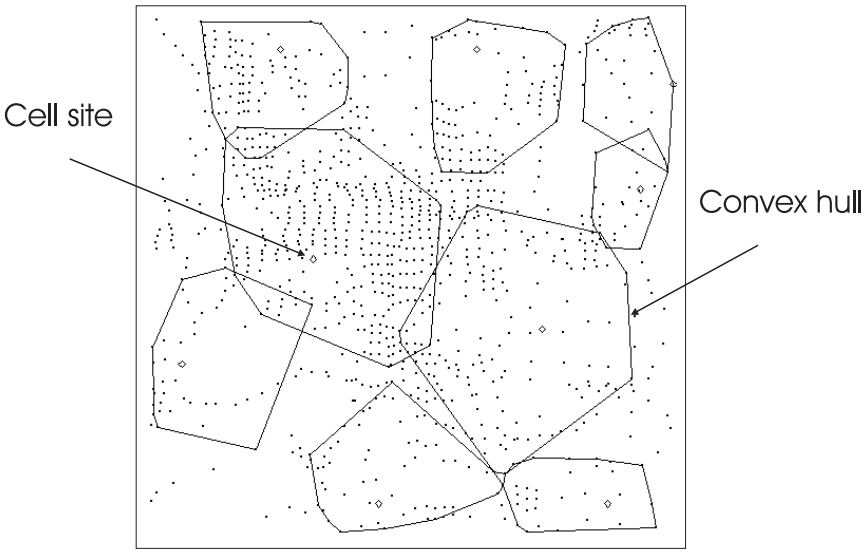


Figure 6.8: *ICEPT planning result: base station locations*

are not obeyed, then the separation constraints have to be increased. If the C/I specifications are met, the network design stops with the output of the cellular radio network configuration.

6.4.4 Planning Result

ICEPT was tested on the topography around the city center of Würzburg. The task was to find the optimal locations for nine transmitters. A typical result of the SCBPA algorithm is depicted in Figure 6.8. The base station locations are marked by a $\diamond$ symbol. The lines indicate the convex hull around the set of demand nodes which are supplied by the base station. In this experiment, ICEPTs' SCBPA algorithm was able to obtain a 75% coverage of the teletraffic of the investigated area. The total computing time for the configuration, including the traffic characterization, was 4min on a SUN Ultra 1/170.

6.5 Interference Minimizing Radio Network Design

The main characteristic of the cellular concept in mobile communication systems is the application of *frequency reuse* for FDMA/TDMA wire-

less networks in order to increase the traffic capacity of the radio part of these systems. Users in geographically separated areas are simultaneously employing the same carrier frequency. However, the frequency reuse introduces co- and adjacent-channel interference which limits the theoretical capacity gain of the reuse if the geographical separation of cells with the same frequency is too small.

The high capacity design of a cellular network requires that, after selecting the cell site, frequencies are allocated to the cells in such a way that the co-channel and the adjacent channel interference in the cells is minimized. Due to the inhomogeneous traffic distribution and the irregular shape of the cell boundaries, however, the frequency allocation procedure is extremely difficult. In order to decrease the complexity of this engineering task, already the selection of cell sites should be carried out with regard to interference. Especially, the worst case scenario of co-channel interference (CCI) has to be addressed at an early stage during the cellular design.

So far, the presented algorithms for solving the base station placement problem are based on the set-oriented view of a cell covering a certain number of demand nodes, cf. Definition 6.1. The coverage metric is defined solely by the maximal eligible path loss from the investigated transmitter to the demand nodes, without regarding the other transmitters.

The inclusion of the interference value as a design criterion for a single base station, however, is quite a challenge: the interference measured at a certain location depends on the one side on the signal disturbance introduced by an investigated new base station, and on the other side on the configuration of other interfering, already located, transmitters. Hence, every candidate transmitter has to be evaluated in context of all other transmitters in the configuration. Thus, the run time complexity of the interference minimizing transmitter locating task increases over the one of the conventional covering problem approach.

In the following section, two efficient methods are proposed which are capable to maximize the average CCI ratio while optimizing the traffic coverage: *a)* a Greedy-based heuristic and *b)* an approach based on Simulated Annealing. Before introducing the interference minimizing design algorithms, the basic RF objectives for automatic radio network engineering and the calculation of the co-channel interference ratio in cellular systems are outlined below.

6.5.1 RF Design Objectives

Most automatic cellular network design algorithms consider the received signal power level $\Omega_{(dB)}$ at certain test points as their main design objective, cf. Calégari et al. (1997) and Chamaret et al. (1997). However, the consideration of this value as the sole design criterion is insufficient. The provision of a usable radio link requires at least the fulfillment of two constraints:

a) the received signal level $\Omega_{(dB)}$ has to obey the threshold $\Omega_{th,(dB)}$, defined by the link budget, cf. Faruque (1996):

$$\Omega_{(dB)} > \Omega_{th,(dB)}, \tag{6.12}$$

b) the co-channel interference ratio $\Lambda_{(dB)}$ is not allowed to exceed the interference threshold $\Lambda_{th,(dB)}$:

$$\Lambda_{(dB)} < \Lambda_{th,(dB)}. \tag{6.13}$$

The threshold $\Lambda_{th,(dB)}$ is defined by the receiver sensitivity.

Co-channel Interference

Figure 6.9 depicts a typical co-channel interference scenario of a cellular system. A mobile station MS in distance d_0 receives the strongest signal from base station Tx_0, which is denoted as the *best server*. The reception of this signal is disturbed by three surrounding *interferers* $\mathrm{Tx}_k, k = 1 \ldots 3$. Additionally, it is assumed that the interferers are transmitting on the same frequency as the best server. The average downlink CCI ratio at the location of the mobile station is, cf. Stüber (1996):

$$\Lambda_{(dB)} \quad = \quad \Omega_{(dB)}(d_0) - 10\log_{10}\{\sum_{k=1}^{N_I} 10^{\Omega_{(dB)}(d_k)/10}\}, \tag{6.14}$$

where $\Omega_{(dB)}(d_k)$ is the received signal power level from transmitter Tx_k, d_k is the distance of the mobile station to the transmitter and N_I is the number of interferers.

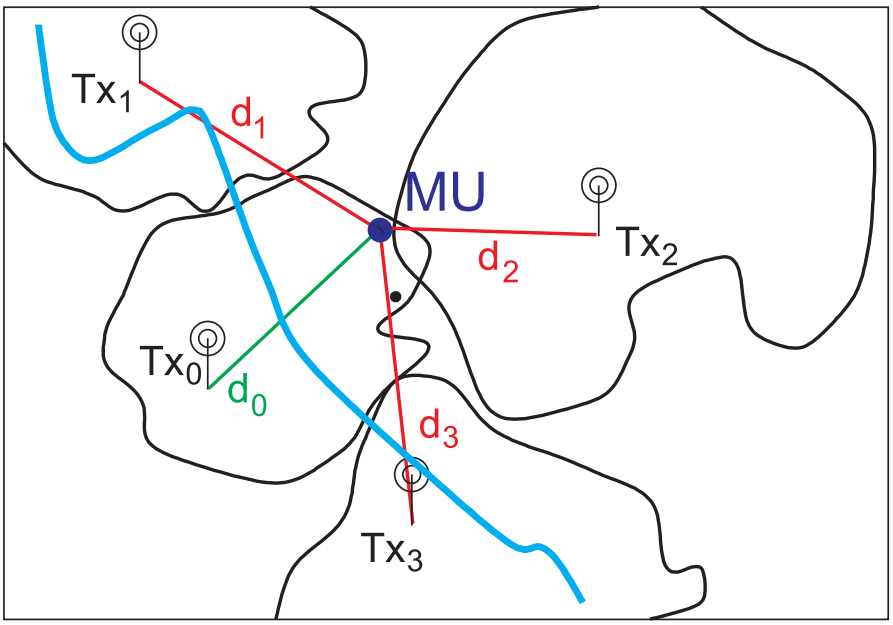


Figure 6.9: *Interferer scenario*

6.5.2 Interference Minimizing Design Algorithms

Automatic cell site selection algorithms facilitate the deployment of mobile systems in two significant ways. First, they are capable of verifying a large number of different sites until the optimal set of sites is found under the given constraints. Second, automatic selection algorithms accelerate the engineering process. A preliminary network configuration synthesized by the algorithms can serve as an immediate starting point for the detailed determination of system parameters. Hence, the network designer does not have to deal with invalid sites. These two advantages of automatic network design algorithms suggest their application in the complex task of interference minimizing radio engineering.

The proposed interference minimizing design algorithms are extensively exploiting the *demand node concept (DNC)*, cf. Chapter 5. However, due to the additional constraint of Eqn. (6.13), it is necessary to enhance the definition of the term *coverage area*:

Definition 6.11 : Coverage Area (with respect to interference)

The coverage area of a transmitter with respect to interference is the set of demand nodes which are provided with a usable radio link according to Eqn.(6.12) and Eqn.(6.13).

An appealing feature of the demand node concept, also taking into account Definition 6.11, is the fact, that the validation of the RF performance of a new base station only requires the calculation of field strength values at positions where it is highly probable to locate a mobile subscriber. It is not necessary any more to compute performance values at every location within the service area. Thus, the DNC leads to a significant speed up of the design of cellular systems.

a) Objective Function for Interference Minimization

An interference minimizing design of a cellular network requires that the decision variable y_i, cf. Eqn.(6.3), which indicates whether a demand node is covered or not, obeys the two constraints given by Eqn.(6.12) and Eqn.(6.13). Hence, the covering criteria is enhanced to:

$$y_i = \begin{cases} 1 & \exists j \in N_i : \quad (\Omega_{(dB)}(i,j) > \Omega_{th,(dB)}) \\ & \qquad\qquad \wedge (\Lambda_{(dB)}(i) < \Lambda_{th,(dB)}(i)) \,, \\ 0 & \text{otherwise} \end{cases} \tag{6.15}$$

where $\Omega_{(dB)}(i,j)$ is the received signal strength at demand node i from transmitter j, and $\Lambda_{(dB)}(i)$ is the co-channel interference according to Eqn.(6.14).

b) Greedy Heuristic Solution

Due to its flexibility, a Greedy heuristic, cf. Section 6.3.1, was selected first as a method for solving the interference minimizing transmitter location task. The algorithm imposes no restriction on the maximum number of potential base station configurations. The enhanced Greedy algorithm is shown in Algorithm 6.4. Its major difference to Algorithm 6.2 is the application of a modified objective function `if_cover()`, depicted in Function 6.1. This function calculates the weighted number of covered demand nodes with respect to interference, cf. Eqn.(6.15).

To obtain the weighted cover, Function 6.1 computes for every demand node the best serving transmitter, i.e. the base station which provides the highest radio signal level. Afterwards, it validates whether the co-channel interference at the location of the demand node is below the given threshold, cf. Eqn.(6.13). The co-channel interference is calculated

Algorithm 6.4 (Optimize configuration under interfernce constraints)

variables:

S_i	*configuration of a transmitter at location with index i*
$\mathcal{S}$	*set of all potential transmitter configurations*
$\mathcal{J}$	*set of selected transmitter configurations*
DN	*set of all demand nodes*

algorithm:

```
1  proc optimize_net() ≡
2    begin
3      S ← all configurations S_i;
4      J ← ∅;
5      find S_i ∈ S : if_cover(J + S_i, DN) is max;
6      J ← C + S_i;
7      S ← S − S_i;
8      if | J |= p ∨ if_cover(J) > required %
9        then return J;
10        else goto 5;
11     fi
12   end
```

Algorithm 6.4: Optimize Configuration under interfernce constraints

by the function `inter()`, which is implementing Eqn.(6.14). The algorithm terminates as soon as the solution consists of p transmitters or the weighted coverage reaches the required percentage, cf. line *8*.

Function 6.1 directly reveals the increase of the run time complexity of the interference minimizing site selection problem over the conventional locating task. The calculation of the weighted sum of covered demand at interference requires that every investigated potential transmitter configuration has to be compared with the already selected base stations. Furthermore, this procedure is repeated in every iteration of the Greedy algorithm.

c) Complexity of Interference Minimizing Models

Due to the increased run time of the above proposed Greedy-based heuristic, one might fear that the complexity of the interference minimizing covering models has increased. Therefore, the evaluation of the complexity of the locating task under the interference constraint has to be reassessed. A complexity theoretical evaluation of the problem was

Function 6.1 (Coverage under interference constraints)

variables:

$\mathcal{T}$	*set of investigated configurations*
DN	*set of all demand nodes*
dn_i	*demand node with index i*
c	*weighted coverage*

algorithm:

```
 1  funct if_cover(T, DN) ≡
 2    begin
 4      c ← 0;
 5      for all dn_i ∈ DN do
 6          best_server ← 0
 7          find T_j ∈ T : Ω_(dB)(i, j) > Ω_th,(dB)
 8                 ∧ Ω_(dB)(i, j) is max; /* Eqn. (6.12) */
 9          best_server ← j;
10          if best_server > 0
11            then Λ_(dB) ← inter(dn_i, best_server, T);
12                                 /* Eqn. (6.14) */
13                 if Λ_(dB) > Λ_th,(dB) /* Eqn. (6.13) */
14                   then c ← c + a_i;
15                 fi
16          fi
17      od
18      return c;
19    end
```

Function 6.1: Compute coverage under interference constraints

undertaken by Glaßer (1998), the results of which are presented as follows. In Glaßer (1998), the interference minimizing transmitter locating problem is denoted as the *Maximize Totally Supplied Nodes (MTSN)*. The term "totally" indicates that both field strength and interference constraint are obeyed. The notation of "MTSN" is preserved in this section for reason of shortness. The first result presented by Glaßer is promising:

Theorem 6.9 :

$\widetilde{\text{MTSN}}$, i.e. the decision version of MTSN is **NP**-complete.

In particular, no polynomial time algorithm exists for MTSN optimization problem unless $\mathbf{NP} \subseteq \mathbf{P}$.

Proof: see Glaßer (1998). □

Thus, the $\widetilde{\text{MTSN}}$ problem is of the same complexity class as the SCP. Regarding the approximation capability of MTSN, Glaßer's proof provides evidence for the followoing results:

Theorem 6.10 :

For every $\epsilon > 0$ the following holds:

- MTSN has no $(n^{1-\epsilon})$-approximation algorithm, unless $\mathbf{NP} \subseteq \mathbf{ZPP}$.
- MTSN has no $(n^{0.5-\epsilon})$-approximation algorithm, unless $\mathbf{NP} \subseteq \mathbf{P}$.

Proof: see Glaßer (1998). □

The abbreviation **ZPP** is denotes the class of *polynomial randomized algorithms with zero probability of error*. These algorithms are also known as *Las Vegas* algorithms, cf. Papadimitriou (1994). At the first stage, they consist of two *Monte Carlo* algorithms which return either a "positive", a "negative" or a "no definitive" answer. In case of a "no definitive" answer, the Las Vegas algorithms repeat the execution of the Monte Carlo algorithms until a definitive answer is obtained. Thus, the probability of obtaining the correct answer asymptotically reaches the value of one.

These results show that it is very unlikely that the interference minimizing transmitter locating problem, as well as the conventional set cover approach, has satisfactory polynomial time approximation algorithms. However, this does not mean that the practical cell site selection problem is an optimization task which is far too complex for attaining feasible solutions. It is highly probable that a good heuristic does exist which can solve the task in most cases with a small error. Furthermore, better approximation algorithms might be obtained by exploiting the metric or the Euclidean space, which are used in practical problems.

d) Simulated Annealing Solution

Due to the increased run time complexity of the Greedy heuristic for the interference minimizing cell site selection problem, Simulated Annealing heuristic (SAH) was chosen as a second option for solving the task. SAH

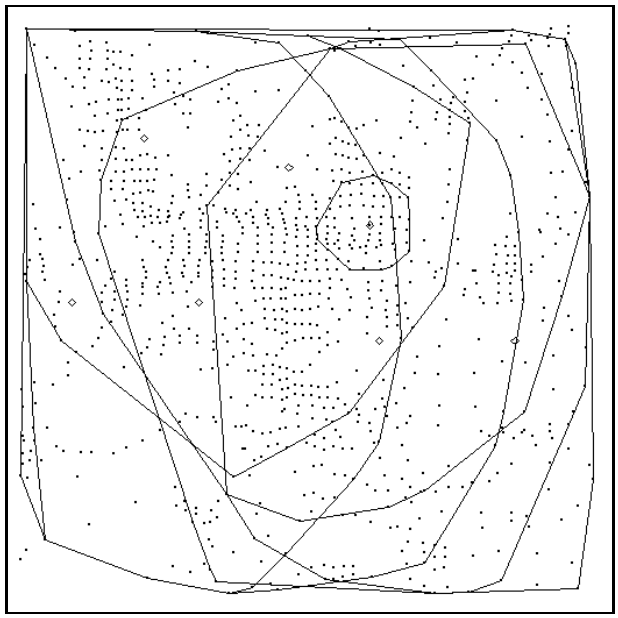
promised to be at least as efficient as the Greedy approach since it constitutes a more general approach to combinatorial optimization problems. The applied Simulated Annealing algorithm is outlined in Section 6.3.5. A drawback of the SAH was its limit to the maximal number of candidate configurations which was about 3000. However, this constraint was imposed by the application of a general Simulated Annealing algorithm.

During the experiments, it turned out that the investigated version of the SAH behaves in an instable way, in the sense that the initialization determines the quality of the final solution. The reason for this is partly inherent in the problem. Good solutions tend to have a small number of base stations each serving a large amount of traffic whereas the remaining base stations only serve traffic in small areas. This is a natural "micro/macro" cell design. However, if once the temperature is low, Simulated Annealing unlikely changes the number of "micro" and "macro" cells in the solution, since this could only be achieved by a temporary huge decrease of the cost function. To overcome the problem the temperature was restricted in such a way that with a probability of at most 1% a more worse state is accepted.

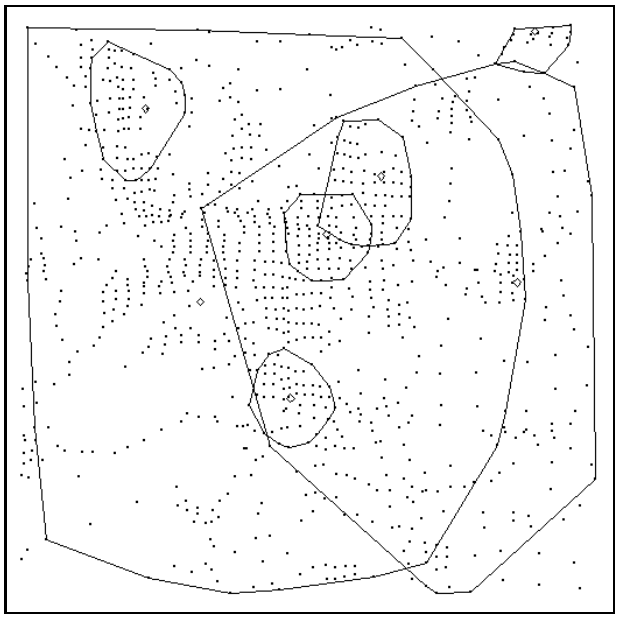
6.5.3 Single Stage Design

To prove their capability, the proposed interference minimizing design methods were integrated into the ICEPT planning tool demonstrator, cf. Section 6.4, and tested in the Würzburg planning scenario. The task is to find optimal locations for seven transmitters in the region around the city center of Würzburg. The value seven for the number of transmitters is the real number of base stations deployed in the region by a German cellular network operator. Serveral case studies were performed with different interference constraints and different sets of potential base station configurations. The algorithms were limited to single design stage. The teletraffic in the service region is described by 1127 demand nodes.

For the optimization, three instances for the set of potential transmitter configurations were considered. In the basic instance, the number of potential configurations, i.e $\mid \mathcal{S}_{basic} \mid$, was 2960. This instance was generated by overlaying the service region with an equally spaced grid of $400m \times 400m$ and using two power levels. The second and the third instance were modifications of the basic one. In the second instance, all potential configurations serving more than 50% of the total demand nodes were eliminated, since such configurations might not be desirable. Thus, a total number of seven base stations were removed. In the third



(a) Experiment 1



(b) Experiment 2

Figure 6.10: *Transmitter locations in single-stage-design using the Greedy heuristic (GRH)*

	GRH	SAH
basic instance	1029 (91.3%)	1033 (91.7%)
modified instance (50% restriction)	1006 (89.3%)	1026 (91.0%)
modified instance (33% restriction)	987 (87.6%)	1025 (90.9%)

Table 6.1: *Demand coverage of Experiment 1*

instance, all configurations serving more than 33% were eliminated, i.e. a total of 74.

In a first series of experiments, the coverage criterion was only defined by Eqn. (6.12), i.e., a demand node was considered as covered if and only if it measures a sufficient signal strength. The results of the Greedy heuristic (GRH) and Simulated Annealing heuristic (SAH) typically look as depicted in Figure 6.10(a). The positions of the selected configurations are marked by the $\diamond$ symbol. The lines indicate the convex hulls around the set of demand nodes which are supplied by the selected configurations. It can be seen that configurations with very large coverage areas are preferred, which yields a large amount of interferences. In the second series of experiments, the coverage criterion was defined by Eqns. (6.12) and (6.13), i.e., interference constraints were considered. Demand nodes that satisfy both requirements are denoted as *totally covered*. A typical result for this situation is shown in Figure 6.10(b). Configurations with large, middle and small coverage areas are selected. The results of GRH and SAH do not differ very much in the first series of experiments, but the gap increases when configurations are eliminated, see Table 6.1. Interestingly enough, SAH achieved almost the same quality in all instances.

In the second series of experiments, GRH and SAH differed not very much for the basic instance, but significantly under the 50% restriction, cf. Table 6.2. This is due to the non-iterative nature of GRH. If the first two or three selected configurations do not allow further good choices, then GRH fails. Therefore, the additional amount of computing effort when using SAH, which was limited to one hour of computing time, is strongly recommended.

6.5.4 Micro/Macro Cell Design

A common method to increase the teletraffic capacity of cellular networks while reducing interference is to use micro- and macro-cells. In areas of high teletraffic, micro-cells should be deployed to reduce interference and

	GRH	SAH
basic instance	839 (74.4%)	852 (75.6%)
modified instance (50% restriction)	732 (65.0%)	821 (72.8%)
modified instance (33% restriction)	769 (68.2%)	803 (71.3%)

Table 6.2: *Total coverage of Experiment 2*

to obtain a higher spatial frequency reuse, whereas macro-cells should be employed for the provision of area coverage. Motivated by the results obtained in the previous section, it is of interest to see how the interference minimizing technique proposed above can be extended to an explicit "micro/macro" cell design, i.e. a method which first deploys small and then large cells.

a) Two-stage Cellular Design

The micro/macro cell engineering principle was transformed into a two-stage-design algorithm for the Greedy heuristic. An extension of this principle to the SAH was not considered since it is a self-steering heuristic. The two-stage Greedy heuristic was integrated into the ICEPT tool demonstrator in two different ways:

- **Stage 1: place micro-cells.**
  A certain number of micro-cells should be placed in such a way that the demand coverage under constraints of Eqn.(6.12) and Eqn.(6.13) is maximized. Micro-cells are defined by the transmitting power and should operate at a low power level.

- **Stage 2: place macro-cells.**
  A given number of "macro" cells should be deployed in such a way that the remaining unsupplied traffic is covered. For the experiments, two versions of Stage 2 were implemented in the ICEPT demonstrator:

  a) only "macro" cells, i.e. cells with a high transmitting power, were allowed to be deployed.

  b) those "macro" and "micro" cells, which were not selected in Stage 1, were allowed to compete.

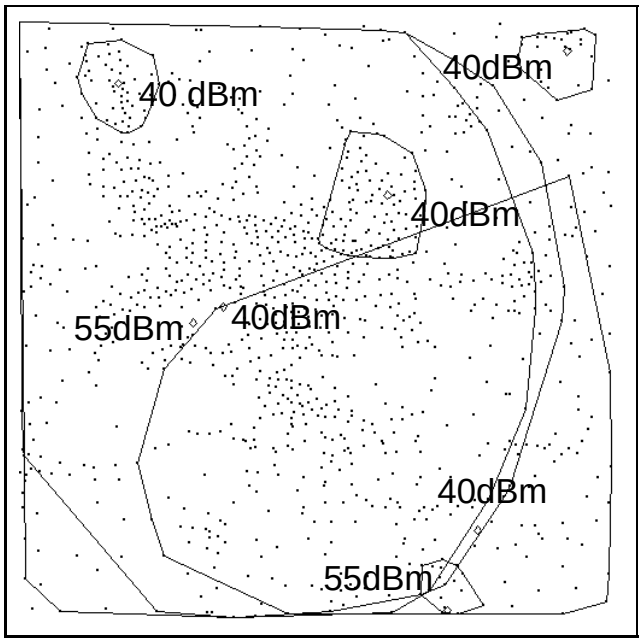


Figure 6.11: *Transmitter locations in a two-stage-design using the Greedy Heuristic (experiment 3)*

b) Results

Two case studies were carried out for the Würzburg planning scenario. In experiment 3, a micro/macro cell design was performed with version *a)* of Stage 2. At the first stage, five transmitters were allowed, using a power level of 40dBm. At the second stage two transmitters were deployed, with a power level of 55dBm. The computed transmitter locations are shown in Figure 6.11. The power levels of the transmitters are noted next to the $\diamond$ symbol. The algorithm selects two macro-cells whose coverage area is very similar to that of the two selected micro-cells, cf. Figure 6.11. Table 6.3 shows that the deployment of the additional high power "macro" cells does not improve the solution. This behavior results from our definition of the term "macro" cell which is based on the distinction of cell types according to their power level. However, a low power cell can also have a large area extension. Thus, using a sophisticated placement, it is possible to obtain the same performance with low power cells as it is possible with high power ones. To prove this statement, experiment 4 was performed using version *b)* of Stage 2. Now the high power macro-cells had to compete with the low power micro-cells. The algorithm obtained

	Exp. 3	Exp. 4
demand coverage (Eqn.(6.12) only)	77.0%	77.1%
demand coverage (Eqn.(6.12) and Eqn.(6.13))	61.4%	67.4%
average CCI ratio at covered demand nodes	11.1 dB	13.3dB

Table 6.3: *Interference minimization in two stage design*

the same solution as in experiment 2. Instead of deploying high power cells, it uses low power ones.

However, it is necessary to mention that the results obtained by the modified two-stage Greedy heuristic are not superior to the results obtained by the SAH. Due to its self-steering ability, SAH chooses automatically the best mix of transmitter configurations.

6.5.5 Summary of the Results

In this section, two automatic cellular network design algorithms have been presented. Both methods are capable to determine the locations of transmitters with respect to covered traffic and co-channel interference (CCI). The proposed algorithms are able to maximize the part of well served traffic, i.e. totally covered demand nodes, while keeping the covered teletraffic demand at a high level. The Simulated Annealing based method selects a good combination of micro and macro cells automatically, however, at the price of large computing times. Additionally, it was investigated how the Greedy heuristic of Section 6.3 can be extended for locating micro- and macro-cells. It turns out that the two-stage-design does not perform better under the given constraints. Nevertheless, it is expected that for a different definition of the term "micro cell" - i.e. a restriction of the area extension or the covered traffic in conjunction with the low power constraint - the two-stage-sequence will perform better than the single-stage-algorithm. This is an open issue and should be investigated in the future.

6.6 Concluding Remarks

This chapter introduces a new demand-oriented design methodology for radio network synthesis and optimization in cellular communications systems. It demonstrates the feasibility of the integration of three different techniques from engineering areas, which are, at first glance, difficult to

combine: *a)* facility location science, *b)* traffic engineering, and *c)* RF engineering.

In the conventional approach to radio network planning, the design areas are addressed separately. The proposed new method is based on the forward-engineering procedure of the *Integrated Approach* to cellular network planning, cf. Chapter 4, and is facilitated by the application of a new discrete population model for the traffic description, the demand node concept. This concept enables the formulation of the transmitter location task as a *maximal coverage location problem (MCLP)*, which is well known in Operations Research to model and solve facility location problems. In contrast to usual facility locating tasks, where the metric is mainly the geographical distance, the radio network planning has to regard the reception of a reliable radio link as a measure for coverage.

Unfortunately, by its nature, the MCLP belongs to a class of very intractable computational optimization problems; it is *NP-hard*. Therefore approximation methods for obtaining a near optimal solution for the locating task have to be investigated. Thus, the *set cover base station positioning algorithm (SCBPA)* was introduced, which is based on a greedy heuristic for solving the MCLP problem. Additionally, a Simulated Annealing procedure has been considered as an alternative approximation method. Both methods have shown to provide efficiently feasible solutions for the maximal coverage location problem.

To demonstrate the applicability of the procedures proposed in this section, the planning tool prototype ICEPT has been implemented. Due to its detailed implementation of all the design steps, ICEPT can be used for the synthesis of radio network configurations in real world design scenarios.

Furthermore, it was shown that the proposed set covering models contribute a large benefit in specific RF optimization tasks. The use of the models facilitates the application of optimization methods for co-channel interference minimization. Hence, a radio network designer receives valuable support by an automatic procedure. It is anticipated that the application of optimization methods accelerate significantly the task of radio network planning.

Due to the demand node concept and the use of efficient approximation algorithms, the *integrated approach* is able to obey all the RF design objectives as well as the capacity and the network deployment constraints. The new method is able to find trade-offs between different design objectives. It can obtain an overall optimized network configuration in acceptable computing time. Thus, the integrated approach meets

the requirements for the planning methods of future generation networks. The automatic network design enables the integrated approach to generate *synthetic networks*.

As a result of the research presented in this chapter, set covering models have found application in complex radio network optimization tasks, e.g. radio network engineering in CDMA systems, cf. Yu et al. (1998). The methodology is now in commercial use.

7 Call Handling Procedures in Cellular Mobile Networks

The *cellular concept* achieves a larger system capacity by the reuse of the same radio frequencies in geographically separated cells, cf. Chapter 2. The drawback of the cellular concept is the increased system complexity. The network must be able to handle *handover* events: on-going calls which move across cell boundaries have to be transfered into the adjacent cell without interruption. The event of a new call in a cell, either originating or terminating at the mobile station, is said to be a *fresh call*, cf. Figure 7.1. The schemes which define the modus of processing the calls are denoted as *call handling mechanisms*.

The increased system complexity of cellular communication networks, requires an adequate design and engineering methodology. In particular, the performance evaluation of call handling mechanisms has to be carefully assessed. The performance of these mechanisms depends mainly on the following design factors.

First, the performance is influenced by the capability of the system architecture, e.g. by the complexity of the handover protocol and by the performance of the involved transport network elements. In the second place, it is affected by the behavior of the customers within the service area. In order to get a good insight into the performance of the system, one has to take into account the distribution of interarrival times and service times, i.e. *traffic mix*, e.g. the fraction of call requests being handover and fresh calls, respectively, the user mobility, etc. And third, the performance depends on the actual configuration of the network. Loosely speaking, if the architecture has been determined, the parameters of the system have to be chosen, e.g. the network planner has to dimension the number of channels allocated to a cell. Furthermore, the strength of the influence of these factors depends not only on their direct effect. The factors interact and amplify each other, e.g. the service availability will

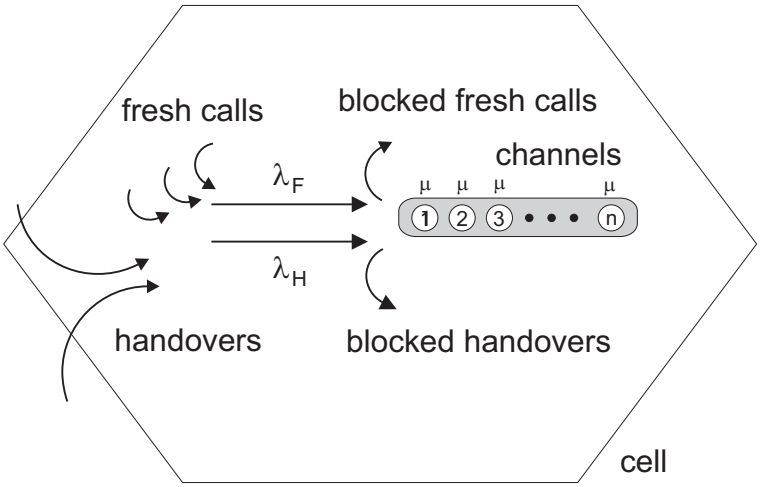


Figure 7.1: *Contenting fresh calls and handovers*

increase the service demand.

So far, the proposed models for the performance analysis of call handling mechanisms are not very accurate. They address only few of the above mentioned factors. Moreover, the factors are are addressed mostly isolated, rather than in a comprehensive way. These deficiencies lead to inefficient system configurations. Hence, it is anticipated that the use of more sophisticated call handling mechanisms can enhance significantly the performance of cellular mobile networks. Apart from this, the application of more detailed traffic models shows that the actual load is considerably higher than estimated by the conventional traffic models. Thus, the real performance of common systems is lower and a larger efficiency requires better call handling mechanisms.

In this chapter, a design method is presented, which is capable to integrate the three above mentioned engineering aspects for call handling procedures. First, in Section 7.1, two advanced handling mechanisms, *guard channel* and *handover retry*, are presented which prioritize handovers, in order to enhance the system performance. However, as opposed to earlier studies on this subject, the presented analysis is using a finite population customer behavior model that enables blocked calls to redial. The customer model is outlined in Section 7.2. In Section 7.3, the conventional traffic engineering method for call handling mechanisms is presented and the constraints for advanced configuration procedures are discussed. Section 7.4 presents a performance analysis of various call handling mechanisms and evaluates their main Quality-of-Service (QoS)

parameters with respect to the enhanced user model. Being able to calculate this QoS, one can choose the configuration, e.g. the number of channels, the number of guard channels, and the parameters of the handover retry procedure. Finally, Section 7.5 concludes this chapter with a summary of the presented results and an outlook to future research work.

In this chapter, the user initiated repeated attempts to occupy a channel are termed *redialing* and the system triggered attempts are termed *retrying*.

7.1 Call Handling Mechanisms

Call handling mechanisms have to perform two tasks. First, they have to manage reliably the admission of fresh calls and handovers to a cell. Second, call handling mechanisms have to ensure that the system is able to operates at high traffic load, i.e. a good handling mechanism has to admit as many calls as possible. In this section, first, the conventional call handling mechanisms used in nowadays cellular systems like GSM, cf. Mouly and Pautet (1992), are outlined. Then, the advanced procedures, which are considered to be included in third generation cellular networks, are discussed.

7.1.1 Conventional Handover Mechanisms

In most cellular networks handover events are managed in the following way: the base station and the mobile station measures regularly the radio signal strength and the mobile station transmits its measurements to the base station. If the base station detects a decrease in radio signal under a minimal secure level d_{urge}, cf. Figure 7.2, it initiates a handover request. Usually this type of call transfer is denoted by the term *rescue handover*. The base station informs the *base station controller* about the request, which than verifies if it is possible to transfer the call into a new cell. To this end, the controller checks whether there is a free channel available in the new cell, or not. Usually, the base station controller does not distinguish between channel requests for fresh calls or handovers. If a handover request can be satisfied, the controller informs the mobile station to switch to the new cell. If no channel is free, the call is interrupted and lost.

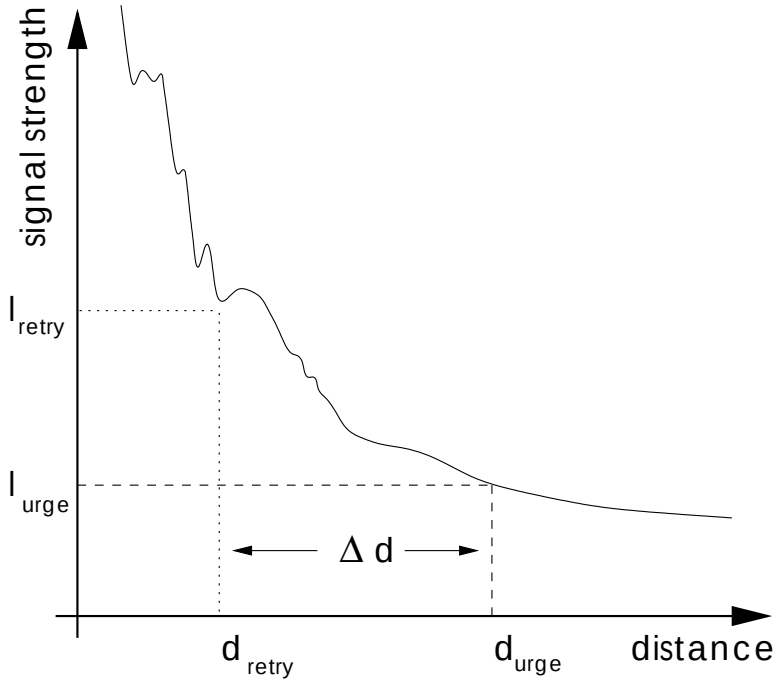


Figure 7.2: *Signal levels for handover commitment*

The drawback of this handover procedure is the fact that the handover requests contend for the same channels which are used for fresh calls, cf. Figure 7.1. Thus, the reliability of this mechanism depends on the load of the new cell. Moreover, since handovers are more valuable than fresh calls, cf. Section 7.1.2, a lost handover decreases the quality of service of the cellular system much stronger than a blocked fresh call.

7.1.2 Advanced Handover Procedures

The objective of advanced handover procedures is to ensure on the one hand a highest possible traffic load of the cellular system, while one the other hand maintaining a certain quality of service. Therefore, in this subsection a definition is given what is considered as the quality of service with regard to call handling and call admission. Then, two advanced call handling mechanisms which are expected to meet these objectives are presented.

a) Quality of Service

In cellular mobile networks, the Quality of Service (QoS) comprises the speech quality as well as the availability of the service within the supplying area. Both factors are mainly determined by the quality of the radio transmission. However, due to the increased traffic in mobile communication networks, the second factor depends more and more on availability of free channels and thus on the appropriate teletraffic configuration of the system.

From the teletraffic point of view the Quality of Service is determined by the probability of the two events which occur due to the occupancy of all available channels:

a) the fresh call blocking probability P_{BF} and,

b) the handover dropping probability P_{BH}.

Because there is a tradeoff between these two performance measures and the configuration, an overall *cost function* C can be defined as the weighted sum of the blocking probabilities:

$$C = \alpha P_{\mathrm{BF}} + (1 - \alpha) P_{\mathrm{BH}}, \tag{7.1}$$

where $\alpha \in [0;\ 1]$. The value of α indicates the priority of handovers relative to fresh calls. This performance measure was first used by Chang et al. (1994).

One aspect of the cost function measure defined in Eqn.(7.1) is its focus on blocked handovers. For a subscriber an interrupted call is much more fretful, than a blocked fresh call. Thus, the handover blocking probability should be kept one magnitude below the value of the fresh call blocking probability, which is typically in the order of one percent. To equally include the two probabilities into one cost function, a value of $\alpha = \frac{1}{11}$ was chosen throughout this study, such that $\alpha/(1-\alpha) = \frac{1}{10}$.

b) Guard Channels

The dependence of the handover procedure on the traffic load of a cell can be reduced by applying channel allocation mechanisms which favor handovers in overload situations. An effective mechanism is the use of *guard channels*. Guard channels are established only when the number of free channels is equal to or less than a predefined threshold g, cf. Figure 7.3. In this case, fresh calls are rejected and only handover request

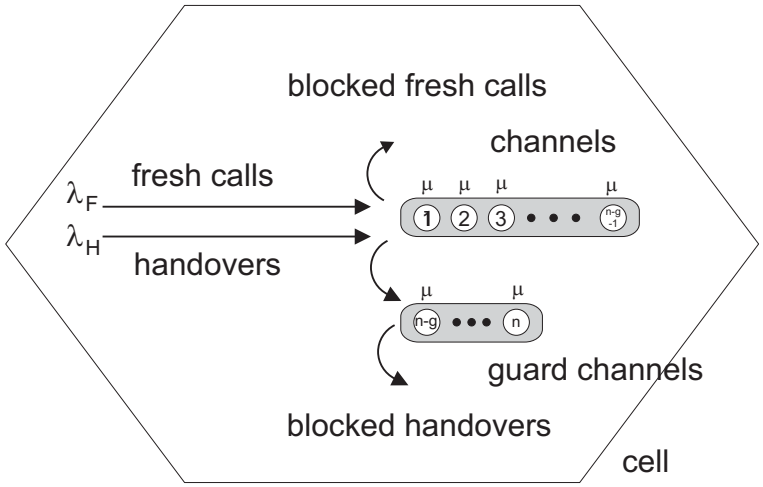


Figure 7.3: *Guard channels for handover request*

are served by the cell until all channels are occupied. As soon as the number of unused channels exceeds the threshold, the cell starts again accepting fresh calls.

The guard channel mechanism directly reduces the probability of dropping a handover. Since guard channels can only be used by handovers, their occupancy depends only on the number of handover requests. However, the reservation of channels for handovers restricts fresh calls from being served and increases their blocking rate.

c) Handover Retry

Handover retry is an other promising mechanism to decrease the blocking probability of handovers. Once a handover has failed, the request enters a retry group, where it attempts a number of retrials before the handover is definitively lost, cf. Figure 7.4. In a real implementation, the number of handovers in the retry group should be limited. If many handovers are waiting in the retry group, it is not very probable that a new handover can be served within the time limit a handover can be postponed.

So far, the handover retry mechanism was not implemented in nowadays networks. However upcoming generations of cellular systems will have two technical features which are using this mechanism. The first feature is denoted as the *two-level handover*. A handover is triggered already when the strength of radio signal falls below a first threshold l_{retry}, even if the signal strength is still sufficient for radio transmission.

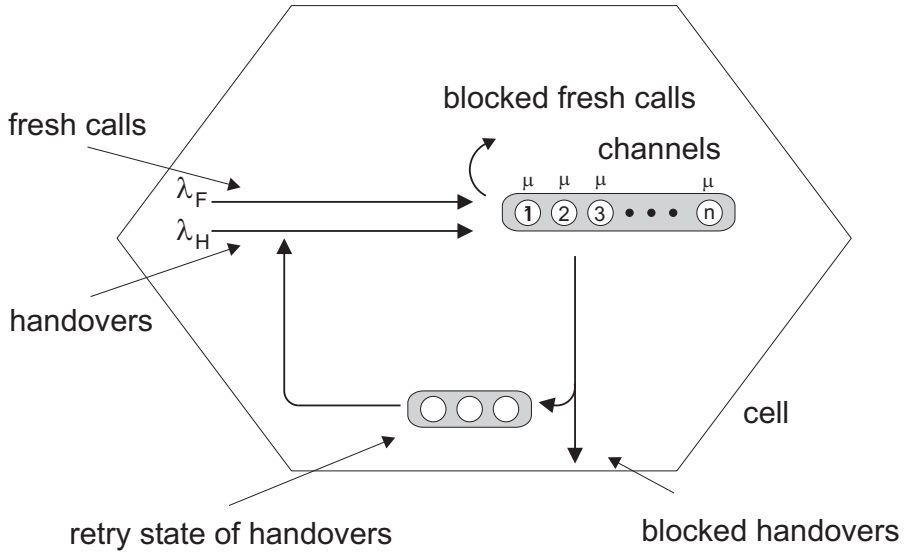


Figure 7.4: *Handover retry mechanism*

The *urgent* handover must be committed, if the radio signal falls below a second threshold l_{urge}, cf. Figure 7.2. The interval, a moving mobile station needs to cover the distance between the threshold points, can be used for processing the request.

The second technical feature is the *call re-establishment*. In a radio mobile environment, a call can be always interrupted because of sudden severe loss of radio signal strength due to obstacles. Often, another cell could be used to continue the transmission and a handover request could be initiated to transfer the call. A similar feature already exists in GSM network specifications, Mouly and Pautet (1992) p. 412ff. However, it is not used due to the high complexity which is required in the controller.

Handover retry can be triggered either by a *centralized* controller or by the handset itself without any user interaction. A centralized controller which manages the retry state is expected to be very efficient. It can occupy resources instantaneously and schedule the requests in an optimal order. However, the mechanism increases the complexity of the controller and is therefore very expensive to implement. A *distributed mechanism* is constitutes if handover retries are triggered by the handsets. For instance, the handset repeats its request every k time units, with a maximum of K times. Here, the complexity of the call handling mechanism is partly moved from the controller and distributed into the handset. The efficiency of such a mechanism is not expected to be opti-

mal, but as the performance analysis in Section 7.4 will show, the mechanism improves the Quality of Service quite well. Furthermore, this type of mechanism can be implemented cost-efficiently in the handset.

7.1.3 Overview on Analytical Models

An early paper on guard channels was published by Guérin in 1988. The model consists of guard channels and a queue of blocked fresh calls. It does not consider any retry state space for handovers. The service of the queue is managed by a central controller. This model was also approached, but from a more computational point of view, by Keilson and Ibe (1995) and Daigle and Jain (1992). Both used a matrix-geometric approach for the analysis.

Chang et al. (1994) considered a more advanced guard channel model. It comprises the guard channel mechanism and two finite queues for waiting fresh calls and handovers. The model allows both types of calls to withdraw from the queues, due to customer impatience or early leaving the handover area, respectively. The queues are managed by a centralized scheduling scheme, i.e. a waiting request is served as soon as a channel becomes idle. A similar model was investigated by Zeng et al. (1994). However, the authors did not consider the reneging of calls due to impatience.

In contrast to the above mentioned authors, Yoon and Un (1993) investigate two call call handling mechanism which prioritize handovers but do not use guard channels. Both models consists of a single queue for fresh calls and handovers. Handovers are allowed, if they can not be served immediately, to drive out fresh calls from the head of the waiting line. The two models differ in the handling of fresh calls. The first model considers a last-in, first-out (LIFO) strategy for the service of fresh calls and the second model applies a first-in, first-out (FIFO) scheme. Yoon and Un (1993) compare both models with a guard channel model without a waiting room. In fact, this is the classical *trunk reservation* model, cf. Roberts (1983).

7.2 Traffic Models for Call Handling Mechanisms

Traffic models for the performance analysis and dimensioning of call handling mechanisms have to address comprehensively the characteristics of the call handling scheme as well as they have to capture accurately the temporal user behaviour. So far these requirements are fulfilled only to a limited extent and in particular the description of the temporal user behaviour is insufficient. In the next section, a comprehensive traffic model for call handling mechanisms is introduced which is capable to meet the requirements.

7.2.1 Single Request Stream Model

The commonly used traffic model for a single cell in mobile communication networks is the *single arrival stream* model. Handovers of on-going calls from adjacent cells and fresh calls are assumed to arrive according to a Poisson process with rates λ_{H} and λ_{F}. Both streams are aggregated in a single sequence of channel requests, cf. Figure 7.1. Again, the aggregated stream is a Poisson process with rate

$$\lambda_{\mathrm{total}} = \lambda_{\mathrm{F}} + \lambda_{\mathrm{H}}. \tag{7.2}$$

In this model, the call duration distribution is assumed to be negative exponential with mean μ^{-1} for both types of traffic.

The major disadvantage of this traffic model is the aggregation of all channel request into one single stream. The model can not be used for the analysis of advanced call handling mechanisms, which distinguish between handovers and fresh calls.

Moreover, the *single stream* model does not consider any temporal user behavior. In overload situations, it is very likely that blocked customer repeat their attempt after a few seconds. Particularly, modern handsets have the capability to execute the redialing of a phone number by just one push of a button. The redial rate is usually one or two magnitudes higher than the normal call arrival rate and thus the offered traffic extremely increases in a very short period of time. It is known from literature on classical switching systems, e.g. Jonin and Sedol (1976) and Macfadyen (1979), that this redial phenomenon dramatically degrades the system performance. An accurate traffic model has to consider this

redial behavior in order to predict the consequences of the phenomenon during system operation.

7.2.2 Repeated Attempts

An important feature of our study is a new user model which considers the impatience of customer and models the redial phenomenon. Therefore, a customer behavior model which was introduced by Tran-Gia (1982) is adopted. It is assumed in the model, that every time a customer is blocked, he waits for the next attempt with probability θ_0 and will give up the request with the probability $1 - \theta_0$. If the customer decides to reattempt, he will try this reattempt after an exponentially distributed time with mean α_0^{-1}. The redial probability θ_0 is not affected by the number of redial attemps. In Tran-Gia (1982) an efficient recursive algorithm is given to analyze a loss model with a finite number of customers. Comprehensive surveys on reattempt queues can be found in Falin (1992) or Yang and Templeton (1987). More recent articles, that specifically study the effect of repeated attempts in cellular mobile networks, are Tran-Gia and Mandjes (1997) and Choi et al. (1995). In contrast to the evaluation presented here, these studies focus on the system degeneration effect of the redial phenomenon.

7.3 Traffic Engineering for Call Handling Mechanisms

A well configured call handling mechanism should enable the cell to operate at high traffic load. However, the disadvantage of running a system close to its limits is the higher vulnerability to a service increase in a very short time scale, e.g. mass calling. Such systems degrade quickly in overload situations. Hence, the application of efficient call handling procedures requires a careful and accurate traffic engineering and dimensioning in these system. In this section, at first, the conventional traffic engineering method for call handling methods in cellular networks is reviewed. Then, the engineering approach for advanced handling mechanisms is discussed.

Conventional Traffic Engineering

The widely used traffic engineering method in cellular mobile networks is the application of the Erlang-B-formula, e.g. Mouly and Pautet (1992):

$$P_{\mathrm{B}} = \frac{(\frac{\lambda_{\mathrm{total}}}{\mu})^n / n!}{\sum_{k=0}^{n} (\frac{\lambda_{\mathrm{total}}}{\mu})^k / k!}. \tag{7.3}$$

The formula relates the offered traffic load $\lambda_{\mathrm{total}}/\mu$ with the number of channels n and the blocking probability P_{B} for all channels requests, under the assumption that the instant of call attempts follow a Poisson process and the call duration is exponentially distributed. The formula does not distinguish between fresh calls and handovers. For cell configuration, it is required that the blocking probability stays below a certain value, in GSM networks typically two percent. For the initial configuration of a cell, where measurements are not available, the offered load λ_{F}/μ is estimated by using the average traffic per subscriber, measured in Erlang, the expected number of subscribers in the cell and the mean call duration. In an operating mobile communication system like GSM, the values of λ_{total} represents the number of call attempts during the *busy hour* per time unit. The busy hour is defined as the four consecutive 15min time intervals with highest number of requests. The mean call duration is also measured during the busy hour.

The major drawback of this traffic engineering method is, that it treats fresh calls and handovers in same way. Therefore the method cannot be used for advanced call handling mechanisms, like the two schemes presented in Section 7.1. Furthermore, since the underlying traffic model does not consider temporal user behavior, the Erlang-B-formula is not even very appropriate in determing the number of required channels for conventional call handling mechanisms. To compensate the insufficiencies, correction factors on the offered traffic λ_{total} can be applied. However, the estimation of these factors is still done by thumb rules and therefore the design remains very inaccurate.

Improved Traffic Engineering

A traffic engineering procedure for advanced call handling mechanisms has to offer the right combination of the parameters n, g, k and K. Fur-

thermore the engineering method has to consider the different objectives, i.e. the two blocking probabilities. Therefore the configuration has to make use *(i)* of an accurate traffic model, like the new one present in Section 7.2.2, and *(ii)* of a detailed analytical model for the mechanisms. The evaluation of such a model is presented in the next section.

7.4 Performance Analysis of Call Handling Mechanisms

In this section, at first, the handling schemes that were presented in Section 7.1, guard channels and handover retry, are described in mathematical terms, using the customer model introduced in Section 7.2. In the second subsection, it is be shown that the network can be modeled as a continuous-time Markov chain. Finally, the section concludes with a graphical evaluation of the efficiency of the advanced call handling mechanisms.

7.4.1 Analytical Model

The analytical model of the investigated advanced call handling mechanism is shown in Figure 7.5. The model comprises a guard channel mechanism as well as a handover retry. The arrival processes are assumed to be Poissonian: the interarrival times of fresh calls and handover calls are exponentially distributed with mean $\lambda_{\mathrm{F}}^{-1}$ and $\lambda_{\mathrm{H}}^{-1}$ and the call duration, i.e., the time before a call terminated or a handover to another cell is attempted is exponentially distributed with mean μ^{-1}. The fresh call redial behavior is modeled as proposed in Section 7.2.2. The handover retry mechanism is similar the one presented in Section 7.1.2. Since this kind of handover retry mechanism is notoriously difficult to analyze, the mechanism is approximated by one which repeats the requests after exponentially distributed time with mean α_1^{-1}. Blocked handovers enter a retry group with probability θ_1, where $\alpha_1 = k^{-1}$ and $\theta_1 = 1 - K^{-1}$. This retry mechanism is similar to the fresh call redial model. The number of handovers in the retry state B is limited, as explained in Section 7.1.2.

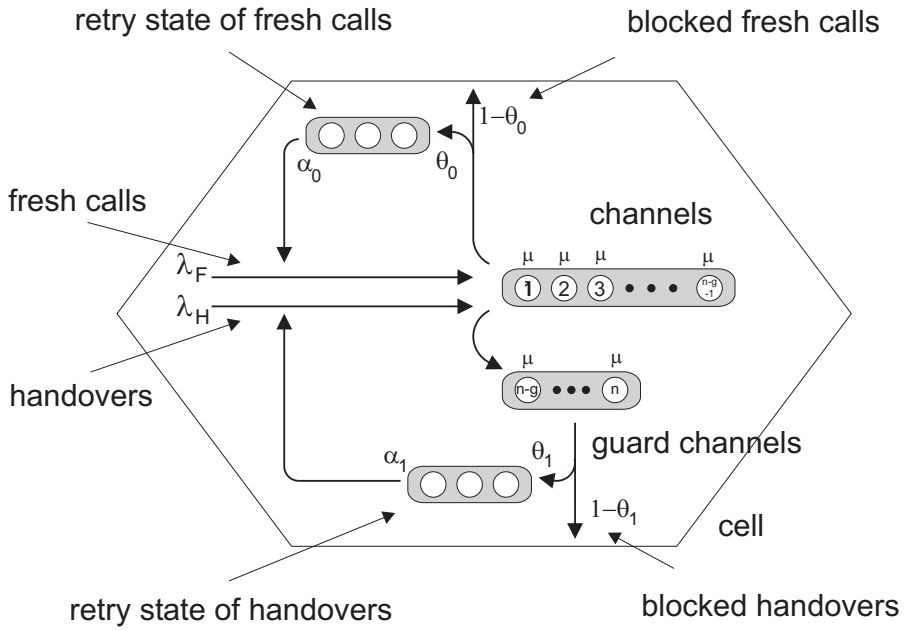


Figure 7.5: *Analytical model of a cell with user redial behavior, guard channels and handover retry*

7.4.2 Markov Chain

For the performance analysis the following model is used. The state of the system is described by the triple (X, Y, Z). The first coordinate X is the number of channels occupied. Obviously, this number varies between 0 and the number n of available channels. The second coordinate Y reflects the number of blocked fresh calls that are redialing. Finally, Z denotes the number of handover calls in the retry queue.

In the following, the state equations are derived, which give the equilibrium distribution of the above Markov chain. Let $x(i,j,k)$ be the equilibrium distribution of this Markov chain, denoted by $P(X = i, Y = j, Z = k)$, where

$$(i,j,k) \in S := \{0,\dots,n\} \times \{0,1,\dots\} \times \{0,\dots,B\}.$$

Let $x(i,j,k) \equiv 0$ when $(i,j,k) \notin S$. Then the state equations are for $i = 0,\dots,n-g-1$:

$$x(i,j,k)(\lambda_{\mathrm{F}} + \lambda_{\mathrm{H}} + i\mu + j\alpha_0 + k\alpha_1) =$$

$$x(i-1,j,k)(\lambda_{\mathrm{F}}+\lambda_{\mathrm{H}})+x(i+1,j,k)(i+1)\mu+$$
$$x(i-1,j+1,k)(j+1)\alpha_0+x(i-1,j,k+1)(k+1)\alpha_1,$$

i.e., in this case in which every new request, both fresh calls and handovers is accepted. For $i \in \{n-g,\ldots,n-1\}$ only handovers are accepted immediately for service. For $i=n-g$, the following equation is obtained:

$$\begin{aligned}x(i,j,k)(\lambda_{\mathrm{F}}\theta_0+\lambda_{\mathrm{H}}+i\mu+j(1-\theta_0)\alpha_0+k\alpha_1)=\\ x(i-1,j,k)(\lambda_{\mathrm{F}}+\lambda_{\mathrm{H}})+x(i+1,j,k)(i+1)\mu+\\ +x(i,j-1,k)\lambda_{\mathrm{F}}\theta_0+x(i-1,j+1,k)(j+1)\alpha_0+\\ x(i,j+1,k)(j+1)(1-\theta_0)\alpha_0+x(i-1,j,k+1)(k+1)\alpha_1.\end{aligned}$$

For $i=n-g+1,\ldots,n-1$, the state transitions are provided by:

$$\begin{aligned}x(i,j,k)(\lambda_{\mathrm{F}}\theta_0+\lambda_{\mathrm{H}}+i\mu+j(1-\theta_0)\alpha_0+k\alpha_1)=\\ x(i-1,j,k)\lambda_{\mathrm{H}}+x(i+1,j,k)(i+1)\mu+\\ x(i,j-1,k)\lambda_{\mathrm{F}}\theta_0+x(i,j+1,k)(j+1)(1-\theta_0)\alpha_0+\\ x(i-1,j,k+1)(k+1)\alpha_1.\end{aligned}$$

For $i=n$, arriving calls of both types cannot be accepted immediately. Thus, the state equation reads:

$$\begin{aligned}x(i,j,k)(\lambda_{\mathrm{F}}\theta_0+\lambda_{\mathrm{H}}\theta_1+i\mu+j(1-\theta_0)\alpha_0+k(1-\theta_1)\alpha_1=\\ x(i-1,j,k)\lambda_{\mathrm{H}}+x(i,j-1,k)\lambda_{\mathrm{F}}\theta_0+\\ x(i,j,k-1)\lambda_{\mathrm{H}}\theta_1+x(i,j+1,k)(j+1)(1-\theta_0)\alpha_0+\\ x(i,j,k+1)(k+1)(1-\theta_1)\alpha_1+x(i-1,j,k+1)(k+1)\alpha_1.\end{aligned}$$

Analogously to the results in Tran-Gia and Mandjes (1997), the fresh call blocking probability and handover blocking probability is given in terms of the equilibrium distribution $x(i,j,k)$. The fresh call blocking probability equals the mean number of blocked fresh calls, i.e., fresh calls that leave the system before being successfully connected per time unit, divided by the mean number of arriving fresh calls per time unit:

$$P_{\mathrm{BF}}=\frac{\sum_{j,k}((1-\theta_0)\lambda_{\mathrm{F}}+j(1-\theta_0)\alpha_0)x(n,j,k)}{\lambda_{\mathrm{F}}}.$$

In the same manner, the handover blocking probability is computed by:

$$P_{\mathrm{BH}} = \frac{\sum_{j,k}((1-\theta_1)\lambda_{\mathrm{H}} + k(1-\theta_1)\alpha_1)(x(n-1,j,k) + x(n,j,k))}{\lambda_{\mathrm{H}}}.$$

To evaluate these performance measures, the distribution of $x(i,j,k)$ must be known. Some comments on how to find this distribution are provided in the next subsection.

7.4.3 Evaluation of the Mechanisms

In this section, the mechanisms described in the previous sections will be graphically evaluated. The following parameters have been chosen: the call termination rate is $\mu = 1/120$ sec^{-1} and the number of channels $n = 15$. The "traffic mix", that is the ratio of the mean number of fresh call requests per unit time and the mean number of handover attempts, is defined to be $\lambda_{\mathrm{F}}/\lambda_{\mathrm{H}} = 24$. For the performance analysis different values for the offered load $\rho = (\lambda_{\mathrm{F}} + \lambda_{\mathrm{H}})/n\mu$, are considered by varying the fresh call arrival rate λ_{F} and keeping the traffic mix $\lambda_{\mathrm{F}}/\lambda_{\mathrm{H}}$ constant.

The evaluation of call handling mechanism is focused on the three performance measures defined in Section 7.1.2: the fresh call blocking probability P_{BF}, cf. Figure 7.6, the handover blocking probability P_{BH}, cf. Figure 7.7, and the cost function C, cf. Figure 7.8, with $\alpha = 1/11$:

$$C = \frac{P_{\mathrm{BF}} + 10P_{\mathrm{BH}}}{11}. \tag{7.4}$$

Subsequently, four analytical models of call handling mechanisms will be considered:

ERL This model does not consider guard channels or handover retries. Blocked fresh calls are not supposed to redial. This model is a special case of the model presented in this section with parameters: $g = 0, \theta_0 = \theta_1 = 0$. The probabilities are calculated by means of the well-known explicit Erlang-B formula, see Section 7.3.

F The second model is the model without any advanced call handling mechanism, neither guard channels nor handover retry, but with redialing of blocked fresh calls. The parameters are set to $\alpha_0 = 0.4$

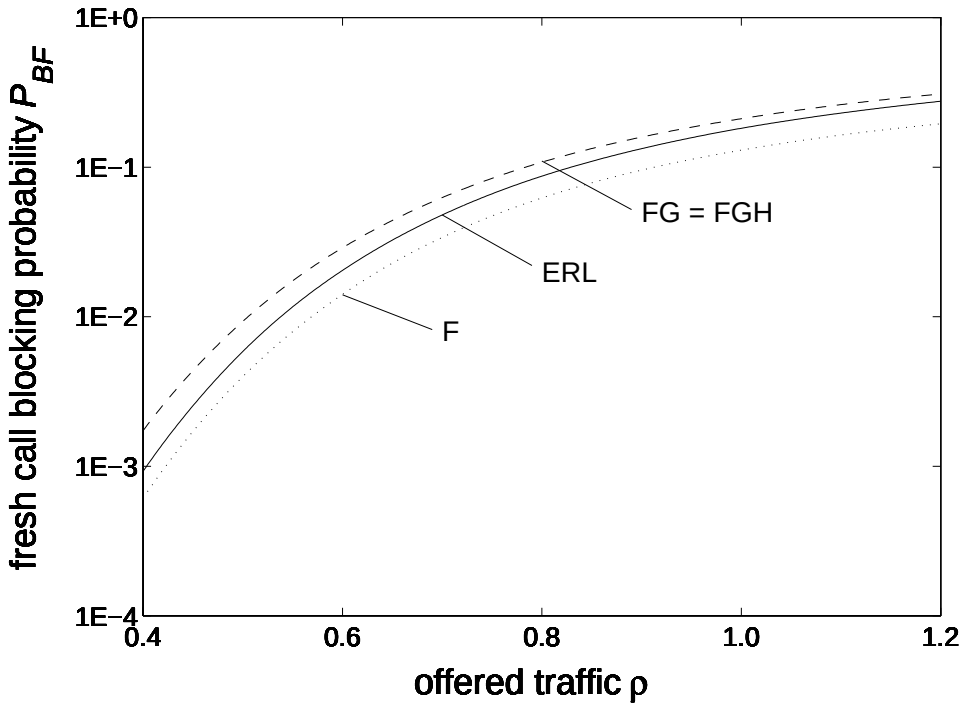


Figure 7.6: *Fresh call blocking probability*

and $\theta_0 = 0.75$. That means that on average four fresh call redials are initiated every 2.5 seconds. The curve with redialing of fresh calls (F) shows that it is very dangerous to neglect the effect of redialing blocked fresh calls: the handover blocking probabilities are considerably larger than for (ERL). On the other hand, of course, fresh call blocking will occur more rarely. However, since the first effect has more impact on the cost function than the second one: the cost function increases increases due to the redialing. The calculations of (F) are done with a recursive algorithm similar to the one presented in Tran-Gia (1982).

FG Two mechanisms were designed to prioritize handovers, as explained in section 7.1. The first is to add guard channels (FG) to the model with redialing of blocked fresh calls (F). The number of guard channels is $g = 1$. The corresponding calculations are done with an algorithm described by Tran-Gia and Mandjes (1997). In this case both handover blocking rates and the cost function decrease.

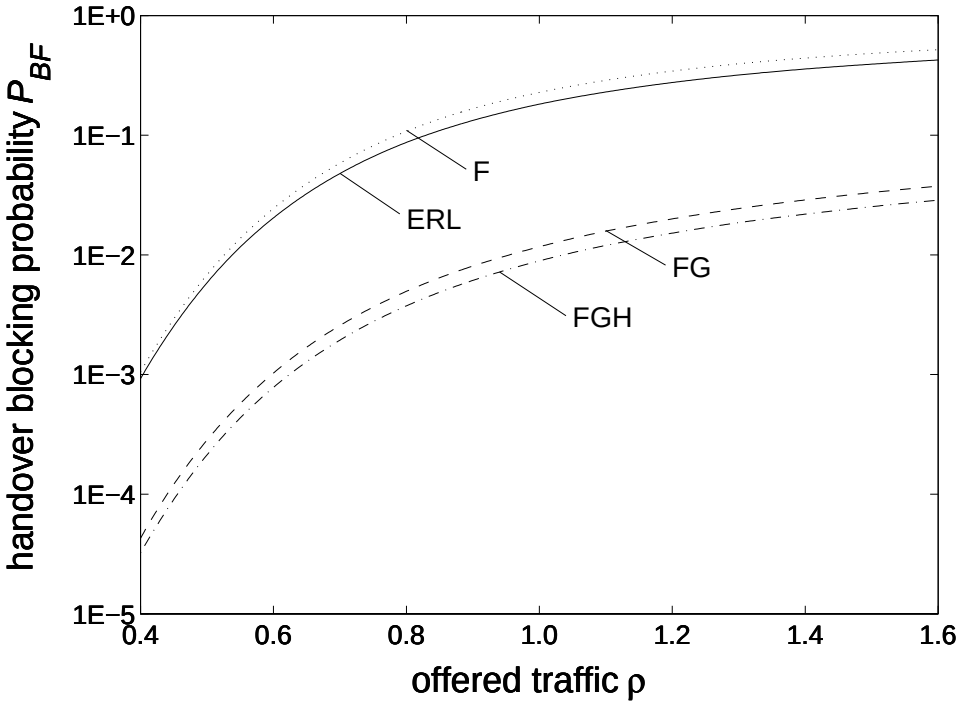


Figure 7.7: *Handover blocking probability*

FGH The second call handling mechanism is the mobile station triggered retry of blocked handovers. The model considered model is shown in Figure 7.5. Again, one guard channel is used, i.e. $g = 1$. Furthermore, the parameters for the retry are $\alpha_1 = 1$ and $\theta_1 = 0.75$. This selection means, that on average a handover repeats its request four times, with a distance of one second. The maximum allowed number of calls being in the retry state, B, is two. The curves (FGH) show the effect of the implementation of this mechanism, in addition to the guard channel, in the model with redialing blocked fresh calls. These results were found by solving the balance equations of the Markov chain of the comprehensive model described in Section 7.4.1. It can be seen that the fresh call blocking probability stays on the same level, whereas the handover blocking rate and cost function C decreases, but not as much as by introducing the guard channel.

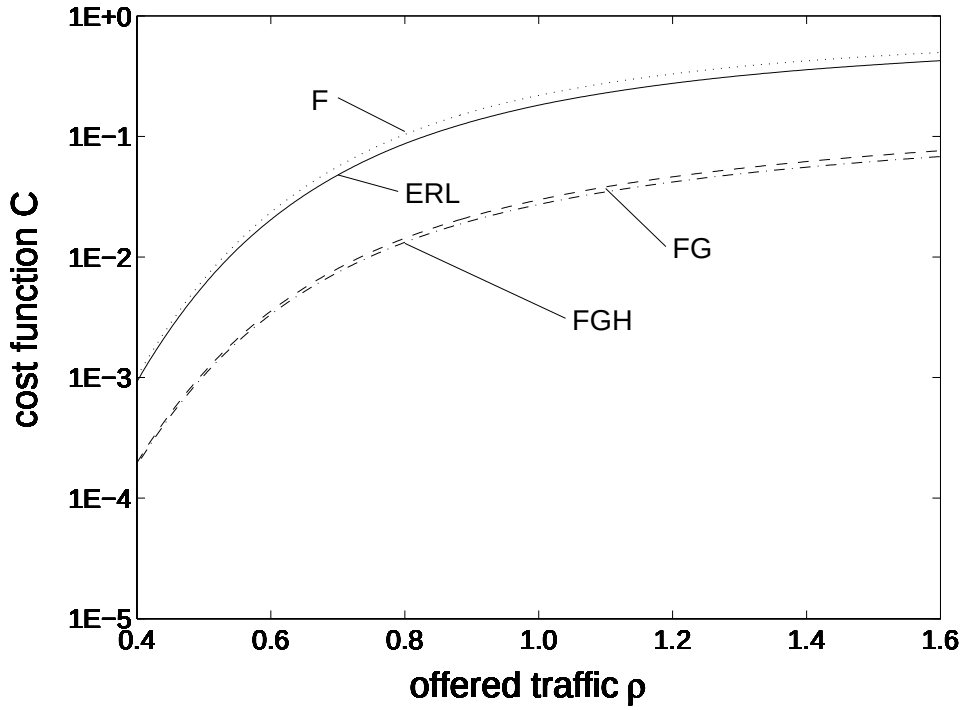


Figure 7.8: *Cost function of the different handling schemes*

Required Number of Guard Channels

Until now, only models have been considered with a single guard channel. Of course, the effect of multiple guard channels is of interest. It is clear that, by increasing the number of guard channels g, the fresh call blocking probability P_{BF} will increase, whereas the handover blocking probability P_{BH} will decrease. Since the cost function C subsummes both, it is not obvious what is the optimal number of guard channels in order to minimize C. In Figure 7.9, the performance of the (FGH) model for different numbers of guard channels is considered. It can be seen that the largest gain is achieved by introducing one guard channel. Using two guard channels instead of one means, for small values of the offered traffic, even a cost function degradation. Of course, this picture depends heavily on the choice of α in the definition of the cost function, but as a general guideline the analysis shows that in most cases one or at most two guard channels are adequate.

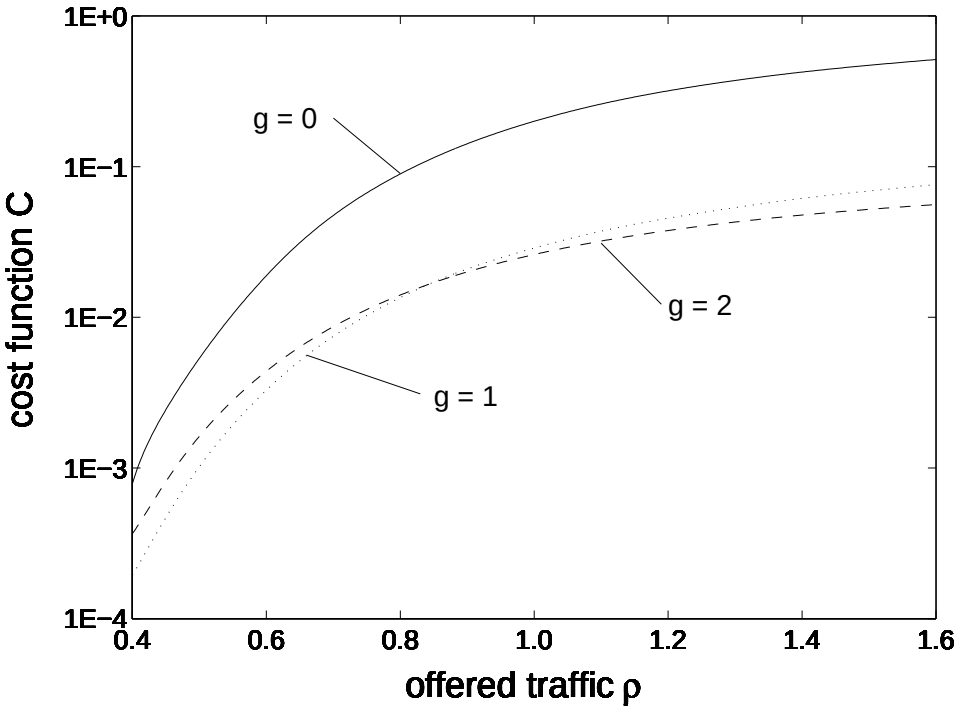


Figure 7.9: *Effect of multiple guard channels*

7.5 Concluding Remarks

The performance of call handling mechanisms in cellular networks depends mainly on three factors: *(i)* the call handling scheme, *(ii)* the user behavior and *(iii)* the proper traffic configuration of the cell. To address the first factor, two advanced call handling schemes, that prioritize handover calls in order to enhance the Quality of Service have been investigated: guard channels and handover retry. As expected and already discussed in the literature, cf. Guérin (1988), the guard channel mechanism improves strongly the performance of the system. Furthermore, the presented performance analysis indicates that in general it is not necessary to use more than one guard channel. An additional performance improvement can be obtained by using a handover retry mechanism. The attractive feature of the mechanism proposed in this chapter is that it is organized in a distributed manner, and therefore does not increase the system complexity considerably. In order to implement this mechanism, only a timer and a counter have to be added in the mobile handsets. The mechanism does not require the controller to manage a centralized

queue.

To address the second factor, the temporal user behavior, a simple but efficient model was presented that can be used to capture the phenomenon of repeated attempts of blocked calls. As discussed in the evaluation of the mechanisms, in order to guarantee the desired blocking rate for handovers, it is critical to neglect the redial behavior of mobile subscriber, as seen from the figures in Section 7.4.

Both adaptations, advanced call handling schemes and improved user modeling, can be used to obtain a more accurate configuration method of the cells in mobile networks. It had been demonstrated in this section that such a system can be modeled as a continuous-time Markov chain and that the solution of its balance equations is mathematically tractable. For different scenarios of the parameters, i.e. number of channels n, number of guard channels g, the probability of a retry of a blocked handover θ_1 and the mean time until the next retry α_1^{-1}, the relevant performance measures can easily be calculated. Thus, a network designer can select them in a way that all service criteria, i.e. the blocking probabilities of fresh calls and handovers as well as the cost function, are kept below a certain predetermined value. The new configuration method increases the efficiency of the cellular system even in high load situations.

Due to the promising capability of the repeated attempt model of Section 7.2.2 to characterize the user behaviour and influenced by the results of the performance analysis of handling mechanisms, the ITU-T's focus group on traffic engineering for personal communications (FG-TEPC) is currently considering to recommend the repeated attempt model as a reference model to be included in the E.750 recommendation series: "Traffic engineering aspects of networks supporting mobile and UPT services", cf. Grillo et al. (1998).

8 ABR Service Engineering in Large Scale ATM Networks

Due to the deregulation of telecommunication markets, cost-efficient service provision is mandatory for network operators. Customers can switch almost instantaneously to the most economical provider. Hence, the network operators have to put a communication service product on the market which is cost-efficient for the provider and offers a considerable benefit to the customer.

In particular, this proposition is valid for ATM network operators. So far, ATM service is highly expensive, due to its high system complexity. Currently, it is not attractive for a large community of customers. To facilitate a low-priced ATM service provision, an additional service category was included in the ATM service architecture, the *Available Bit Rate (ABR)* service category, cf. ATM Forum (1994). The basic idea of ABR is to exploit the excess bandwidth in the network which is not used by other service classes, on an "on-availability" and "best-effort" basis. At first, the ABR service category was intended for sporadic use. Mainly the conventional categories CBR and VBR are supposed to be booked by customers. However, due to its low price, ABR is expected to attract a large number new customers. Thus, the provision of ABR service might become, from a commercial point of view, as important as offering the conventional service categories.

At a first glance, ABR service planning states a contradiction. In conventional ATM network planning, the reduction of the unused capacity is the major design objective. In addition, its almost impossible to predict the availability of unused bandwidth, since the exact amount depends on the instantaneous use of the network. In contrast to this, the provision of ABR service to a large number of users requires a high service availability and an assurance of a certain quality for the service.

These objectives can only be achieved by allocating dedicated bandwidth to ABR and by appropriate service planning. The purpose of this chapter is the presentation of a new approach for ABR service engineering in large scale ATM networks.

In contrast to earlier investigations of the ABR service category, cf. Chen et al. (1996) or Ritter (1998), the focus of the work presented here, is on connection level engineering. Therefore, the investigated performance criteria for ABR are, from a customer's viewpoint, the connection blocking probability and the throughput.

The chapter is organized as follows. Section 8.1 is devoted to the foundations of the ABR service category. In Section 8.2 the ABR service model investigated in this chapter is outlined. In Section 8.3 three methods, two exact and one approximate, for the performance evaluation of the ABR service on connection are introduced. Section 8.4 presents a case study for ABR service planning in large ATM networks. The chapter is concluded by Section 8.5 which presents a summary of the results.

8.1 The ABR Service Category

The VBR and CBR service categories are supporting a large variety of applications with high service requirements, cf. ATM Forum (1994). The VBR service class offers connections up to a specified peak cell rate for variable-bandwidth real-time or non-real-time applications, e.g. data or compressed video. Cells can be generated at arbitrary intervals and are delivered by the ATM network within defined limits for the cell delay and the *cell loss ratio (CLR)* as required by the application. In addition to the *peak cell rate (PCR)*, the burstiness of VBR sources can be indicated by two optional parameters, denoted as *sustainable cell rate (SCR)* and *burst tolerance (BT)*. Together, these values specify upper bounds on the average rate and burst length.

The CBR service category can be viewed as a special case of the VBR class with the PCR equal to the average rate. CBR services are used for emulating fixed-bandwidth circuits for real-time applications, e.g. voice, or reliable point-to-point connections. Cells generated at regular periodic intervals are delivered within strict bounds for the end-to-end cell delay and the cell delay variation.

In contrast to the VBR and CBR services, the UBR service category is intended for applications which have minimal service requirements. The UBR service does not require any prior information about the ex-

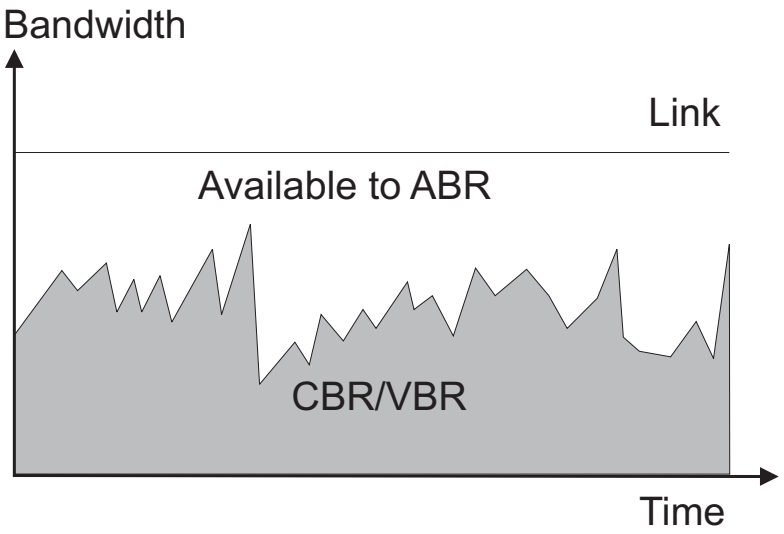


Figure 8.1: *Example of excess bandwidth for ABR traffic*

pected traffic characteristic except the PCR. No commitments for the cell delay, the cell loss ratio, or the minimal allocated bandwidth are made by the ATM network. The UBR service category offers a pure *best-effort* service. By selecting this category, the user accepts whatever bandwidth and cell loss can be provided at the instance the cell is transmitted by the system, cf. Ritter (1998).

The ABR service category is designated for a more specific class of applications, cf. Chen et al. (1996). It is intended to fill the gap between the strict service guarantees of VBR/CBR and the very loose service provision of UBR. The ABR service category is designed for applications which are able to adapt to time-varying bandwidth, and can tolerate unpredictable end-to-end cell delays, but need a minimum throughput, cf. ATM Forum (1995). For example, this category is suited for the transport of TCP/IP data traffic, since the performance of the TCP/IP protocol is highly sensitive on cell loss but robust against variations in cell delay. The ABR service provides a strict guarantee on the cell loss ratio but does not specify any limits on the cell delay or the cell delay variation.

The basic idea of the ABR service is to exploit the bandwidth which exists in the network in excess of CBR/VBR traffic. Figure 8.1 depicts one approach which is used to define the excess bandwidth in a physical link. The available bandwidth for the ABR service will fluctuate

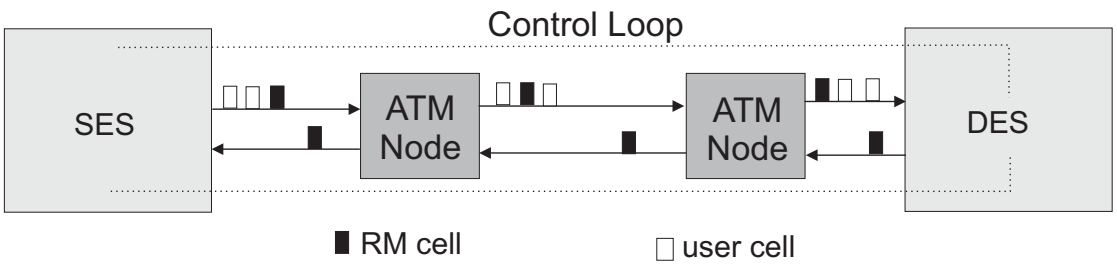


Figure 8.2: *Control loop for an ABR connection*

dynamically due to the randomness in the CBR/VBR traffic.

The share of available bandwidth for a single ABR connection is dynamic and may diminish down to the *minimum cell rate (MCR)*, which is specified by the user. However, the user is not required to always transmit at MCR. The network ensures that the bandwidth available to the connection will vary between the peak rate and the MCR, but not below the MCR. Hence the ABR sources have to adapt continuously the rates to their share of the time-varying excess bandwidth. This share is denoted as the *allowed cell rate (ACR)*.

The rate adoption is performed by a *feedback flow control mechanism* between the source end system (SES) and the destination end system (DES), cf. Figure 8.2. The traffic source adapts to the changing load characteristics by receiving feedback information from *resource management (RM)* cells which are send periodically by the source and looped back by the destination system. The RM cells are modified when they notice congestion. Due to the transfer delay of the RM cells, there is a potential mismatch between the available bandwidth in the ATM network and the source rate. Thus, sufficient buffers are required in the switches to absorb the ABR traffic during the momentary intervals of network congestion. The end-to-end performance in course of an ABR connection highly depends on the performance of the flow control procedure and the proper dimensioning of the switches and their buffers.

For the rate adoption, the ATM Forum has developed a *rate-based* flow control mechanism with two different methods for signaling the congestion, cf. ATM Forum (1995): *a)* the flow control method with congestion indication and *b)* the explicit rate flow control mechanism. The flow control method with congestion indication is using a single bit in the RM cell for signaling, whereas the explicit rate flow control mechanism notifies the source end system with the *explicit rate (ER)* for

the cell emission. Both methods differ in their performance, however they provide a reasonably fast convergence in adapting the source rate to the available bandwidth. The convergence time can be neglected in the time frame of the investigated ABR network planning model, cf. Section 8.2. A detailed description and performance analysis of the ABR feedback flow control mechanisms in ATM systems is provided by Ritter (1998).

The UBR service class and ABR service class represent a substantially different approach of using ATM networks. Whereas CBR and VBR connections require the transmission capacity to be exclusively allocated, UBR and ABR connections will share the remaining bandwidth resource. However, UBR service provides no guarantees on the quality of service. Therefore this service category is limited in its commercial use when reliability is an important issue. In contrast to this, ABR service combines the advantage of a best-effort service with providing assured minimal reliability. The ABR service facilitates an economic use of ATM networks. Due to this characteristic, network operators are highly interested in offering this service to customers despite its increased complexity. In order to handle the higher complexity, new planning procedures are required for ABR service provision.

8.2 ABR Network Model

Without the loss of generality, the investigated ATM model for ABR service planning comprises the basic service categories CBR, VBR, and ABR. The first two services classes, i.e. CBR and VBR, are considered in context of ABR service planning to be the background traffic. The traffic of the ABR source is regarded as the foreground traffic. Furthermore, the ABR connections are distinguished into two different types of calls: *a)* ABR connections with *well known holding time* at call set-up, denoted as *time-oriented traffic*, e.g. LAN-to-LAN inter-connection service during business hours, and *b)* ABR connections with a *fixed amount of data* to be transmitted, denoted as *volume-oriented traffic*, e.g. traditional file transfer. Of course, the connection holding time of the second class of ABR calls depends on the current network load.

Example Network

The considered ATM core network model consists of *core edge switches* and *transit nodes* which are interconnected by ATM links. The

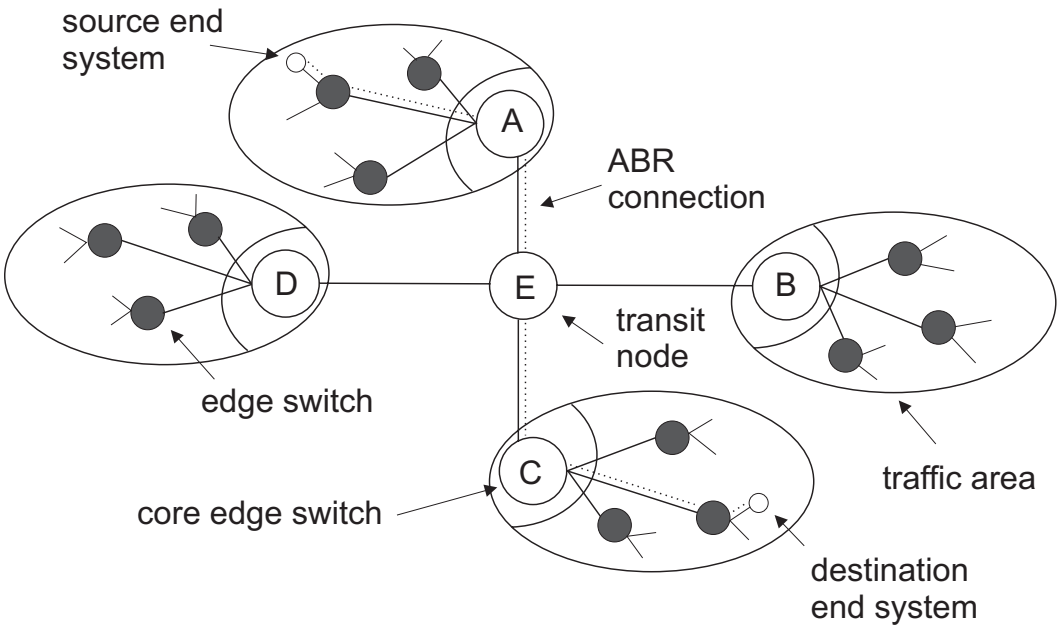


Figure 8.3: *ATM network model for ABR service planning with a star-like structure*

ATM links are defined by their transmission capacity and the capability of the switches at the endpoints of the links. It is assumed that the switches are configured with a sufficiently large number of buffers so that virtually no cell loss will occur in the system due to congestion. The core edge switches are connected to at least one *edge switch*, whereas the transit nodes are not connected to any edge switch. The transit nodes are pure switching facilities; no connection commences or terminates at these locations. An example network is depicted in Figure 8.3. An ABR connection starts from the source end system, passes on to an edge switch and further to a core edge node. From the core edge node, the connection is transferred through the core network, which may consists out of transit nodes and core edge systems, and on to the terminating core edge node. In the investigated model, the subnetwork common to one core edge system is denoted to be a single *traffic area*. The set of all links of the network is denoted by $\mathcal{L}$. For the ABR planning scenario, only *virtual connections* between the core edge nodes and their corresponding traffic areas are considered. It is not relevant over which edge systems the SES are connected to the core edge nodes. This simplification is based on the assumption that usually the ABR service provider has no influence on

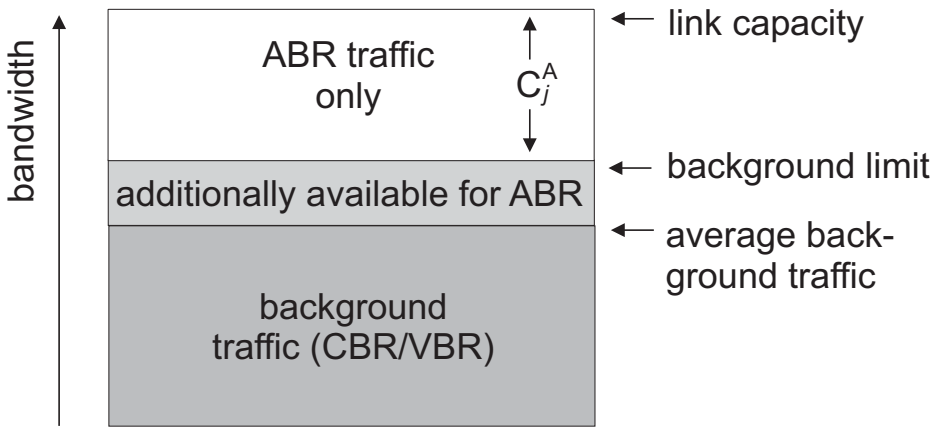


Figure 8.4: *ABR service trunk model*

the planning of the user premise equipment.

ABR Service Model

The service model for ABR network engineering comprises the above mentioned CBR, VBR, and ABR services categories. Each link is supposed to be divided into two main parts: *a)* a part assigned for CBR/VBR traffic, i.e. the background traffic, and *b)* a fraction of the trunk devoted only to ABR traffic, i.e. the foreground traffic. Figure 8.4 shows an example of the engineering model a for single trunk in the ABR planning scenario. A large portion of the link, depicted in the lower part of Figure 8.4, is reserved for the background traffic. The remainder of the link, the upper part in Figure 8.4, is devoted to ABR traffic. Here, a distinction into a segment exclusively devoted to ABR connections (most upper part), and a segment which is borrowed if available by the ABR connections from the CBR/VBR traffic. The capacity on link $l_j \in \mathcal{L}$ exclusively reserved for ABR traffic is denoted by C_j^A. The frontier between the background traffic and the foreground traffic, i.e., the background limit, is permeable in the direction from the foreground traffic to the background traffic. This means that the ABR connections are allowed to exploit the excess bandwidth if and only if it is not used by CBR/VBR connections. The sum of the MCRs of all ABR connections passing through the trunk may not exceed the portion which is exclusive for ABR. Without the loss of generality, the background traffic on a trunk can be defined by two parameters, the sum of the *guaranteed bandwidth* and the sum of the *average bandwidth* used on the trunk.

Since CBR and VBR connections are not always using the reserved bandwidth completely, sophisticated traffic engineering procedures can exploit this feature. In economic network design only a fraction of the guaranteed bandwidth for CBR/VBR traffic is actually reserved on the trunks. In this way, network resources are saved and can be used for the transport of other connections. The ratio between the guaranteed bandwidth and the average used bandwidth considered in network design is known as the overbooking factor:

Definition 8.1 : Overbooking factor

The *overbooking factor* n is the ratio of the guaranteed bandwidth and the bandwidth which is reserved in network planning.

However, due to average case assumptions, the application of the overbooking factor is rather difficult in ATM network design. On the one hand, if the overbooking factor is too small, the network is not efficient enough. On the other hand, i.e. if overbooking is too large, the probability of violating the assured QoS parameter will remarkably increase.

Foreground Traffic

The ABR traffic in the model, i.e., the foreground traffic as defined above, is described by a set of virtual connections which are active at a certain instant in time. Each of the virtual ABR connections departs and terminates at a traffic area, respectively at their corresponding core edge nodes. Connections departing and terminating at the same core edge are not considered. The route of the virtual connection is fixed for its duration. Hence, every ABR connection is described by its MCR and its path.

All virtual connections with identical path and equivalent MCR are accumulated to a set of connections, denoted as a *connection class*:

Definition 8.2 : Connection class

A *connection class* v_i is a set of connections with identical MCR and equivalent path.

The decision operator $\delta_i(j)$ indicates whether connection class v_i is using

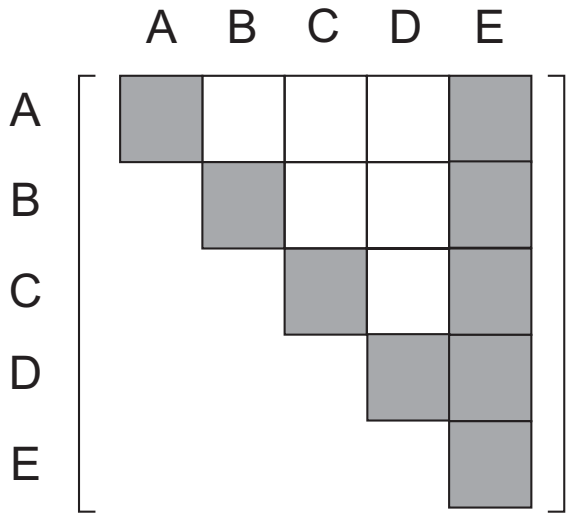


Figure 8.5: *End-to-end traffic matrix*

link l_j or not:

$$\delta_i(j) = \begin{cases} 1 & : \quad l_j \in \text{ path of connection class } v_i \\ 0 & : \quad \text{otherwise} \end{cases} . \tag{8.1}$$

Next, the *end-to-end traffic matrix* $\mathcal{V}$ can be defined. The element $\mathcal{V}_{ij}$ of the traffic matrix denotes all connections with a source in traffic area i and destination in traffic j. An additional condensation of the traffic matrix is obtained by a combination of connection classes with equivalent path on the main path and on reverse direction. Due this symmetrical traffic assumption, the traffic matrix takes the shape of an upper triangle matrix with an empty diagonal. Figure 8.5 depicts the structure of the end-to-end traffic matrix corresponding to the network of Figure 8.3. White squares stand for valid entries, grey squares invalid ones.

The other ABR parameters, like the cell loss ratio (CLR), the additive increase rate (AIR), or the rate decrease factor (RDF) are not considered in the investigated ABR model, since they mainly refer to the flow control mechanisms. The PCR of a connection of class v_i is assumed to be the capacity in the link along the path of the connection with the least bandwidth available for ABR. Therefore, the peak cell rate (PCR)of a connection of class v_i is:

$$PCR_i = \min_{\forall l_j : \delta_i(j)=1} \{C_j\};\ j = 1, \ldots, M\,, \tag{8.2}$$

where C_j is the capacity on link l_j available for ABR (including the capacity not used by CBR/VBR connections) and M is the number of links in the network.

Bandwidth Sharing

The above introduced ABR service model assumes an efficient sharing of the ABR fraction of the trunk among the active ABR connections. To share the available bandwidth, different policies have been proposed in literature, cf. Roberts et al. (1996). The bandwidth sharing policy adopted for the investigated model is *complete sharing without bandwidth reservation*.

8.3 Methods for ABR Service Design

The economic provision of ABR service to a large number of customers requires a careful design of the ATM system. The increased complexity of the ABR mechanism with its sophisticated bandwidth sharing policy has to be addressed during the network planning phase. Since the ABR traffic model differs significantly from the traffic scenario of the conventional CBR/VBR services in ATM networks, in this section first the investigated ABR traffic model is outlined and secondly approaches for ABR service planning are discussed.

8.3.1 ABR Traffic Model

So far, most network planning cases assume *static* traffic scenarios, i.e. the number of connections and the amount of traffic transferred over the network is constant throughout the modeling time. In ATM networks with a large amount of ABR traffic, however, this assumption can not be applied anymore. Due to its nature, the ABR service can be requested by the customer on a very short notice. Additionally, the arrival rate for ABR service requests is expected to be much higher than the rate of establishing conventional CBR/VBR connections. Thus, the foreground traffic scenario changes continuously, whereas the background traffic can be regarded as static.

The considered traffic scenario for ABR service engineering must be of a detailed *stochastic* nature. ABR service requests and releases at a core edge node have to be modeled in such a way that they follow

a stochastic arrival process. Since the investigated ABR service model, presented in Section 8.2, comprises two different types of ABR connections, time-oriented and volume-oriented calls, this distinction has to be included into the ABR traffic model. The definition of the term connection class has to be extended by the stochastic attributes of the classes.

Time-oriented Stochastic ABR Traffic

In the time-oriented stochastic ABR traffic model, the connection holding time does not depend on the amount of transmitted data. The holding time is given directly by a distribution function. It is assumed that the holding time T_H follows a negative-exponential distribution with parameter μ. The mean of the connection holding time is $E[T_H] = 1/\mu$. The arrival process of ABR connections at the core edge switch is modeled as a stationary Poisson process with parameter λ. For easier notation the offered traffic load $a = \lambda/\mu$ is introduced. Thus, in time-oriented stochastic ABR traffic model a connection class v_i is described by a 3-tupel consisting of:

- the path of the connection,
- the MCR denoted by MCR_i, and
- the offered load a_i.

Volume-oriented Stochastic ABR Traffic

In the volume-oriented stochastic ABR traffic model the connection holding time is not known at call set-up. The holding time depends on the amount of data to be transmitted as well as on the network load during transmission. The amount of data is modeled by a random variable which is negative-exponentially distributed with mean v^{-1}. As in the time-oriented model, the call set-up process is assumed to be a stationary Poisson process with parameter λ. Hence, in the time-oriented stochastic ABR traffic model a connection class v is described by a 4-tupel:

- the path of the connection,
- the MCR denoted by MCR_i,
- the call arrival parameter λ_i, and

- the mean data volume v^{-1}.

It is important to state at this point that all transmitted cells are assumed to be payload cells. The signaling and resource management cells are neglected. Moreover, the overhead produced by the ATM cell header is already included in the transfer volume described by the distribution function. The allocated rate for a connection at any time is used completely to decrease the data volume. Volume-oriented ABR sources are saturated for the whole transfer time. Similar assumption are also made for the time-oriented traffic model.

8.3.2 Common ABR Service Planning

At a first glance, the term "ABR service planning" seems to state a contradiction in itself. ABR's basic idea is to exploit the excess bandwidth which remains after conventional CBR/VBR traffic engineering by an "on-availability" principle. In this context, ABR service planning can be referred to as ABR service engineering at the moment when an ABR connection is requested. Therefore, the most important issue in conventional ABR service planning on connection level scope is the fair and fast allocation of the available bandwidth to the various virtual ABR connections currently active. Usually, in this context, a static traffic scenario is considered, i.e., the foreground traffic as well as the background traffic is assumed to be fixed during the considered time. The traffic engineering for the background traffic is usually done by applying the overbooking principle, cf. Section 8.2.

Fair Rate Allocation

The ATM forum adopted the well known *max-min* fairness criterion for the proper allocation of the excess bandwidth in ATM networks, cf. ATM Forum (1995). An intuitive definition of max-min fairness is provided in Bertsekas and Gallager (1987):

Definition 8.3 : Max-min fairness

Max-min fairness in rate allocation is the maximization of the allocated bandwidth of each connection i subject to the constraint that an incremental increase in i's allocation does not cause a decrease in some other connection's allocation that is already as small as i or smaller.

In the context of the ABR model considered in this chapter, the max-min criterion attempts to equally allocate the available bandwidth among all connection classes *bottlenecked* at a link. Fairness is achieved by allocating an equal share of link bandwidth to all connection classes provided they can use this fair share. The term *bottlenecked connection classes* defines those connections classes which are unable to achieve their fair (equal) share of bandwidth at a link because of constraints imposed by their PCR requirements or, most likely, by the limited bandwidth available at other links. The efficiency of the max-min principle is the maximization of the throughput for all connections that have minimum allocations in the network. The *fair share (FS)* can be computed as follows, cf. Arulambalam et al. (1996):

$$\textit{Fair Share} \;=\; \frac{C_j - \sum \text{ rates of connec. classes bottlenecked elsewhere}}{N_j - \sum \text{connec. classes bottlenecked elsewhere}} \tag{8.3}$$

where C_j is the capacity of link l_j available for sharing and N_j is the number of connection classes using link l_j.

Definition 8.3 provides a blue print for the formulation of an iterative max-min fair share rate allocation algorithm:

Step 1 Find the equal share for the connection classes on each link

Step 2 Find the connection class(es) with minimum allocated rate

Step 3 Subtract this rate at the link and eliminate the connection classes with minimum allocation

Step 4 Recompute an equal share of each link in the reduced network

Step 5 Repeat procedure 2-4 until all connection classes are eliminated

The result of the max-min rate allocation algorithm is the fair share of the capacity to be distributed. The explicit rate is the sum of the fair share and the MCR.

An important criterion for rate allocation algorithms is their *convergence* behaviour. It is required to be sufficiently fast in order to assure an efficient data transmission. In general there exist two approaches for rate

allocation procedures: *a)* approximation algorithms, revealing a fast convergence which is traded for accuracy, and *b)* exact fair rate calculation procedure, which obtains optimal rates, but require global information. Without the loss of generality, the ABR model investigated in this chapter considers an *exact fair rate algorithm*. The detailed description of the fair rate allocation algorithm used in this study is provided in Staehle (1999).

8.3.3 Design Objectives

The main objective of large scale ABR service engineering is to facilitate the provision of the ABR service to a commercially significant number of customers. Hence, ABR service engineering has to deal with a considerably large number of on-going ABR connections in the network. This characteristic requires an appropriate traffic engineering which has to be based on an accurate performance analysis of the ABR service model.

From a customers' viewpoint, the provision of reliable ABR service requires a high *availability* of the service. The users' connection request has to be satisfied by the network at any time with a sufficiently high probability. Furthermore, the chance of using the ABR service is substantially increased if the users' prospect of obtaining *additional bandwidth*, in excess to the MCR, is considerably large.

From the perspective of the ABR service provider, large scale ABR service engineering, on the one hand, has to meet the customer requirements. On the other hand, it has to ensure a high network utilization. This means that network engineering is required to obtain a high link utilization as well as to enable as many ABR connections as possible to be in the system at the same time.

Customers' View

In detail, two major design objectives can be specified from the viewpoint of the user: *a)* a low connection blocking probability and *b)* a large sustainable cell rate.

Due to the application of the complete sharing without a bandwidth reservation policy, cf. Section 8.2, the event of blocking a connection of class v_i in the considered ABR model is defined as:

Definition 8.4 : Blocking of a connection of class v_i

Let the parameter C_j^A denote the capacity on a link l_j dedicated exclusively for ABR traffic. Let n'_k be the number of ABR connections of class v_k at the instance of the arrival of a new connection request and N the total number of connection classes. The arriving ABR connection request of class v_i will be blocked iff:

$$\exists j : \left(\delta_i(j) = 1 \wedge (C_j^A - (\sum_{k=1}^{N} \delta_k(j) \cdot n'_k \cdot MCR_k) < MCR_i) \right) \quad (8.4)$$

otherwise the connection is accepted.

The resulting blocking probability for connections of class v_i is denoted by B_i. Blocked connection requests are assumed to be cleared from the system and do not enter the system again.

The second design objective from a customer perspective is the obtained sustainable cell rate (SCR). The SCR parameter describes the average source rate which is allocated to a connection by the maxmin fair share algorithm. In the investigated ABR model, the SCR is the average explicit rate (ER) of the explicit rate flow control mechanism, cf. Section 8.1.

Providers' View

The main design objective from the provider's viewpoint is a high utilization of the network or the links respectively. In the context of the ABR service model considered in this chapter, the utilization of a link can be characterized by two different parameters: *a)* the MCR utilization and *b)* the ABR utilization.

The *MCR utilization* of link l_j describes the used fraction of the part of the trunk which is exclusively reserved for ABR traffic. The average MCR utilization is of great interest in network engineering. This parameter characterizes whether the capacity exclusively reserved for ABR has been appropriately planned or not.

The *ABR utilization* of a link l_j characterizes the usage of the overall capacity available for ABR, i.e. the capacity exclusive for ABR and the remaining portion which is used not by CBR/VBR traffic. An ABR utilization of less than one indicates that capacity is wasted on the trunk.

Of course, the SCR is also an important engineering objective for the service provider. From his point of view, the SCR should be chosen in

such a way that the customer notices a benefit by using the ABR service, however, without this decreasing the possible revenue in the system.

8.3.4 Engineering for Stochastic Time-oriented Traffic

The main objective of ABR service engineering assuming stochastic time-oriented traffic is the evaluation of the connection blocking probabilities with regard to the specific user traffic behavior. From the blocking probabilities, additional Quality-of-Service parameters can be derived. To calculate the connection blocking probabilities it is necessary to compute the *state probabilities* of the system. In this section, first, two exact methods to calculate the state probabilities are introduced. However, since their high computational complexity prohibits their application in large scale ABR service engineering, an approximative approach is additionally presented.

a) Product-Form Solution

The first method considered for the performance evaluation of the ABR service model is a *product-form* solution, which is well known in queueing network literature. The approach assumes that the state of the network model is changed when an ABR connection is established or terminated. Hence, the state vector in this model is defined as:

$$\mathbf{n} = (n_1, \ldots, n_N) \tag{8.5}$$

where n_i is the number of connections of class v_i currently active in the network and N is the total number of connection classes. The state transitions in the model are denoted by the successive state:

$$\mathbf{n}_i^+ = (n_1, \ldots, n_{i-1}, n_i + 1, n_{i+1}, \ldots, n_N) \tag{8.6}$$

$$\mathbf{n}_i^- = (n_1, \ldots, n_{i-1}, n_i - 1, n_{i+1}, \ldots, n_N), \tag{8.7}$$

where the state $\mathbf{n}_i^+$ is obtained when a connection of class v_i is established and $\mathbf{n}_i^-$ is the state after a connection of class v_i is released. The set of

all possible states $\mathcal{S}$ is defined by:

$$\mathcal{S} = \{\mathbf{n} \mid n_i \geq 0, \text{ for } i = 1, \ldots, N \text{ and } \sum_{i=1}^{N} n_i MCR_i \delta_i(j) \leq C_j^A, \forall l_j \in \mathcal{L}\}, \tag{8.8}$$

with $\mathcal{L}$ denoting the set of all links in the network, C_j^A the capacity on link l_j exclusively available for ABR, and MCR_i is the minimum cell rate request by connections of class v_i.

According to Kaufman (1981), the local balance equation for the state probabilities of the model is given by:

$$p(\mathbf{n}) = \frac{\lambda_i}{n_i \mu_i} p(\mathbf{n}_i^-) = \frac{a_i}{n_i} p(\mathbf{n}_i^-), \text{ for } \mathbf{n}, \mathbf{n}_i^- \in \mathcal{S}, \tag{8.9}$$

where the parameter λ_i is denoting the arrival rate of connections of class v_i and μ_i^{-1} is the corresponding average connection holding time, cf. Section 8.3.1. The probability $p(\mathbf{n})$ is denoting the probability of the system to be in state $\mathbf{n}$. The recursion of Eqn. (8.9) is solved by initializing $p(\mathbf{0}) = p(0, \ldots, 0) = 1$. The resulting unnormalized state probabilities are:

$$\tilde{p}(\mathbf{n}) = \prod_{i=1}^{N} \frac{a_i^{n_i}}{n_i!}, \forall \mathbf{n} \in \mathcal{S}. \tag{8.10}$$

The state probabilities are obtained from the unnormalized probabilities via the normalization constant $G(\mathcal{S})$:

$$p(\mathbf{n}) = \frac{\tilde{p}(\mathbf{n})}{G(\mathcal{S})} \tag{8.11}$$

with $G(\mathcal{S}) = \sum_{\mathbf{n} \in \mathcal{S}} \tilde{p}(\mathbf{n})$.

At last, the blocking probability B_i for connections of class v_i is given by the sum of the probabilities of the states in which an arriving connection of class v_i is blocked:

$$B_i = \sum_{\mathbf{n} \in \mathcal{S}_i} p(\mathbf{n}). \tag{8.12}$$

The set $\mathcal{S}_i$ of the states in which connections of class v_i are blocked is defined by:

$$\mathcal{S}_i = \{\mathbf{n} \in \mathcal{S} \mid \mathbf{n}_i^+ \notin \mathcal{S}\}. \tag{8.13}$$

Computational complexity of the product-form solution

The time complexity of the product-form approach is determined by the size of the state space $\mathcal{S}$. The cardinality of $\mathcal{S}$ is roughly estimated by, cf. Staehle (1999):

$$\mid \mathcal{S} \mid \leq \prod_{i=1}^{N} \frac{C_{min}^{A}}{MCR_i} \tag{8.14}$$

where $C_{min}^{A} = \min_j \{C_j^A \mid \delta_i(j) = 1\}$ and $\mid \cdot \mid$ denoting the size of set $\mathcal{S}$. Hence, the growth of $\mid \mathcal{S} \mid$ can be assumed as exponentially. Therefore the time complexity of the product-form solution is in, cf. Pinsky and Conway (1992):

$$O\left(\left\lceil \frac{\max\limits_{j}\{C_j^A\}}{\min\limits_{i}\{MCR_i\}} \right\rceil\right). \tag{8.15}$$

The high computational complexity of the product-form solution prohibits its use in the performance analysis of large ATM networks. Only small and medium sized systems can be evaluated with this method. Therefore, a less computational demanding procedure for calculating the blocking probabilities is urgently needed.

b) Kaufman & Roberts Solution

The second method considered for calculating the ABR connection probabilities is a *multi-dimensional* extension of the Kaufman and Roberts method for obtaining the blocking probabilities in multi-service models, cf. Kaufman (1981) and Roberts (1981). The basic idea of Kaufman and Roberts is the reduction of the cardinality of the state space. This method promises to be more applicable as the product-form solution. Hence, in the Kaufman and Roberts method, *macro states* are applied in the model which describe the bandwidth already allocated for the ABR service on a link. This constrasts the product-form approach where the number of connections is used to describe the state of the model. Therefore, a quantization of the allocated bandwidth on every link is introduced, since link capacities and MCRs are not always integer values. In the Kaufman and Roberts approach a macro state $\mathbf{c}^u$ with

respect to the quantization vector $\mathbf{u}$, cf. Eqn.(8.23), is defined as:

$$\mathbf{c}^u = (c_1^u, \ldots, c_M^u), \tag{8.16}$$

where c_j^u is the capacity occupied in link l_j, M is the number of links, and $j = 1, \ldots, M$. Thus, a macro state $\mathbf{c}^u$ in the extended Kaufman and Roberts model is equivalent to a set of states $\mathbf{n}$ in the product-form solution:

$$\mathbf{c}^u \hat{=} \left\{ \mathbf{n} \mid \sum_{i=1}^{N} n_i \cdot b_{ij} \cdot u_j = c_j, \text{ for } j = 1, \ldots, M \right\}. \tag{8.17}$$

The state space of the investigated Kaufman and Roberts model is:

$$\mathcal{Z} = \{\mathbf{c}^u \mid 0 \leq c_j^u \leq C_j^u, \text{ for } j = 1, \ldots, M\}. \tag{8.18}$$

The structure of the state space is an *M-dimensional* hypercube with edge length of C_j^u; the variable M is denoting the number of links in the network. Therefore, this approach is also denoted as the *multi-dimensional Kaufman and Roberts method*.

The blocking probability B_i for connections of class v_i is the sum of the probabilities of the system to be in a state $\mathbf{c}^u$ where the free capacity for ABR on at least one trunk is less than the MCR_i, i.e.:

$$B_i = \sum_{\mathbf{c}^u \in \mathcal{Z}_i} p(\mathbf{c}^u) \tag{8.19}$$

with $\mathcal{Z}_i = \{\mathbf{c}^u \mid \mathbf{c}^u \in \mathcal{Z} \wedge (\mathbf{c}^u + \mathbf{MCR}_i) \notin \mathcal{Z}\}$ denoting the set of the states in which a connection of class v_i is blocked.

The bandwidth quantum variable u_j on a link l_j is the *greatest common divisor (gcd)* of the link capacity and the MCRs of all connection classes passing through l_j:

$$u_j = gcd\left\{ C_j^A \cup \{MCR_i \mid \delta_i(j) = 1\} \right\} \tag{8.20}$$

The capacity C_j^u on a link l_j exclusively available for ABR expressed in capacity units is:

$$C_j^u = \frac{C_j^A}{u_j}, \text{ for } j = 1, \ldots, M\,, \tag{8.21}$$

and the bandwidth requirement b_{ij} of a connection of class v_i on link l_j in capacity units of l_j is denoted as:

$$b_{ij} = \frac{MCR_i \delta_i(j)}{u_j}, \text{ for } , i = 1, \ldots, N; \;\; j = 1, \ldots, M. \tag{8.22}$$

The corresponding vectors for the capacity units u_j and the bandwidth requirement b_{ij} are:

$$\mathbf{u} = (u_1, \ldots, u_M) \tag{8.23}$$
$$\mathbf{b}_i = (b_{i1}, \ldots, b_{iM}), \text{ for } i = 1, \ldots, N. \tag{8.24}$$

The state probability $p(\mathbf{c})$ of a macro state in the multi-dimensional Kaufman and Roberts model is obtained by solving the local balance equations, cf. Staehle (1999):

$$\lambda_i p(\mathbf{c}^u - \mathbf{b}_i) = \bar{n}_i(\mathbf{c}) \mu_i p(\mathbf{c}^u), \text{ for } i = 1, \ldots, N, \tag{8.25}$$

with

$$\bar{n}_i(\mathbf{c}) = E\left[n_i \mid \sum_{k=1}^{N} n_k b_{kj} = c_j^u, \text{ for } j = 1, \ldots, M \right]$$

as the average number of active connections of class v_i under the condition that the sum of the allocated bandwidth units in each link l_j is exactly c_j^u. The recursion to determine the state probabilities is obtained by multiplying Eqn.(8.25) with $\sum_{j=1}^{M} b_{ij}/\mu_i$ and then summing over all connection classes:

$$\begin{aligned}
&\sum_{i=1}^{N} \sum_{j=1}^{M} b_{ij}\, a_i p(\mathbf{c}^u - \mathbf{b}_i) = \left(\sum_{i=1}^{N} \sum_{j=1}^{M} b_{ij} \bar{n}_i(\mathbf{c}^u) \right) p(\mathbf{c}^u) \\
&= E\left[\sum_{i=1}^{N} \sum_{j=1}^{M} b_{ij} n_i \mid \sum_{i=1}^{N} n_k b_{kj} = c_j^u, \text{ for } j = 1, \ldots M \right] p(\mathbf{c}^u) \\
&= \left(\sum_{j=1}^{M} c_j^u \right) p(\mathbf{c}^u).
\end{aligned} \tag{8.26}$$

The resulting closed formula to calculate the unnormalized state probabilities is given by:

$$\tilde{p}(\mathbf{c}^u) = \begin{cases} 1 & : \quad \mathbf{c}^u = \mathbf{0} \\ \sum_{i=1}^{N} \frac{\sum_{j=1}^{M} b_{ij}}{\sum_{j=1}^{M} c_j^u} a_i \tilde{p}(\mathbf{c}^u - \mathbf{b}_i) & : \quad \mathbf{0} < \mathbf{c}^u \leq \mathbf{C}^u \\ 0 & : \quad \mathbf{c}^u < \mathbf{0} \end{cases} . \tag{8.27}$$

The vector operator "$<$" is defined as: $\mathbf{c} < \mathbf{c}' \Leftrightarrow \exists i : c_i < c_i'$. The normalized state probabilities are:

$$p(\mathbf{c}^u) = \frac{\tilde{p}(\mathbf{c}^u)}{\sum_{\mathbf{c}^u \in \mathcal{Z}} \tilde{p}(\mathbf{c}^u)}. \tag{8.28}$$

Finally, the blocking probability for a connection of class v_i is obtained by Eqn.(8.19).

Computational Complexity of the Kaufman & Roberts Approach

In Kaufman and Roberts approach, the cardinality of the state space $\mathcal{Z}$ is directly given by:

$$\mid \mathcal{Z} \mid = \prod_{j=1}^{M} C_j^u \tag{8.29}$$

To determine the blocking probabilities, the normalization constant has to be calculated, and therefore all state probabilities are required. The probability of each state depends on N other states, and thus the computing time is within a complexity of:

$$O\left(N \prod_{j=1}^{M} C_j^u \right), \tag{8.30}$$

on the basis that enough storage is available, cf. Staehle (1999)

As well as in the case of the product-form approach, the disadvantage of the Kaufman and Roberts method is the significantly high computational complexity and the large cardinality of the state space. Even the reduction of the states does not ease the complexity. Thus, the Kaufman and Roberts approach is also prohibited from application in large scale ABR service engineering.

c) Knapsack Approximation

The third approach investigated for computing the blocking probabilities is the *Knapsack* approximation. The basic idea of the approximation is to consider each link in an isolated but exact manner and then to combine the results of the links for computing the approximated probabilities for the network. The combination of the results is done by a *reduced load* approximation, cf. Chung and Ross (1993).

Reduced Load Approximation

To motivate the reduced load approximation, first a single trunk is assumed. If a connection of class v_i is blocked on link l_j with probability B_{ij} then the load on the link, i.e. the mean number of active connections in l_j, will be reduced to:

$$\alpha'_{ij} = a_i(1 - B_{ij}) \tag{8.31}$$

where a_i is the load offered by connection class v_i. Now, the complete path of connection class v_i is considered. Supposing that the blocking events occur independently from link to link, then the reduced load α_{ij} from connection class v_i on link l_j is:

$$\alpha_{ij} = a_i \prod_{k=1,k\neq j}^{M} (1 - B_{ik}\delta_i(k))\,. \tag{8.32}$$

Exploiting the link independence assumption, the link blocking probability B_{ij} is defined by a fixed point equation:

$$B_{ij} = \beta(C_j^u;\ a_i \prod_{k=1,k\neq j}^{M} (1 - B_{ik}\delta_i(j)), \tag{8.33}$$

where $\beta(\cdot;\ \cdot)$ is the function to calculate the single link blocking probability. Hence, the blocking probability B_i of connection class v_i for the whole network is approximated by, cf. Chung and Ross (1993):

$$B_i = 1 - \prod_{j=1}^{M} (1 - B_{ij}\delta_i(j)). \tag{8.34}$$

In the Knapsack approximation, the single link blocking probability in Eqn.(8.33) is calculated using the method of Kaufman and Roberts. The single link system consists of a trunk with capacity C_j^u exclusively available for ABR and the set $\{v_i \in \mathcal{V} \mid \delta_i(j) = 1\}$ of connection classes using link l_j with the corresponding bandwidth requirement b_{ij} and the offered load a_i. This system is denoted as the *stochastic Knapsack* due to its resemblance to the Knapsack model in combinatorial optimization. The function for computing the link blocking probability is denoted in the Knapsack approximation by K_{ij}. The blocking probability B_{ij} for connection class v_i in the isolated link l_j is exactly given by:

$$B_{ij} = K_{ij}[C_j^u;\ \alpha_{ij}, \delta_i(j) = 1] = 1 - \frac{\sum_{n=0}^{C_j^u - b_{ij}} p_j(n)}{\sum_{n=0}^{C_j^u} p_j(n)}, \tag{8.35}$$

where $p_j(0) = 1$ and

$$p_j(n) = \frac{1}{n} \sum_{i=1}^{N} b_{ij}\,\alpha_{ij}\,p_j(n - b_{ij}),\ \text{for } n = 1, \ldots, C_j^u\,. \tag{8.36}$$

Finally, the blocking probability for connections of class v_i for the whole network is given by Eqn.(8.34).

An iterative algorithm for solving the fixed point equation (8.35) and obtaining the connection class blocking probabilities is depicted in Algorithm 8.1. The parameter ϵ specifies the convergence criterion of the reduced load approximation. The iteration stops if the relative error on the reduced load approximation becomes sufficiently small, i.e. the deviation for all load values between the iterations is less than ϵ.

Algorithm 8.1 (Knapsack approx. algorithm for connection blocking probabilities)

variables:

a_i	*offered load by connection class i*
l_j	*link with index j*
C_j^u	*capacity on link l_j exclusively available for ABR in bandwidth units*
ϵ	*convergence criterion*

algorithm:

```
1  proc approximate_blocking()
2    ≡
3   begin
4     α_ij ← a_i, ∀i = 1,...,N ∧ ∀j = 1,...,M;
5     for j = 1 to M do
6         determine p(n) for n = 0,...,C_j^u for link l_j using Eqn.(8.36);
7         compute B_ij according to Eqn.(8.35);
8     end
9     calculate α_ij^new according to Eqn.(8.33);
10    if (| α_ij^new − α_ij |≥ εα_ij) for at least one i, j with δ_i(j) = 1 then
11      α_ij = α_ij^new, ∀i = 1,...,N ∧ ∀j = 1,...,M;
12      goto 5;
13    fi
14    compute B_i for i = 1,...,N;  according to Eqn.(8.34);
15  end
```

Algorithm 8.1: Knapsack approximation algorithm for the connection blocking probabilities

Computational Complexity of the Knapsack Approximation

The computational complexity of the Knapsack approximation is significantly below the complexity of the exact methods. Since the Knapsack approximation uses the Kaufman and Roberts method for obtaining the single link blocking probabilities, the complexity is reduced to the hardness of solving a single link system. The calculation of the probabilities has to be performed for every trunk and therefore the time complexity grows in a linear fashion with the number of links in the network. this differs from the exact methods where the hardness mainly depends on the number of connections classes. The Knapsack approximation states a feasible approach for obtaining the connection blocking probabilities in large scale ATM networks.

d) QoS Values

The major Quality-of-Service (QoS) parameters for ABR service planning assuming time-oriented traffic can be derived from the connection blocking probabilities. The average number of active connections of class v_i is directly given by Little's theorem:

$$\bar{n}_i = a_i(1 - B_i) \tag{8.37}$$

As outlined above, the term *utilization* is ambiguous in the context of ABR service engineering. Therefore, two utilization parameters have been introduced, cf. Section 8.3.3: *a)* ABR utilization and *b)* MCR utilization.

The average MCR utilization ρ_j^{MCR} for link l_j is determined by:

$$\rho_j^{MCR} = \frac{\sum_{i=1}^{N} \bar{n}_i MCR_i \delta_i(j)}{C_j^A}. \tag{8.38}$$

To compute the average ABR utilization ρ_j^{ABR} on link l_j it is required to compute the mean explicit rate $\overline{ER}_i$, also denoted as the SCR, allocated to the source of class v_i. This rate is the sum of the constant MCR and the additional bandwidth allocated to the source by the maxmin fair share algorithm, cf. Section. 8.3.2.

In case of the product from solution, the mean explicit rate $\overline{ER}_i$ for connections of class v_i can directly be derived from the state probabilities:

$$\overline{ER}_i = \frac{\sum_{\mathbf{n}\in\mathcal{S}} p(\mathbf{n}) ER_i(\mathbf{n})}{p(n_i > 0)} = \frac{\sum_{\mathbf{n}\in\mathcal{S}} p(\mathbf{n}) ER_i(\mathbf{n})}{\sum_{\mathbf{n}\in\mathcal{S}, n_i>0} p(\mathbf{n})}, \tag{8.39}$$

where $ER_i(\mathbf{n})$ is the result of the maxmin fair share algorithm in state $\mathbf{n}$ a connection of class v_i and n_i is the number of active connections of class v_i.

In the Kaufman and Roberts approach, however, the state probability $p(\mathbf{n})$ is not explicitly known. Hence, it is not possible to compute the mean explicit rate $\overline{ER}_i$ exactly. Only an approximation can be provided using the mean number of active connections:

$$\overline{ER}_i \approx ER_i(\bar{\mathbf{n}}), \tag{8.40}$$

where $\bar{\mathbf{n}} = (\bar{n}_1, \ldots, \bar{n}_N)$. In detail the approximation is based on the following assumptions:

$$\begin{aligned} ER_i(\bar{\mathbf{n}}) &= ER_i\left(\sum_{\mathbf{n}\in\mathcal{S}} p(\mathbf{n})n_1, \ldots, \sum_{\mathbf{n}\in\mathcal{S}} p(\mathbf{n})n_N\right) \\ &\overset{(1)}{\approx} \left(\frac{\sum\limits_{\mathbf{n}\in\mathcal{S}} p(\mathbf{n})n_1}{p(n_1>0)}, \ldots, \frac{\sum\limits_{\mathbf{n}\in\mathcal{S}} p(\mathbf{n})n_N}{p(n_N>0)}\right) \\ &\overset{(2)}{\approx} \frac{\sum\limits_{\mathbf{n}\in\mathcal{S}, n_i>0} p(\mathbf{n})ER_i(\mathbf{n})}{\sum\limits_{\mathbf{n}\in\mathcal{S}, n_i>0} p(\mathbf{n})} = \overline{ER}_i\,. \end{aligned}$$

The first assumption (1) can be made since the probability, that at least one connection is active, approaches one for heavy traffic. The second assumption (2) is based on a linearization.

The average ABR utilization ρ_j^{ABR} of link l_j is approximated by:

$$\rho_j^{ABR} = \frac{\sum\limits_{i=1}^{N} \overline{ER}_i \bar{n}_i \delta_i(j)}{C_j^A} \approx \frac{\sum\limits_{i=1}^{N} \overline{ER}_i(\bar{\mathbf{n}})\delta_i(j)}{C_j^A}. \tag{8.41}$$

Finally, the bandwidth benefit of a customer using the ABR service is characterized by the *relative additional rate* $\overline{ER}^+$ allocated to the source:

$$\overline{ER}_i^+ = \frac{\overline{ER}_i - MCR_i}{MCR_i} \cdot 100. \tag{8.42}$$

This parameter allows a rough comparison of ABR service and CBR service. Its value describes the ratio of additional bandwidth allocated in average to an ABR connection and its MCR. From the view point of CBR service, the MCR value can be regarded as the constant rate of a CBR connection. In this context, the value of $\overline{ER}^+$ represents the benefit of requesting an ABR connection instead of using a CBR service with PCR equal to the MCR of the ABR connection.

8.3.5 ABR Planning for Volume-oriented Traffic

ABR service engineering assuming stochastic volume-oriented traffic differs substantially from the service engineering in the time-oriented case. The connection holding time depends on the amount of connections currently active in the network, i.e. the connection holding time is *state dependent*. Therefore, the above introduced methods for computing the connection blocking probabilities can not be applied in this case.

In this section, a method for obtaining the connection blocking probabilities and other QoS parameters will be outlined. However, the computational complexity of the approach is very large. Nevertheless, a presentation of the method in this chapter might initiate some additional work towards a procedure with less computational requirements.

State Probabilities

As in the product-form approach, the state of the model for ABR planning with volume-oriented traffic is denoted by $\mathbf{n}$ and characterizes the number of active connections of class v_i:

$$\mathbf{n} = (n_1, \ldots, n_N). \tag{8.43}$$

Hence, the state space, cf. Eqn.(8.8), the state transitions for establishing and terminating an ABR connection, cf. Eqn.(8.6) and Eqn.(8.7), and the event of blocking a connection of class v_i, cf. Eqn.(8.13), are all defined in the same way as in the product-form approach which is applied in the time-oriented model.

The difference between the two models is the state transition rate for the case of terminating an ABR connection. As described above, the transition rate for connection release depends on the current load of the network. Hence, the transition rates are defined as:

$$q(\mathbf{n}, \mathbf{n}_i^+) = \begin{cases} \lambda_i & : \quad \mathbf{n}_i^+ \in \mathcal{S} \\ 0 & : \quad \text{else} \end{cases}, \text{ for } \mathbf{n} \in \mathcal{S}, \tag{8.44}$$

$$q(\mathbf{n}, \mathbf{n}_i^-) = \begin{cases} n_i \frac{\nu_i^{-1}}{ER_i(\mathbf{n})} & : \quad \mathbf{n}_i^- \in \mathcal{S} \\ 0 & : \quad \text{else} \end{cases}, \text{ for } \mathbf{n} \in \mathcal{S}. \tag{8.45}$$

The explicit rate $ER_i(\mathbf{n})$ is assigned to the source of class v_i by the maxmin fair share algorithm, cf. Section 8.3.2. Consequently, the connection holding time for a call of class v_i can be approximated by the

assigned rate $ER_i(\mathbf{n})$ divided by the the average data volume ν_i^{-1} of calls in this class.

Due to the dependence of the transition rate on the system state, it is not possible to apply the product-form approach nor the Knapsack approximation. The remaining method to obtain the state probabilities is to solve the system of equilibrium balance equations for each state $\mathbf{n} \in \mathcal{S}$:

$$\left[\sum_{i=1}^{N} \theta_i^+(\mathbf{n})\lambda_i + \sum_{i=1}^{N} \theta_i^-(\mathbf{n}) n_i \frac{\nu^{-1}}{ER_i(\mathbf{n})}\right] p(\mathbf{n}) = \sum_{i=1}^{N} \theta_i^-(\mathbf{n})\lambda_i p(\mathbf{n}_i^-) + \sum_{i=1}^{N} \theta_i^+(\mathbf{n}) n_i \frac{\nu^{-1}}{ER_i(\mathbf{n}_i^+)} p(\mathbf{n}_i^+)\,, \tag{8.46}$$

$$\theta_i^-(\mathbf{n}) = \begin{cases} 1 & : \quad \mathbf{n}_i^- \in \mathcal{S} \\ 0 & : \quad \text{else} \end{cases} \quad \text{and} \tag{8.47}$$

$$\theta_i^+(\mathbf{n}) = \begin{cases} 1 & : \quad \mathbf{n}_i^+ \in \mathcal{S} \\ 0 & : \quad \text{else} \end{cases} \quad . \tag{8.48}$$

By obeying the normalization constraint $\sum_{\mathbf{n}\in\mathcal{S}} p(\mathbf{n}) = 1$, a unique solution for the state probabilities is obtained.

The computational complexity of this method for obtaining the state probabilities is in the same order as the complexity of the product-form solution, cf. Eqn.(8.14) and Eqn.(8.15). Hence, also these methods can not be applied in large scale ABR service engineering.

QoS Values

The advantage of solving the system of balance equations is the exact knowledge of the state probabilities. Therefore, an accurate calculation of the Quality-of-Service values can be carried out. The blocking probability B_i for connections of class v_i is:

$$B_i = \sum_{\mathbf{n}\in\mathcal{S}, \mathbf{n}_i^+ \notin \mathcal{S}} p(\mathbf{n}) = 1\,, \tag{8.49}$$

and the average transmission rate $\overline{ER}_i$ for connection of class v_i is given by:

$$\overline{ER}_i = \sum_{\mathbf{n}\in\mathcal{S}} \frac{ER_i(\mathbf{n}) \cdot p(\mathbf{n})}{\sum\limits_{\mathbf{n}\in\mathcal{S}, n_i>0} p(\mathbf{n})}, \tag{8.50}$$

where $ER_i(\mathbf{n})$ is computed by the maxmin fair share algorithm.

Finally, the average time required to transmit the data, i.e. the average connection holding time for calls of class v_i is of great interest. This parameter is given by:

$$E[T_i] = \sum_{\mathbf{n}\in\mathcal{S}, n_i>0} p(\mathbf{n}) \frac{\nu_i^{-1}}{ER(\mathbf{n})}. \tag{8.51}$$

8.4 Case Study

The remainder of this chapter is devoted to a case study on the applicability of the Knapsack approximation for the evaluation of ABR service in large scale ATM networks. For the case study, a network structure is assumed which is close to the network of a major ATM service provider in the USA. Additionally, at the end of this section, some aspects in ABR service planning are discussed by using this example.

8.4.1 Backbone Network

The example ATM network considered in the case study is depicted in Figure 8.6. The network comprises eleven core edge nodes. The capacity printed next to a link is denoting the overall capacity on the link. On each link, one fifth of the total trunk capacity is exclusively reserved for ABR traffic. The background traffic is supposed to utilize its part completely. Hence, no additional capacity is available to be allocated by the fair share algorithm.

The traffic scenario investigated in the case study assumes that each of the core edge nodes is connected to another core edge node by at least one connection class. However, not every possible path between the nodes is considered. For the MCRs of the connections classes, five different levels are allowed: $0.2Mbps$, $0.4Mbps$, $1Mbps$, $2Mbps$, and $4Mbps$.

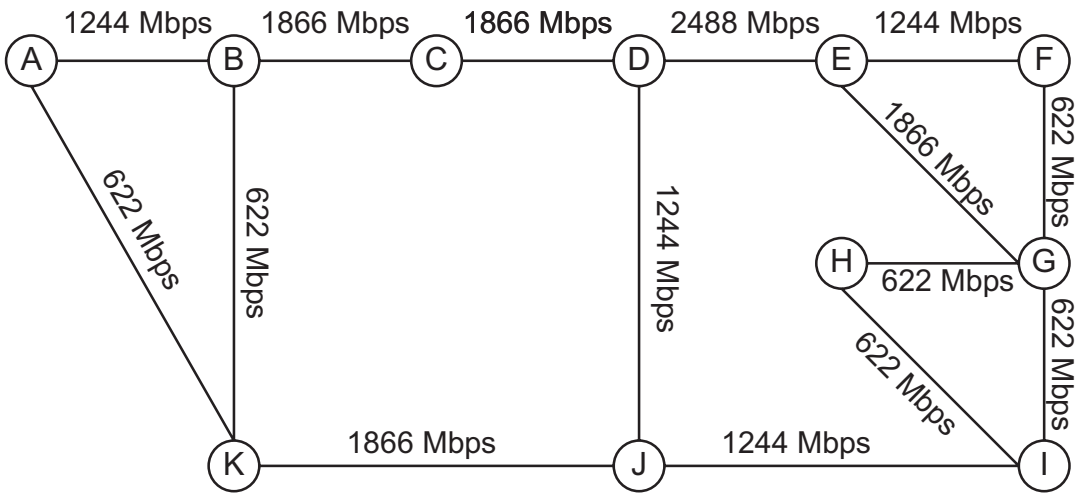


Figure 8.6: *Example network of the case study*

The offered load was arbitrarily chosen. All in all, a total number of 107 connection classes is considered, cf. Table 8.1 and Table 8.2. The classes are numbered by their MCRs. Classes with small indices indicate a low MCR, e.g. class v_1 has $MCR_1 = 0.2Mbps$, and classes with high numbers represent large MCRs, e.g. v_{107} has $MCR_{107} = 4Mbps$. The traffic scenario described above is referred to as *benchmark traffic*.

8.4.2 Performance Evaluation for Time-oriented Traffic

The example network of Section 8.4.1 is evaluated by four major QoS parameters: *a)* the connection blocking probability, *b)* the sustainable cell rate, *c)* the bandwidth benefit, and *d)* the MCR utilization.

Connection Blocking Probability

The blocking probabilities for all connection classes of the benchmark traffic are depicted in Figure 8.7. The approximated probabilities are indicated by bars. These values are compared with results obtained from a discrete event simulation of the ATM network. In this simulation, the events were defined as the arrival or as the release of an ABR connection request. The blocking probabilities obtained by the simulation are indicated by means of their 95% confidence interval. The probabilities

path	class index	MCR [Mbps]	offered load	class index	MCR [Mbps]	offered load	class index	MCR [Mbps]	offered load
A-B	31	0.4	20	74	2	25	103	4	20
A-B-C	19	0.4	20	64	2	5			
A-B-C-D	12	0.4	20	56	2	5			
A-B-C-D-E	28	0.4	20	81	2	5			
A-B-C-D-E-F	31	0.4	20	86	2	5			
A-B-C-D-E-G	37	0.4	20	72	1	5			
A-K-J-I-H	32	0.4	20	47	1	5			
A-K-J-I	45	1	6						
A-K-J	27	0.4	20	73	2	5			
A-K	30	0.4	20	70	2	25			
B-C	24	0.4	20	105	4	45			
B-C-D	16	0.4	20	61	2	45			
B-C-D-E	40	0.4	20	83	2	5			
B-C-D-E-F	33	0.4	20	93	2	4			
B-C-D-E-G	29	0.4	20	86	2	5			
B-C-D-E-G-H	9	0.4	20	49	1	4			
B-K-J-I	97	2	9						
B-K-J	36	0.4	20	78	2	5			
B-K	6	0.4	20	90	2	4	102	4	20
B-A-K	55	2	20						
C-D	1	0.2	20	53	2	40			
C-D-E	21	0.4	20	66	2	5			
C-D-E-F	26	0.4	80	70	2	5			
C-D-E-G	22	0.4	20	67	2	5			
C-D-E-G-H	25	0.4	20	69	2	5			
C-D-E-G-I	44	1	6						
C-D-J	20	0.4	20	65	2	5			
C-B-K	23	0.4	20	68	2	5			

Table 8.1: *Connection classes of the benchmark traffic (part a)*

path	class index	MCR [Mbps]	offered load	class index	MCR [Mbps]	offered load	class index	MCR [Mbps]	offered load
D-E	14	0.4	20	58	2	45			
D-E-F	18	0.4	20	63	2	5			
D-E-G	50	1	20	59	2	35			
D-E-G-H	17	0.4	20	62	2	5			
D-E-G-I	52	2	9						
D-J	13	0.4	20	57	2	45			
D-J-K	15	0.4	20	60	2	35			
E-F	43	1	20	85	2	35			
E-G	51	1	20	107	4	35			
E-F-G	54	2	20						
E-G-H	41	0.4	20	84	2	5			
E-G-I	89	2	9						
E-D-J	34	0.4	20	76	2	5			
E-D-K	39	0.4	20	82	2	5			
F-G	5	0.4	20	88	2	5	100	4	20
F-G-H	11	0.4	20	94	2	4			
F-G-I	42	1	9						
F-E-D-J	38	0.4	20	81	2	5			
F-E-D-J-K	8	0.4	20	92	2	4			
G-H	4	0.4	20	87	2	5	104	4	10
G-I	95	2	39						
G-E-D-J	35	0.4	20	77	2	5			
G-E-D-J-K	2	0.4	20	48	1	5			
H-I	98	2	9	101	4	20			
H-I-J	37	0.4	20	79	2	5			
H-I-J-K	7	0.4	20	91	2	4			
I-J	80	2	54						
I-J-K	99	2	54						
J-K	68	2	20	106	4	40			

Table 8.2: *Connection classes of the benchmark traffic (part b)*

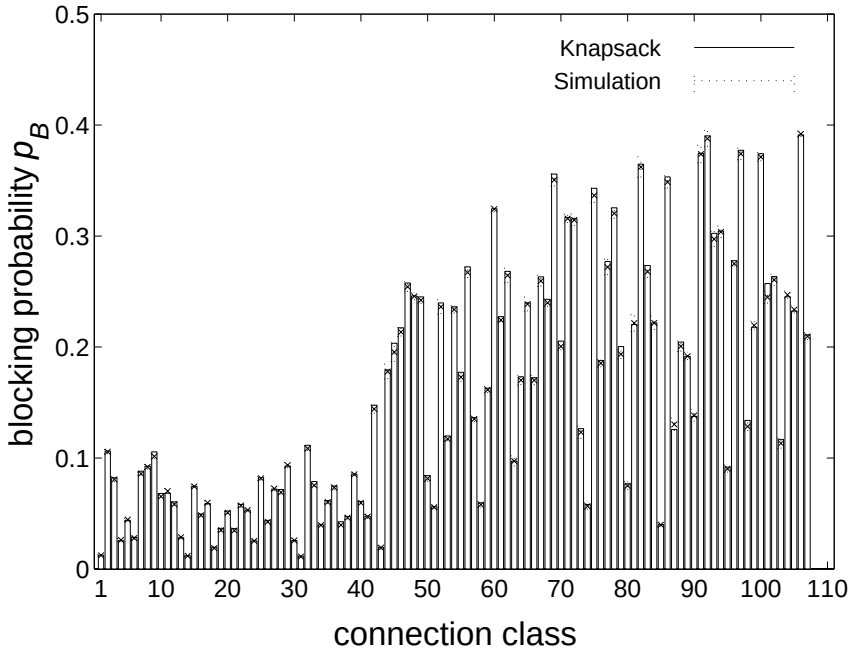


Figure 8.7: *Blocking probabilities for the connection classes of the benchmark traffic*

computed by the approximation are surprisingly close to those obtained by simulation. Furthermore, a close look reveals that the approximation is overestimating the blocking probabilities, however only by a small amount. This characteristic is inherent to the assumption of the Knapsack approximation of independent link blocking, cf. Section 8.3.4.

For selected connections the approximated blocking probabilities are investigated for various load scenarios, cf. Figure 8.8. Again, the values of the approximation are compared with the probabilities obtained by the simulation. In context of the case study, the benchmark traffic is supposed to be a constitute a load of 100 % in the network. A decreased network load assumes that the offered load is equally reduced for all connection classes. A reduced network load of 75% implies that the offered load of every class is only 75% of the offered load of the benchmark traffic. The results for connection classes v_1, v_9, v_{76}, and v_{84} are depicted in Figure 8.8. These classes have been chosen with regard to the amount of their blocking probability and their deviation from the simulation results. The approximation of the probabilities of classes v_1 and v_{84} is very accurate for the benchmark traffic whereas the deviation of classes v_9 and v_{76} is quite strong. Classes with high indices have

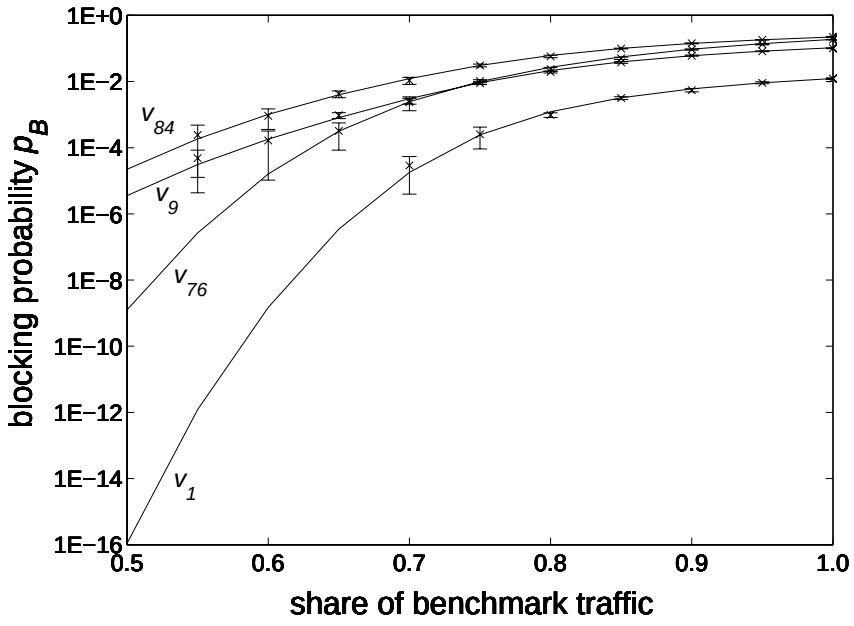


Figure 8.8: *Blocking probabilities for selected connection classes at varying network load*

larger MCRs and consequently an increased blocking probability. As the network load reduces, the blocking probabilities are also decreasing. A difference between the approximated and simulated results can hardly be recognized.

Sustainable Cell Rate

The second QoS parameter investigated in this case study is the sustainable cell rate of an ABR source, cf. Section 8.3.4. Its value is approximated by the application of Eqn.(8.40). The results for the SCR obtained by the approximation and by the simulation are depicted in Figure 8.9. The results show that the SCRs are sufficiently close approximated. However, some SCRs values are underestimated. This behavior can be observed when a connection class only passes through a single link. Such a connection cannot be bottlenecked by other links. Thus, the connection always receives the optimal share of the trunk.

Figure 8.10 depicts the SCRs of selected connection classes for decreasing network load; the benchmark traffic is assumed to be a load of 100%. The lines are denoting the approximated values and are indicating the crosses the results obtained by simulation. As expected, the SCRs

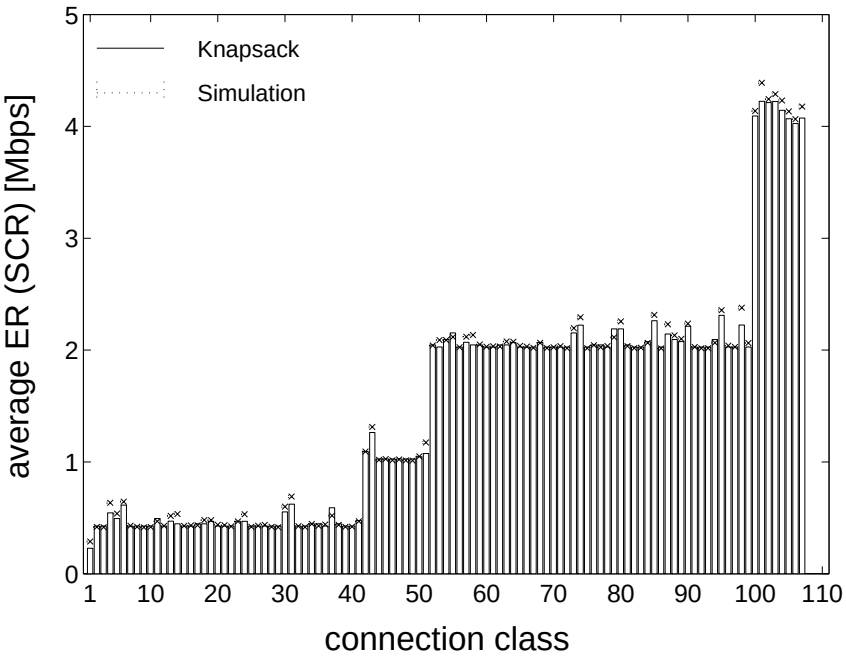


Figure 8.9: *Approximated and simulated SCRs of the benchmark traffic*

increase with reduced load. The deviation between the approximation and the simulation increases slightly for connections classes v_4 and v_{98} which only use a single link. In contrast to this, the accuracy of the approximation for the multi-link connection classes v_2 and v_{64} remains same for different load values.

Bandwidth Benefit

A basic objective of ABR service engineering is to provide the customer with as much bandwidth as possible; of course, under the constraint of supporting as many ABR connections as possible at the same time. A better indication of the bandwidth benefit than the SCR, however, is the relative additional rate $\overline{ER}^{+}$. The results for the example network at 75% load are shown in Figure 8.11. The relative additional rate varies between 2% and 170%. Due to their small MCRs, the connection classes v_1 to v_{41} receive a proportional stronger benefit than the classes with high MCR, e.g. the classes v_{52} to v_{107}.

From the provider point of view, the customer should get some but not too much bandwidth benefit. Therefore, service planning should be

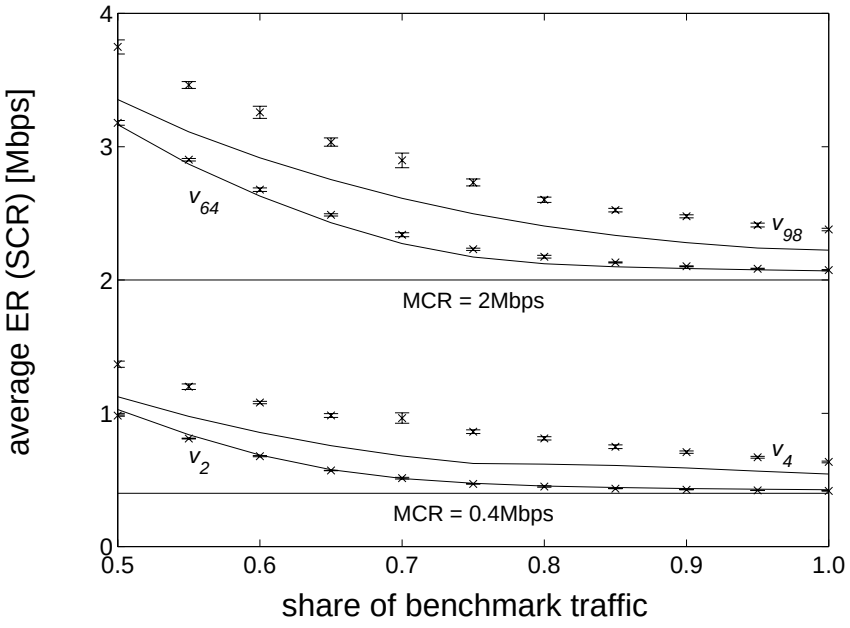


Figure 8.10: *Approximated and simulated SCRs for varying traffic*

carried out in such a way that the benefit should be kept in a predefined range. In Figure 8.11 two range delimiters are indicated: a low boundary at an additional source rate of 5% and an upper a limit at 20%. The majority of the relative additional rates are located within this range.

MCR Utilization

The MCR utilization for the links at the 75% traffic scenario is shown in Figure 8.12. The utilization lies within a reasonable range of 75% to 95%. If the MCR utilization would be too large, then the connection blocking probabilities would have been increased. If the utilization is too low, the links are not efficiently used.

8.4.3 ABR Network Dimensioning

The methods for ABR service engineering introduced so far, are well suited for the evaluation of the QoS parameters of a given network structure and a corresponding traffic scenario. In practical telecommunication network design, however, the initial system structure is neither defined nor it is dimensioned. This is the core task of the network design engineer.

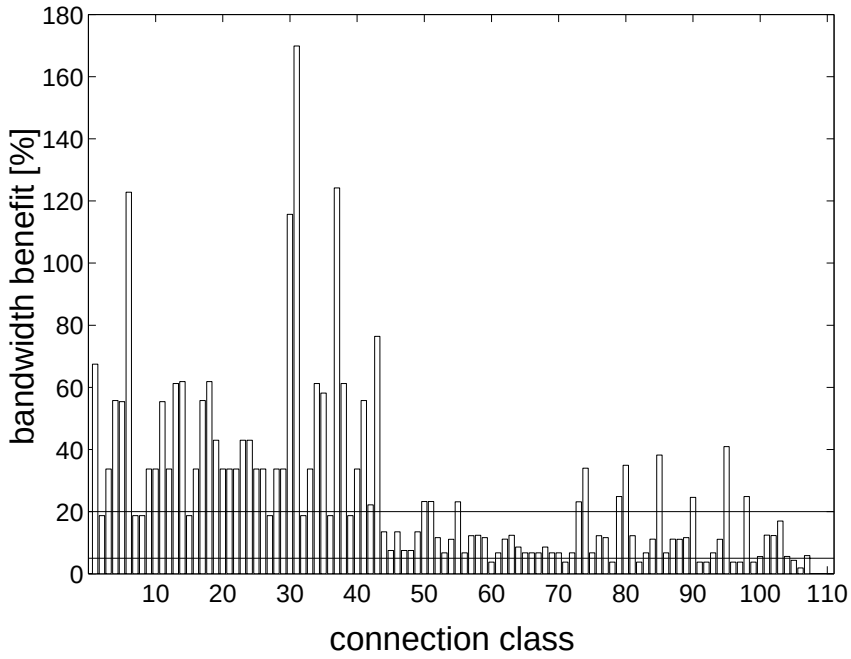


Figure 8.11: *Bandwidth benefit for 75% of the benchmark traffic*

He has to specify the network topology and the link capacities according to the expected traffic scenario and the specific QoS parameters.

The definition of the network topology is a task of very high computational complexity. In fact it is an NP-hard optimization problem. Even the task of dimensioning the trunks is far from being simple. In this section an algorithm for dimensioning the trunks in an ABR service network is proposed. For a given network topology, traffic scenario and connection blocking probability, the method is capable to determine the capacity on a link which has to be exclusively reserved for ABR traffic.

The basic idea of the algorithm is to initialize the link capacity with too much bandwidth and then to reduce the capacity while maintaining the required blocking probabilities. The procedure is depicted in Algorithm 8.2. A detailed description is provided in Staehle (1999). The algorithm comprises two stages. In the first stage the link capacities are overestimated by regarding the single link blocking probabilities. The single blocking probabilities for each connection class B_{ij} is obtained by the method of Kaufman and Roberts, cf. Eqn.(8.35). The single link probabilities have to obey the condition that they have to be less than the required connection blocking probability $B_i^{R'}$ for connection class v_i.

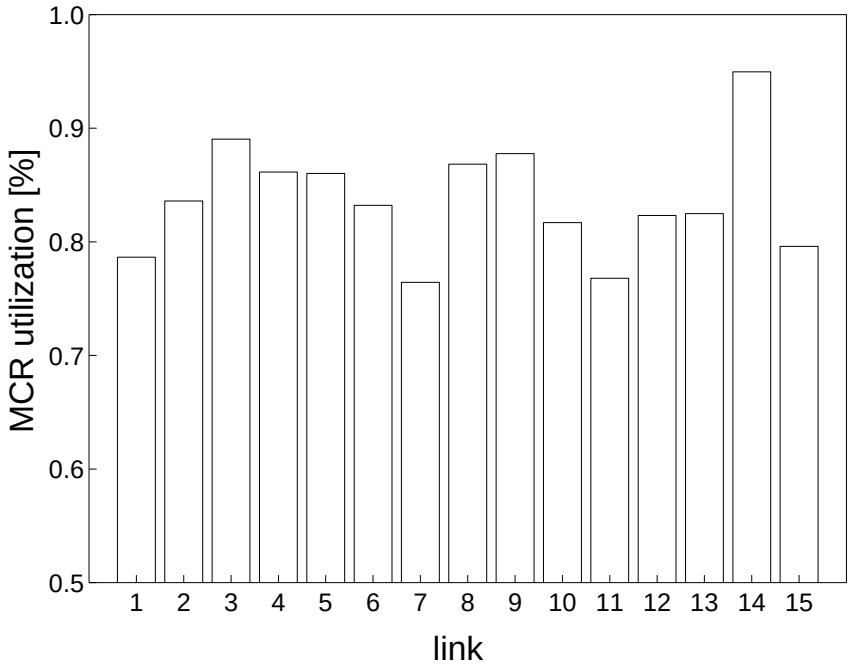


Figure 8.12: *MCR utilization at 75% load of the benchmark traffic*

This value of $B_i^{R'}$ is assumed to be equal for every link. The value of $B_i^{R'}$ is approximated by exploiting independence blocking assumption, cf. Eqn.(8.34):

$$B_i^{R'} = 1 - \sqrt[l(i)]{1 - B_i^R}, \tag{8.52}$$

where B_i^R is the required blocking probability of connection class v_i and $l(i)$ is the number of links used by connections of class v_i. As a result of the first stage, provisional link capacities are obtained.

In the second stage the provisional link capacities are decreased while keeping the connection blocking probabilities B_i lower than the desired level B_i^R. The values of B_i are now computed by the Knapsack approximation. The link capacities are decreased in such an order that the link with the smallest utilization is chosen first. The capacity decrease is stopped as soon as the connection blocking probability does not meet the required probability anymore.

The result of the dimensioning algorithm is depicted in Table 8.4. For the validation of the algorithm, the network topology of the example network (without the link capacities) and the benchmark traffic of

Algorithm 8.2 (Dimensioning of link capacities)

variables:

$\mathcal{L}$	*set of all links in the network*
$\mathcal{K}$	*set of dimensioned links*
C_j	*capacity on link l_j reserved exclusively available for ABR*
B_i^R	*Blocking probability required for connection class v_i*

algorithm:

```
 1  proc approximate_blocking() ≡
 2    begin
 3      B_i^R' ← 1 − ⁽ˡ⁽ⁱ⁾⁾√(1 − B_i^R) for all v_i;
 4      for each link l_j ∈ L do
 5          C̃_j ← Σ_{i=1}^{N} MCR_i · a_i;
 6          do
 7              C̃_j ← C̃_j + 1;
 8              calculate B_ij by Eqn.(8.35); /*using Kaufman and Roberts*/
 9          while(∃i : B_ij > B_i^R')
10      end
11      calculate B_i for all v_i; /* using the Knapsack approx. */
12      repeat
13        calculate ρ_j^MCR for all l_j ∈ L by Eqn.(8.38);
14        ρ_j*^MCR ← min_{l_j ∈ L} {ρ_j^MCR}
15        C̃_j* ← C̃_j* − 1;
16        calculate B_i for all v_i; /* using the Knapsack approx. */
17        if (∃i : B_i > B_i^R) then
18          C̃_j* ← C̃_j* + 1;
19          K ← K + l_j;
20          if (K = L) then goto step 26
21                     else ρ_j*^MCR ← min_{j ∈ C} {ρ_j^MCR};
22                          goto step 12
23          end
24        end
25      forever
26      C_j ← C̃_j for all l_j ∈ L;
27    end
```

Algorithm 8.2: Approximation algorithm for the dimensioning the link capacities

MCR[*Mbps*]	B^R[%]	MCR[*Mbps*]	B^R[%]
0.2	1	2.0	5
0.4	1	4.0	10
1.0	3		

Table 8.3: *Required connection blocking probabilities*

Link	capacity [*Mbps*]	Link	capacity [*Mbps*]	Link	Capacity [*Mbps*]
A-B	290	D-E	600	G-I	151
B-K	169	E-F	289	H-I	176
B-C	499	E-G	478	I-J	320
C-D	473	F-G	880	J-K	563
D-J	317	G-H	170	K-A	158

Table 8.4: *Link capacities obtained by the dimension algorithm*

Section 8.4.1 was considered. The required connection probabilities for connection classes were related to the MCRs, cf. Table 8.3. It shows the calculated link capacities. The cross validation of the result of the dimensioning algorithm is obtained by examining the connection blocking probabilities. The blocking probabilities are depicted in Figure 8.13. The dashed line is the Quality-of-Service limit which has to be met by the blocking probabilities. All blocking probabilities obey these requirements. However, some of the connection classes are very close to the limit. A close examination reveals that the capacity on at least one link in the path of these connection classes cannot be decreased any more without offending the single link blocking probability requirement. This characteristic indicates that the quality of the dimensioning is reasonable good.

8.5 Concluding Remarks

Due to the explicit devotion of a significant fraction of the ATM trunk to ABR, the proposed model for ABR service engineering differs substantially from the view that ABR can be only used for exploiting the excess bandwidth. The presented ABR service engineering model states an efficient approach for large scale ABR service provision in ATM networks

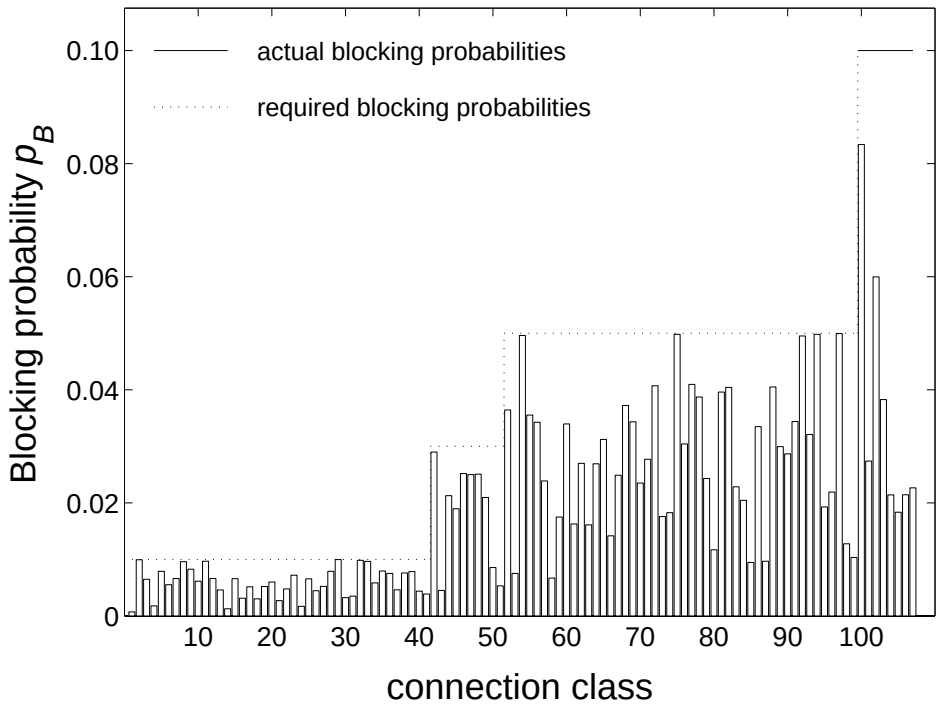


Figure 8.13: *Blocking probabilities for the dimensioned network*

where ABR traffic is, from the commercial view point, as important as CBR/VBR traffic. Hence, the contradiction between the nature of the ABR service and ABR planning is resolved.

For ABR service engineering in large ATM networks, three procedures for the performance evaluation of time-oriented ABR service have been investigated: two exact methods and an approximation procedure. Additionally, an evaluation method for volume-oriented traffic has been outlined. Besides the connection blocking probabilities, additional QoS parameters can easily be derived through the application of these methods.

Finally, the capability of the approximation method for planning the ABR service in large scale ATM network was investigated in a case study. The results of the approximation for time-oriented traffic are promising. They have been proved to be sufficiently accurate. Hence, the proposed approximation for connection blocking probabilities facilitates ABR service engineering in large scale ATM networks. In this way, the network operators would be able to design a reliable ABR service and thus are capable to provide low-priced ATM service to their customers.

9 Conclusion and Outlook

The success of future telecommunication networks depends on intelligent network planning to achieve superior service quality, high capacity, and efficient network management. To facilitate this aim, the planning of the systems needs to focus in particular on three major design issues: *a)* the increased demand for telecommunication services, *b)* the emerging of new transmission technologies, and *c)* the increased competition among providers due the deregulation of the telecommunication market.

Conventional telecommunication system design methods do not sufficiently accomplish this task, since their procedures are mostly based on reverse engineering approaches. The methods are often adjusted to isolated design aspects. The application of forward engineering and integrated design procedures can resolve these deficits. Despite their high requirements to system modeling, the new methods are able to achieve demand-oriented, efficient, and optimal network configurations. In this way, the forward engineering design paradigm eliminates some of the heuristics widely applied in networks design and accelerates the deployment of new networks.

This monograph dealt with the implementation of demand-oriented design procedures. The aim was to develop new engineering methods based on forward design paradigm which considers the demand for telecommunication services as a key input factor to system design.

Spatial Customer Traffic Estimation and Characterization

For demand-oriented engineering of wireless and mobile communication networks, the distribution of customers in the service area is of prime importance. Thus, an efficient framework for the estimation and characterization of the expected spatial demand distribution was developed. The traffic estimation of the framework is based on the *geographic traffic model*, which obeys the geographical and demographical factors in the

service area for the teletraffic demand estimation. The characterization of the spatial distribution was facilitated by the application of discrete points, denoted as *demand nodes*. For the generation of the demand node distributions, two hierarchical clustering algorithms, a *recursive partitional* algorithm and an *agglomerative* method, were introduced and evaluated. In particular, the agglomerative clustering procedure meets the requirements of real world planning cases. The *demand node concept (DNC)* permits a matching of theory and practice as claimed by Grillo et al. (1998).

In addition, the work on user clustering presented in this monograph stimulated further investigations of the impact of customer clumping on cellular system performance. New results indicate that a neglect of the clustering effect can lead to critical results: it can significantly decrease the system performance, cf. Tran-Gia et al. (1998), Remiche and Leibnitz (1999).

Demand-oriented Radio Network Synthesis

The application of the demand node concept also facilitates the demand-oriented radio network design in cellular communications systems. The use of the DNC in combination with a new definition of the term *coverage area*, cf. Definition 6.1, permits the formulation of the transmitter location task as a specific facility location optimization problem: the *maximal coverage location problem (MCLP)*.

Unfortunately, by its nature, the MCLP belongs to the class of *NP-hard* optimization problems. Hence, approximation methods are required for obtaining a near optimal solution. Thus, two approximation methods were investigated: a greedy heuristic for solving the MCLP problem and a Simulated Annealing procedure. Both methods were shown to provide efficiently feasible solutions for the maximal coverage location problem. The greedy heuristic turned out here to be more flexible.

To demonstrate the applicability of the proposed procedures for radio network design, the planning tool prototype ICEPT was implemented. Due to its detailed implementation of all the design steps, ICEPT can support the *synthesis* of radio network configurations for real world design scenarios.

The application of covering models in radio network design, like the MCLP, enables the integration of three design techniques from engineering areas, which are, at a first glance, difficult to combine: *a)* facility location science, *b)* traffic engineering, and *c)* RF engineering. The op-

timization procedures which were applied, are able to attain a trade-off between the design objectives. Moreover, a radio network designer receives valuable support from an automatic procedure. In this way, the *integrated approach* to cellular system design meets all requirements for the planning methods of future generation mobile communication networks.

As a result of the research presented in this monograph, *set covering models* have found application in complex radio network optimization tasks, e.g. radio network engineering in CDMA systems, cf. Yu et al. (1998). The methodology is now being transferred into a commercial procedure.

Call Handling Procedures in Cellular Mobile Networks

In addition to spatial teletraffic engineering, the performance of cellular communication systems largely depends on the temporal user behaviour and on the effectiveness of the procedures used to cope with the partitioning of the service area. Therefore, an efficient user model, which can capture the phenomenon of *repeated attempts* of blocked calls, was investigated in combination with two advanced *call handling* mechanisms, namely *guard channels* and *handover retry*.

Each of the advanced mechanisms can help to increase the system performance. However, a combination of the two provides the most effective way of processing calls. The attractive feature of the combined mechanism is that it is organized in a distributed manner, and therefore does not increase the system complexity. In addition, it was demonstrated that reattempt phenomenon significantly decreases the system performance. It is critical to neglect this effect during system design. Both adaptations, the advanced call handling schemes and the improved user modeling, can be used to obtain a more accurate configuration method and an increased performance of the cells in mobile networks.

ABR Service Engineering in Large Scale ATM Networks

Demand-oriented design procedures also gain more and more importance in the engineering of wireline data networks. In particular, the provision of ABR service in large scale ATM system requires a careful network design. Although, ABR service planning, at a first glance, states a contradiction. The objective of providing the ABR service to a large number of customers while assuring certain QoS values can only be attained by

devoting a significant fraction of the ATM trunk to ABR. In order to determine the dedicated bandwidth, appropriate service planning is required.

Three procedures for the performance evaluation of time-oriented ABR service have been investigated: two exact methods and an approximation procedure. However, the large computational complexity of the exact methods prohibits their application in large scale ABR service engineering. Only the approximation method can be applied in practical network planning. Additionally, an evaluation method for volume-oriented traffic has been considered. This method, however, is also limited to small networks.

The capability of the approximation method for planning the ABR service in large scale ATM network was investigated in a case study. The results of the approximation for time-oriented traffic were sufficiently accurate. Hence, the proposed approximation for the connection blocking probabilities facilitates ABR service engineering in large scale ATM networks. In this way, the network operators would be able to design a reliable ABR service and thus would be capable of providing a low-priced ATM service to their customers.

Outlook

At the end of this monograph, there should be some space to raise a basic question. The demand-oriented design procedures presented in this work as well as conventional teletraffic engineering procedures both rely on the assumption that the amount of the teletraffic generated by a customer is determined "a priori", which means that the demand for teletraffic can be related to self-evident propositions. The traffic estimation procedure presented in Chapter 5, for example, widely exploits this concept. However, it is likely that future telecommunication network and service paradigms will question this assumption, since the availability and the quality of the service subsequently also influence the ways in which the networks are going to be used. A fast changing demand for teletraffic for instance will rapidly alter the system performance. This can be illustrated by the Internet where the sheer speed of the network stimulates an increase in the users rate for requesting information. Such a large number of requests significantly changes the system's performance. Thus, new teletraffic models need to be developed which are capable of capturing this behaviour. In addition, new design concepts are required which can scale with the anticipated levels of service provision and the

expected customer behaviour.

References

Aarts, E. and J. Korst (1990). *Simulated Annealing and Boltzmann Machines*. Chichester: John Wiley & Sons.

ANSI (1997). Personal communications services PCS1900 - Air interface specification J-STD-007. Technical reference, American National Standards Institute, New York, USA.

Arulambalam, A., X. Chen, and N. Ansra (1996). Allocating fair rates for available bit rate service in ATM networks. *IEEE Communications Magazine 34*(11), 92–100.

ATKIS (1991). Amtliches Topographisches Kartographisches Informations System; Bavarian land survey office, Munich, Germany.

ATM Forum (1994). User-network interface specification - version 3.1. Technical reference, The ATM Forum Technical Committee, Mountain View, CA, USA.

ATM Forum (1995). Traffic management specification - Version 4.0. Technical reference, The ATM Forum Technical Committee, Mountain View, CA, USA.

Baccelli, F., M. Klein, M. LeBourges, and S. Zuyev (1997). Stochastic geometry and architecture of communication networks. *Telecommunication Systems 7*, 209–227.

Baum, D. (1998). On Markovian spatial arrival processes for the performance analysis of mobile communication systems. In *First Annual Report of the Center of Network Optimization, Würzburg, Germany*.

Bertsekas, D. and R. Gallager (1987). *Data networks*. Englewood Cliffs, NJ: Prentice Hall, Inc.

Calégari, P., F. Guidec, P. Kuonen, and D. Wagner (1997). Genetic approach to radio network optimization for mobile systems. In *Pro-*

ceedings of the IEEE/VTS 47th Vehicular Technology Conference, Phoenix. USA.

Chamaret, B., S. Josselin, P. Kuonen, M. Pizarroso, N. Salas-Manzanedo, and D. Wagner (1997). Radio network optimization with maximum independent set search. In *Proceedings of the IEEE/VTS 47th Vehicular Technology Conference*, Phoenix. USA.

Chang, C.-J., T.-T. Su, and Y.-Y. Chiang (1994). Analysis of a cut-off priority cellular radio system with finite queueing and reneging/dropping. *IEEE/ACM Transactions on Networking 2*(2), 166–175.

Chen, T. M., S. S. Liu, and V. K. Samalam (1996). The available bit rate service for data in ATM networks. *IEEE Communications Magazine 34*(5), 56–71.

Cheung, J. C. S., M. A. Beach, and J. P. McGeehan (1994). Network planning for third-generation mobile radio systems. *IEEE Communications Magazine 32*(11), 54–59.

Chlebus, E. (1993). Analytical grade of service evaluation in cellular mobile systems with respect to subscribers' velocity distribution. In *Proceedings 8th Australian Teletraffic Research Seminar*, pp. 90–101.

Chlebus, E. and W. Ludwin (1995). Is handoff traffic really poissonian. In *Proceedings of the 1995 Fourth IEEE International Conference on Universal Personal Communications*, pp. 348–353.

Choi, B. D., K. B. Choi, and Y. W. Lee (1995). M/G/1 retrial queueing systems with two types of calls and finite capacity. *Queueing Systems 19*, 215–229.

Chung, S.-P. and K. W. Ross (1993). Reduced load approximations for multirate loss networks. *IEEE Transactions on Communications 41*(8), 1222–1231.

Church, R. L. and C. ReVelle (1974). The maximal covering location problem. *Regional Science 30*, 101–118.

Chvatal, V. (1979). A greedy heuristic for the set-covering problem. *Mathematics of Operations Research 4*(3), 233–235.

Cressie, N. (1991). *Statistic for Spatial Data*. New York: Wiley.

Daigle, J. N. and N. Jain (1992). A queueing system with two arrival streams and reserved server with application to cellular telephone. In *Proceedings of INFOCOM'92*, pp. 2161–2167.

Day, J. D. and H. Zimmermann (1983). The OSI reference model. *Proceedings of the IEEE 71*, 1334–1340.

Drezner, Z. (1995). *Facility Location: A Survey of Applications and Methods*. New York: Springer.

El-Dolil, S. A., W.-C. Wong, and R. Steele (1989). Teletraffic performance of highway microcells with overlay macrocell. *IEEE Journal on Selected Areas in Communications 7*(1), 71–78.

ETSI (1995a). European digital cellular telecommunications system (phase 1); Base station system (BSS); Equipment specification (GSM 11.20). Technical reference, European Telecommunications Standards Institute, Sophia Antibolis, France.

ETSI (1995b). European digital cellular telecommunications system (phase 1); GSM DCS 1800 Base station specification (GSM 11.20 - DCS-1800). Technical reference, European Telecommunications Standards Institute, Sophia Antibolis, France.

Falin, G. (1992). A survey of retrial queues. *Queueing Systems 7*, 127–168.

Faruque, S. (1996). *Cellular Mobile Systems Engineering*. Norwood, MA: Artech House Publishers.

Feige, U. (1996). A threshold of ln n for approximating set cover. In *Proceedings of the 28th Annual ACM Symposium on the Theory of Computing*, Philadelphia, Pa. USA.

Floriani, L. C. and G. R. Mateus (1997). An optimization model for the BST location problem in outdoor cellular and PCS systems. In *Proceedings of the 15th International Teletraffic Congress - ITC 15*, Washington DC. USA.

Fortune, S. J., D. M. Gay, B. W. Kernighan, O. Landron, R. A. Valenzuela, , and M. H. Wright (1995). Wise design of indoor wireless systems: practical computation and optimization. *IEEE Computational Science & Engineering 2*, 58–68.

Foschini, G. J., B. Gopinath, and Z. Miljanic (1993). Channel cost of mobility. *IEEE Transactions on Vehicular Technology 42*(4), 414–424.

Fritsch, T. and S. Hanshans (1993). An integrated approach to cellular mobile communication planning using traffic data prestructured

by a self-organizing feature map. In *Proceedings of the 1993 International Conference on Neural Networks*, pp. 822D–822I. IEEE: IEEE Service Center.

Fritsch, T., K. Tutschku, and K. Leibnitz (1995). Field strength prediction by ray-tracing for adaptive base station positioning in mobile communication networks. In *Proceedings of the 2nd ITG Conference on Mobile Communication '95, Neu Ulm*. VDE.

Gamst, A., E.-G. Zinn, R. Beck, and R. Simon (1986). Cellular radio network planning. *IEEE Aerospace and Electronic Systems Magazine* (1), 8–11.

Garey, M. R. and D. S. Johnson (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness*. New York: W.H. Freemann and Company.

Gerlich, N. (1997). On the spatial multiplexing gain of SDMA for wireless local loop access. In *Proceedings of the 2nd European Personal Mobile Communication Conference*, Bonn, Germany.

Ghosh, A. and S. L. McLafferty (1987). *Location Strategies for Retail and Service Firms*. Lexington, MA: Heath.

Glaßer, C. (1998). On thee approximability of problems for cellular networks. In *First Annual Report of the Center of Network Optimization, Würzburg, Germany*.

Grahovac, G. and M. LeBourges (1996). A stochastic geometry model of civil engineering infrastructure in the local loop. In *Proceedings of the 7th International Network Planning Symposium (Networks 96)*, Sydney, Australia.

Grasso, S., F. Mercuri, G. Roso, and D. Tacchino (1996). DEMON: A forecasting tool for demand evaluation of mobile network resources. In *Proceedings of the 7th International Network Planning Symposium (Networks 96)*, Sydney, Australia, pp. 145–150.

Grillo, D., S. T. S. Chia, and N. E. Rouelle (1995). The european path towards advanced mobile systems. *IEEE Personal Communications 2*(1).

Grillo, D., R. A. Skoog, S. Chia, and K. K. Leung (1998). Teletraffic engineering for mobile personal communications in ITU-T work - the need for matching practice and theory. *IEEE Personal Communications 5*(6), 38–58.

Guérin, R. (1988). Queueing–blocking system with two arrival streams and guard channels. *IEEE Transactions on Communications 36*(2), 153–163.

Harris, R. (1998). Private Communication, Royal Melbourne Institute of Technology, Melbourne, Australia.

Hata, M. (1980). Empirical formula for propagation loss in land mobile radio services. *IEEE Transactions on Vehicular Technology 29*(3), 317–325.

Hochbaum, D. S. (1996). *Approximation algorithms for NP-hard problems.* Boston, Ma: International Thomson Publishing Company.

Hong, D. and S. S. Rappaport (1986). Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures. *IEEE Transactions on Vehicular Technology VT–35*(3), 77–92.

Horowitz, E. and S. Sahni (1978). *Fundamentals of Computer Algorithms.* Rockville, Md: Computer Science Press.

Ibbetson, L. J. and L. B. Lopes (1997). An automatic base site placement algorithm. In *Proceedings of the IEEE/VTS 47th Vehicular Technology Conference*, Phoenix. USA.

ITU-T (1992a). Recommendation ITU-T E.490 - Traffic measurement and evaluation - General survey. Technical reference, International Telecommunication Union, ITU, Place des Nations, CH-1211 Geneva 20, Switzerland.

ITU-T (1992b). Recommendation ITU-T E.506 - Telephone network and ISDN - Quality of service, network management and traffic engineering - Forecasting international traffic. Technical reference, International Telecommunication Union, ITU, Place des Nations, CH-1211 Geneva 20, Switzerland.

ITU-T (1993a). Recommendation ITU-T E.600 - Terms and definitions of traffic engineering. Technical reference, International Telecommunication Union, ITU, Place des Nations, CH-1211 Geneva 20, Switzerland.

ITU-T (1993b). Recommendation ITU-T E.800 - Quality of service and dependability vocabulary. Technical reference, International Telecommunication Union, ITU, Place des Nations, CH-1211 Geneva 20, Switzerland.

ITU-T (1993c). Recommendation ITU-T Q.780 - Signalling system No. 7 test specification - General description. Technical reference, International Telecommunication Union, ITU, Place des Nations, CH-1211 Geneva 20, Switzerland.

Jain, A. K. and R. C. Dubes (1988). *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice Hall.

Jonin, G. L. and J. J. Sedol (1976). Telephone systems with repeated calls. In *Proceedings of the 8th International Teletraffic Congress - ITC8*, Munich.

Kaufman, J. S. (1981). Blocking in a shared resource environment. *IEEE Transactions on Communications 29*(10), 1474–1481.

Keilson, J. and O. C. Ibe (1995). Cutoff priority scheduling in mobile cellular communication systems. *IEEE Transactions on Communications 43*(2/3/4), 1038–1045.

Kleinrock, L. (1975). *Queueing Systems, Vol. 1: Theory*. New York: Wiley.

Latouche, G. and V. Ramaswami (1997). Spatial point patterns of phase type. In *Proceedings of the 15th International Teletraffic Congress - ITC 15*, Washington DC. USA.

Leibnitz, K., P. Tran-Gia, and J. E. Miller (1998). Analysis of the dynamics of CDMA reverse link power control. In *Proceedings of the IEEE Globecom '98*, Sydney, Australia.

Leung, K. K., W. A. Massey, and W. Whitt (1994). Traffic models for wireless communication networks. *IEEE Journal on Selected Areas in Communications 12*(8), 1353–1364.

MacDonald, V. H. (1979). The cellular concept. *Bell System Technical Journal 58*, 15–49.

Macfadyen, N. W. (1979). Statistical observation of repeated attempts in the arrival process. In *Proceedings of the 9th International Teletraffic Congress - ITC9*, Torremolinos.

Madhavapeddy, S. and K. Basu (1994). Optimal paging in cellular mobile telephone systems. In *Proceedings of the 14th International Teletraffic Congress - ITC14*, Antibes Juan-les-Pins, pp. 493–502.

Martin, J. (1990). *Telecommunications and the Computer*. Englewood Cliffs, New Jersey: Prentice Hall.

Mathar, R. and T. Niessen (1997). Optimum positioning of base stations for cellular radio networks. Technical report, Department of Mathematics, Aachen University of Technology.

Maxwell, K. (1996). Asymetric digital subscriber line: Interim technology for the next forty years. *IEEE Communications Magazine 34*(10), 100–106.

Mobile Systems International (MSI) Plc. (1996). PlaNET Technical Reference. London, England.

Mouly, M. and M.-B. Pautet (1992). *The GSM System for Mobile Communications*. 4, rue Elisée Reclus, F-91120 Palaiseau, France: published by the authors, ISBN: 2-9507190-0-7.

Okabe, A., B. Boots, and K. Sugihara (1992). *Spatial Tesselations*. Chichester: John Wiley & Sons.

Okumura, Y., E. Ohmori, T. Kawano, and K. Fukuda (1968). Field-strength and its variability in VHF and UHF land mobile radio service. *Review of the Electrical Communication Laboratory 16*(9-10), 825–873.

Papadimitriou, C. H. (1994). *Computational Complexity*. Reading, Ma: Addison-Wesley.

Papadimitriou, C. H. and K. Steiglitz (1982). *Combinatorial Optimization: Algorithms and Complexity*. Englewood Cliffs: Prentice-Hall.

Parsons, D. (1992). *The Mobile Radio Propagation Channel*. London: Pentech Press.

Pinsky, E. and A. Conway (1992). Exact computation of blocking probabilities in state-dependent multi-facility blocking models. In *Proceedings of the IFIP WG 7.3 International Conference on the Performance of Distributed Systems and Integrated Communication Networks*, Amsterdam. The Netherlands, pp. 383 – 392. North-Holland.

Rechtin, E. and M. W. Maier (1997). *The Art of Systems Architecting*. Boca Raton, Florida: CRC Press.

Remiche, M.-A. (1998). Efficiency of an $IPhP^3$ illustrated through a model in cellular networks. Temporary Document COST257TD(98)13, COST.

Remiche, M.-A. and K. Leibnitz (1999). Adaptive soft-handoff thresholds for CDMA systems with spatial traffic. In *Proceedings of the 16th International Teletraffic Congress - ITC 16*, Edinburgh. UK.

Ritter, M. (1998). *Modeling of Flow Control Mechanisms for the Available Bit Rate Service*. Ph. D. thesis, Insitute of Computer Science, University of Würzburg.

Roberts, J., U. Mocci, and J. Virtamo (1996). *Broadband Network Teletraffic - Final Report of Action COST 242*. Berlin: Springer.

Roberts, J. W. (1981). A service system with heterogeneous user requirements - application to multi-service telecommunications systems. In G. Pujolle (Ed.), *Performance of Data Communications Systems and their Applications*, pp. 423 – 431. North Holland-Elsevier Science Publishers.

Roberts, J. W. (1983). Teletraffic models for the telecom 1 integrated services network. In *Proceedings of the 10th International Teletraffic Congress - ITC10*, Montreal. Canada.

Schwarz da Silva, J., B. Arroyo-Fernandez, B. Barani, J. Pereira, and D. Ikonomou (1997). *Mobile and Personal Communications: ACTS and Beyond*, pp. 379–414. Dordrecht: Kluwer Academic Publishers.

Sherali, H. D., C. M. Pendyala, and T. S. Rappaport (1996). Optimal location of transmitters for micro-cellular radio communication system design. *IEEE Journal on Selected Areas in Communications 14*(4), 662–673.

Staehle, D. (1999). Approximate analytic methods for planning and dimensioning the ABR service category in ATM networks. Master's thesis, Insitute of Computer Science, University of Würzburg.

Stoyan, D. and H. Stoyan (1992). *Fraktale — Formen — Punktfelder: Methoden der Geometrie-Statistik*. Berlin: Akademie-Verlag.

Stüber, G. L. (1996). *Principles of Mobile Communication*. Boston: Kluwer Academic Publishers.

Sung, C.-W. and W.-S. Wong (1995). A graph theoretic approach to the channel assignment problem in cellular systems. In *Proceedings of the IEEE/VTS 45th Vehicular Technology Conference*, Chicago. USA.

T-Mobil (1996). Pegasos. Wittichstraße 6, D-64276 Darmstadt, Germany.

Tran-Gia, P. (1982). Überlastprobleme in rechnergesteuerten Fernsprechvermittlungssystemen - Modellbildung und Analyse. Bericht 36, Institut für Nachrichtenvermittlung und Datenverarbeitung, Universität Stuttgart. In German.

Tran-Gia, P. (1988). Zeitdiskrete Analyse in verkehrstheoretischer Modelle in Rechner- und Kommunikationssystemen. Bericht 46, Institut für Nachrichtenvermittlung und Datenverarbeitung, Universität Stuttgart. In German.

Tran-Gia, P. and N. Gerlich (1996). Impact of customer clustering on mobile network performance. Research Report Nr. 143, University of Würzburg, Institute of Computer Science.

Tran-Gia, P., N. Jain, and K. Leibnitz (1998). Code division multiple access wireless network planning considering clustered spatial customer traffic. In *Proceedings of the 8th International Network Planning Symposium (Networks 98)*, Sorrento, Italy.

Tran-Gia, P. and M. Mandjes (1997). Modeling of customer retrial phenomenon in cellular mobile networks. *IEEE Journal on Selected Areas in Communications 15*(8), 1406–1414.

Veeravalli, V., A. Sendonaris, and N. Jain (1997). CDMA coverage, capacity and pole capacity. In *Proceedings of the IEEE/VTS 47th Vehicular Technology Conference*, Phoenix. USA.

Vohra, R. V. and N. G. Hall (1993). A probabilistic analysis of the maximal covering location problem. *Discrete Applied Mathematics 43*, 175–183.

Wirth, P. E. (1997). The role of teletraffic modeling in the new communication paradigms. *IEEE Communications Magazine 35*(8), 86–93.

Wright, M. H. (1998). Optimization methods for base station placement in wireless applications. In *Proceedings of the IEEE/VTS 48th Vehicular Technology Conference*, Ottawa. Canada.

Yang, T. and J. G. C. Templeton (1987). A survey on retrial queues. *Queueing Systems 2*, 203–233.

Yoon, C. H. and C. K. Un (1993). Performance of personal portable radio telephone systems with and without guard channels. *IEEE Journal on Selected Areas in Communications 11*(6), 911–917.

Yu, C., S. Subramanian, and N. Jain (1998). CDMA cell cite optimization using a set covering algorithm. In *Proceedings of the 8th International Network Planning Symposium (Networks 98)*, Sorrento, Italy.

Zeng, Q.-A., K. Mukumoto, and A. Fukuda (1994). Performance analysis of mobile cellular radio system with priority reservation handoff procedures. In *Proceedings of the IEEE/VTS 44th Vehicular Technology Conference*, Stockholm, Sweden, pp. 1829–1833.

ISSN 1432 – 8801