Cell Design in a Cellular System
Using Guard Channels, Call Queueing and Channel Borrowing.

by

Nikhil Jain

Dissertation submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Industrial and Systems Engineering

APPROVED

Prof. John D. Daigle: Co-Chairman

Prof. Hamid D. Sherali: Co-Chairman

Prof. Monte B. Boisen

Prof. Joel A. Nachlas

Prof. Jeffrey Tew

Prof. John E. Kobza

December, 1993

Blacksburg, Virginia

Cell Design in a Cellular System Using Guard Channels, Call Queueing and Channel Borrowing

by

Nikhil Jain

Committee Chairman: Prof. John N. Daigle and Prof. Hanif D. Sherali

Industrial and Systems Engineering

Abstract

This dissertation develops an analytic framework to undertake cell design in a cellular system. The cell is modeled in a broader sense than ever done before. In our analytical model, we incorporated the use of guard channels, queueing of new calls, and hybrid channel allocation. A numerically stable and efficient solution to a queueing system with two arrival streams having reserved and borrowable servers has been developed. This queueing system is used to model the cell behavior. The model provides valuable insights into the behavior of the cell, and this in turn has been used to devise an efficient stochastic optimization algorithm for determining the minimum number of channels required by the cell.

Our techniques are stable, easy to implement for practical systems and produce optimized solutions quickly. This is particularly useful because we expect that future designs of cellular systems may execute such algorithms on cell-site processors.

Acknowledgements

I express my deepest appreciation to Professor John N. Daigle, my advisor at Virginia Polytechnic Institute and State University, for his guidance and encouragement. In him, I found a perfect advisor and could not have asked for anything more. I thank and congratulate him for a job well done.

I thank Prof. Hanif D. Sherali for being a role model, for driving me towards challenging ideas, for reminding me, in his humble way, of the need to strive for the best and for being my co-chairman.

I am grateful to Prof. Jeffrey Tew for his lessons on simulation and for providing support and advise when I needed it. I thank the rest of my committee for their guidance and encouragement.

I thank my family for standing by me in good and bad times and for providing me with support and love whenever I needed it. Last, but not the least, I thank my friends for their help.

Table of Contents

List of Figures

List of Tables

Chapter 1

The Cellular Telephone System : An Introduction

1.1 INTRODUCTION

The cellular telephone has added a new dimension to our communication capabilities. It offers the security and convenience of keeping in touch with people that are important to us. In 1991 there were roughly 3.5 million users of cellular service. Today, these customers are provided access to cellular service using the 832 channels provided by the FCC by reusing a frequency over and over again at distances far enough so that interference can be avoided.

The existing cellular system is a tribute to the good systems analysis and design undertaken by engineers and scientists. The first step in the design process is to divide a geographical area into cells. Each cell is then assigned a set of frequency channels. These channels are reused over the geographical area. Thus, the system supports a large user population with a limited number of channels by reusing channels. The process of first dividing the coverage area into cells and then allocating channels to calls that need them is a difficult problem. The solution procedure involves partitioning the overall problem of cellular system design into three subproblems. These are the economic problem, the cell design problem and the channel assignment problem. These problems are described in Chapter 3. In this dissertation, we first define the overall cellular system design problem and then develop techniques to efficiently tackle the cell design subproblem.

This dissertation is divided into seven chapters. The first two chapters present background material. Chapter 3 defines the cellular system design problem. Chapters 4, 5 and 6, develop solution techniques and Chapter 7 presents conclusions and summary. The remainder of this chapter presents the history of the cellular telephone system and then describe how a typical system works.

1.2 DEVELOPMENT OF CELLULAR TELEPHONE SYSTEMS

The history of mobile communications started in the 1880s with the first experiments of the radio pioneers. In 1897, Marconi transmitted a message to a tugboat located 18 miles away from

the Isle of Wight. During World War I, radio communication was used sparingly and, in 1921, the first land-based radio-telephone system was installed for use by the Detroit police department. The New York police department acquired a similar system in 1932.

The initial radio-telephone systems operated in the 2-MHz frequency band. In 1933, the Federal Communications Commission (FCC) authorized four channels in the 30-40 MHz band on an experimental basis. In 1938, this authorization was extended for regular service. World War II temporarily halted the installation of commercial systems. However, the technological advances made during the War made it possible to exploit higher frequencies. Thus, in 1945, experiments directed specifically towards mobile systems in the 150 MHz range were started at Bell Telephone Laboratories.

In 1946, the FCC allocated a few channels in the 45 and 150 MHz ranges. Commercial service began with installations of systems at Green Bay, Wisconsin (45 MHz) and St. Louis, Missouri (150 MHz). Operation was a simple "push-to-talk" and the call placement was handled by a mobile telephone operator. The mobile customer had to search manually for an idle channel before placing a call. In 1956, similar service was introduced in other places with newly authorized channels in the 450-MHz band.

Bell Laboratories continued to work on the radio-telephone with the twin objectives of improving service and pushing to a higher frequency. In 1964, a 150-MHz Bell Telephone system became available. It provided full duplex operation, automatic channel search, and dialing to and from the mobile station. This was followed in 1969 with the same kind of improved service in the 450 MHz region.

In the years following World War II, many kinds of mobile telephone systems were introduced. These generally operated at frequencies below 460 MHz and provided services to specialized groups such as public safety, industrial and land transportation, and private individuals. Citizen bands in the 460MHz were established and had a user population of about 800,000 in 1970. In May of 1970, the FCC allocated bandwidth in UHF band from 806 to 881 MHz for domestic public land mobile services (Docket 18262). This allocation resulted in a forty fold increase in bandwidth.

On January 4, 1979, the FCC authorized Illinois Bell Telephone Company (IBT) to install a developmental cellular telephone system in Chicago and offer limited commercial cellular service to the public. In addition, American Radio Telephone Service, Inc., (ARTS) was authorized to operate a cellular system in the Washington, D.C.-Baltimore, MD area.

In 1980, the FCC reconsidered its one-system-per-market strategy and allocated frequencies so that two companies could provide cellular service within a given market. Today, each telephone company is either assigned frequencies from band A or band B. Each band has 416 channels. As stated previously, in December 1991, there were about 3.5 million subscribers of cellular telephone service in the US. This number is expected to double by 1994 (Lee 1991). Similar growth is expected in other countries around the world.

In recent days, controversy about the health hazards of cellular telephone have surfaced. There have been a few cases of brain cancer among people who were intensive users of cellular telephone. The health hazards of the cellular telephone has not been completely understood. The connection between brain cancer and electromagnetic waves emitted by a cellular telephone continues to be investigated.

1.3 THE CELLULAR TELEPHONE SYSTEM

In this section, we first identify the important components of a cellular system and then describe how a call is setup, maintained, and broken down. This section is divided into four subsections. In Subsection 1.3.1, we describe the basic components of the cellular system. Subsections 1.3.2 thru 1.3.4 describe how calls are setup, maintained, and broken down.

1.3.1 Basic Components

The basic cellular system consists of the mobile telephone switching office (MTSO), the mobile units, and the cell-sites. These are discussed below.

The MTSO, which is the central coordinating element for the entire system, contains the cellular processor and the cellular switch. The MTSO interfaces with the wireline local telephone company,

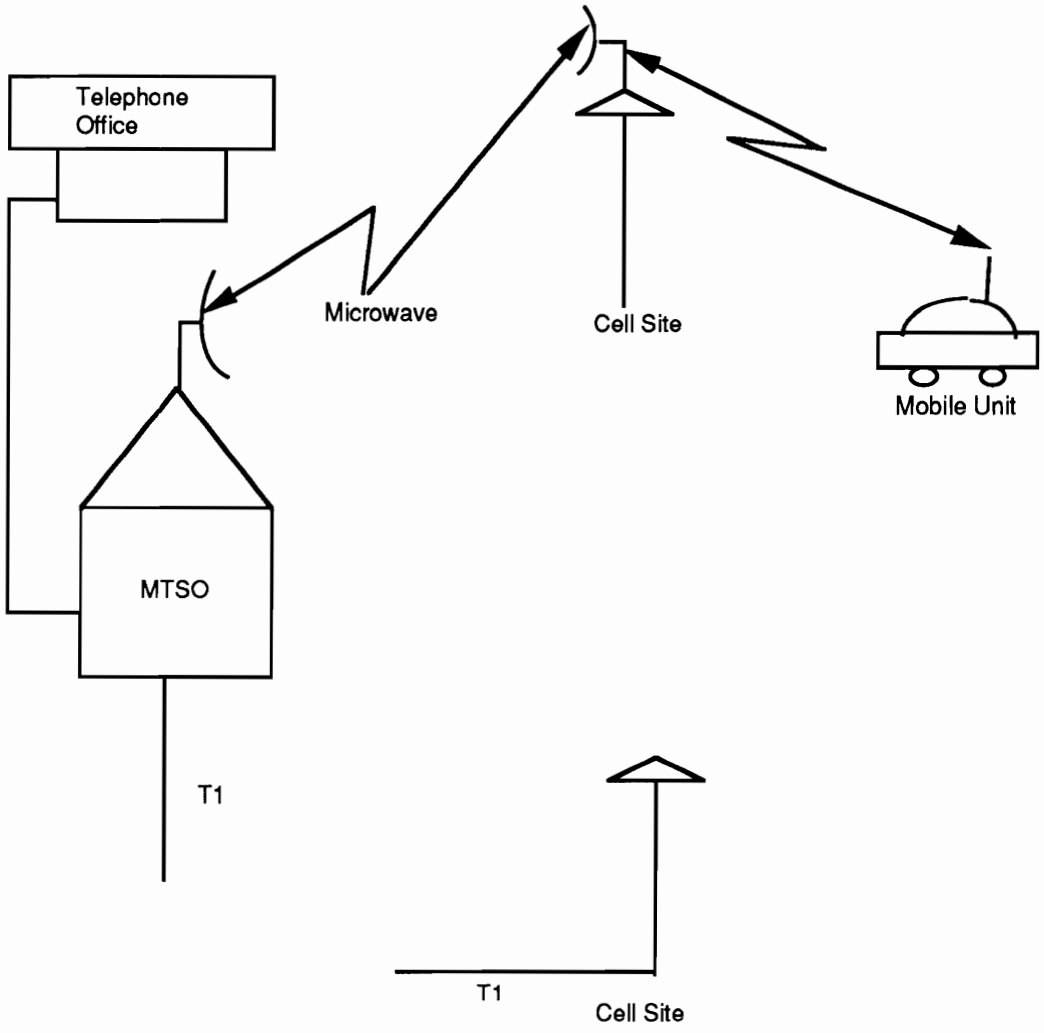

Figure 1.3.1 Cellular system components

controls call processing, and handles billing activities (Figure 1.3.1).

The mobile unit is either hand-held or vehicle-mounted. The customer uses the mobile unit to communicate with either another mobile customer or a wireline telephone user. This communication occurs with the help of a cell-site.

The cell-site serves as an interface between the mobile unit and the MTSO. The cell-site has radio equipment, antennas, a power plant, and data terminals.

Band	Mobile	Base	Two System Market
Band A	824-835, 845-846.5	869-880, 890-891.5	Non-wire-line
Band B	835-845, 846.5-849	880-890, 891.5-894	Wire-line

Figure 1.3.2 Bandwidth allocation

The cellular system is made up of cells. These cells are grouped into groups of k cells called clusters. Dividing the coverage area into cells enable a limited bandwidth system to service a very large number of users. This is accomplished by the concept of frequency reuse. The set of frequency channels that is used in one cell is reused in another cell separated by a distance of D. The cells that use the same set of frequency channels in the system are called co-channel cells. An important design parameter of the cellular system is the ratio $\frac{D}{R}$ where R is the radius of the cell. From the system's $\frac{D}{R}$ ratio, we can determine K, the number cells in a cluster (Lee[1989]).

The cellular frequency band is partitioned into a total of 832 full duplex voice channels. These are divided between two competing cellular communication providers each having 416 channels occupying about 25 MHz of spectrum in the 800 MHz region (Figure 1.3.2).

Each duplex channel consists of a forward and a reverse link. The forward link operates in the direction from the cell-site to the mobile user and the reverse link operates from the mobile user to the cell-site. These links operate 45 MHz apart and require a bandwidth of 30 KHz each. Thus each channel uses a total bandwidth of 60 KHz. Among the 416 channels allocated to each cellular communication provider, 21 are control channels and the remaining 395 are voice channels. The control channels (also called setup and paging channels) are used to set up calls, pass control information and page mobile units.

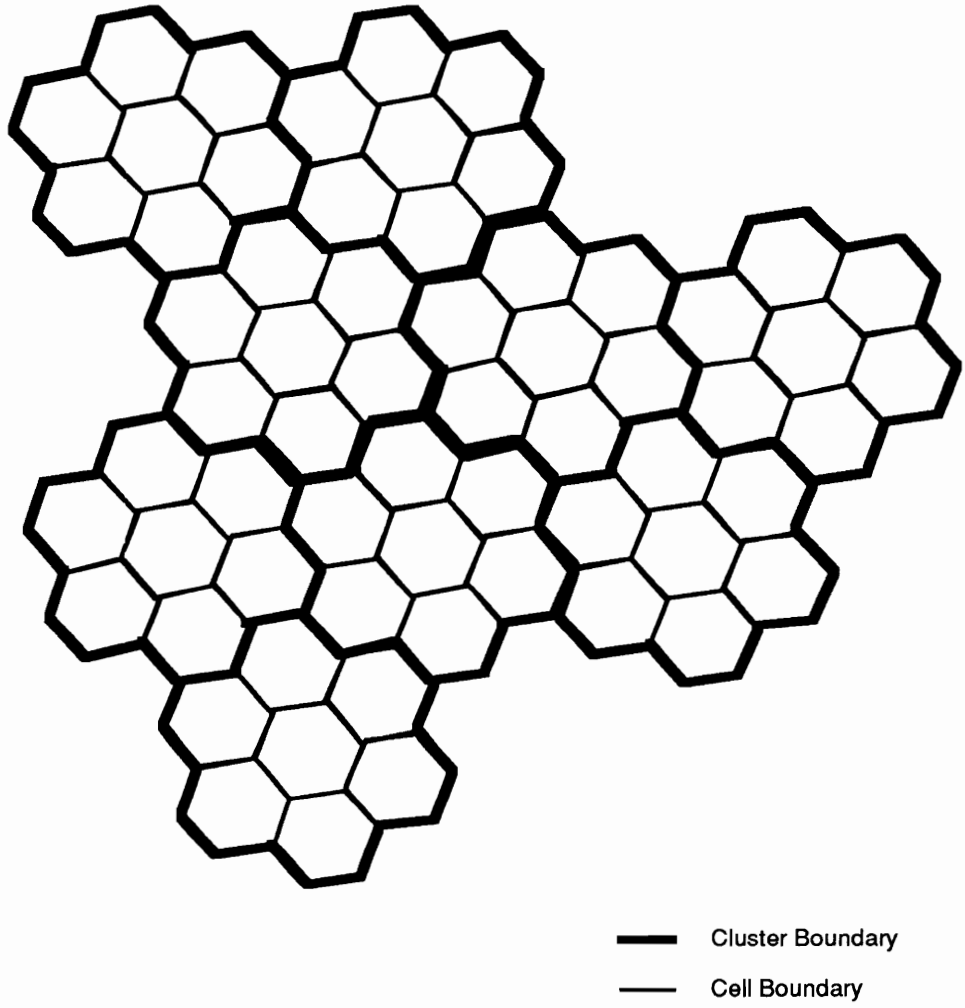

Figure 1.3.3 $K = 7$ cell geometry.

If the value of $K = 7$, then the system has a 7-cell-cluster geometry. This implies that about 56 voice channels (about one seventh of 395) are assigned to each of the seven adjoining cells in a cluster. This seven cell geometry is then replicated over the whole region In Figure 1.3.3 we show a system with a 7-cell-cluster geometry.

Calls may be initiated either by a cellular telephone or a wireline telephone. In the next section, we describe how calls are set up and maintained.

1.3.2 Call Set Up

We first consider the case when the mobile user initiates a call by turning the mobile unit on. The mobile unit scans all the forward control channels, then locks on the strongest one and receives the system status information. Meanwhile, the user enters the desired telephone number and presses the "send" key. The mobile unit verifies the entered number and then sends the number dialed, its own identification number, and a request for a voice channel using the previously selected control channel. The MTSO checks the caller's access privileges, allocates a voice channel and routes the call to the public telephone network. When the called party answers, the MTSO instructs the cell site to begin the call and gives the number of the channel the cell-site has to use, to the cell-site. The cell-site then selects the best antenna to begin transmitting the call. Once the mobile unit begins communication on the allocated voice channel, the setup channel is released for other users to use. At the end of the call, the mobile customer issues an on-hook signal by pressing the "end" button. This releases the voice channel and the mobile unit commences scanning the strongest control channel.

Mobile units automatically register their presence as they move between cell-sites. Their locations are known to be within a collection of cell-sites known as the *paging area.* Now consider the case when a non-mobile, or wireline, telephone user calls a mobile customer. If a wireline user dials a number for a mobile telephone, the telephone exchange recognizes the number as being a cellular telephone's number and forwards it to the MTSO. The MTSO analyzes the number and determines the likely cell-sites where the mobile unit may be found. The MTSO then sends out a paging call to all the cell sites within that paging area. The mobile unit would hear the page and then respond on the control channel. This tells the MTSO which cell-site to use to start the call. A free voice channel is selected and the mobile unit is switched to the voice channel's frequency. The mobile unit is then instructed to switch on the alert tone, which the mobile user can hear. The mobile user answers the call, this causes the mobile unit to sends an "off hook" signal, the voice path is switched on, and the telephone call commences. At the end of the conversation, the voice channel is released

and the mobile unit begins monitoring the strongest forward control channel.

1.3.4 Call Maintenance

During the course of a cellular telephone call, the circumstances of the call may alter. For example, the mobile unit may move out of the area covered by one cell to another, or the quality of the channel may deteriorate, or the mobile unit may change its distance from the cell-site. In each of these instances the system has to take corrective actions to restore the quality of the cellular call. These corrective actions require that control information be passed between the mobile unit and the MTSO. This information is sent through the voice channel using part of the voice channel's bandwidth. The MTSO is responsible for initiating the corrective actions necessary to maintain the quality of the call. This is because the MTSO knows the state of the entire system and can ensure that its actions will not impact the quality of other cellular calls. The two important corrective actions taken by the MTSO to improve quality of communications are transmission power control and handoff. These are discussed below.

Radio signals received by a mobile receiver in a cellular system undergo a propagation loss of 40dB/decade. In other words, a signal reduction of 40dB occurs when the mobile increases its distance from the cell-site by a factor of ten. Simultaneously the mobile unit is subjected to multi-path phenomena. The signal gets reflected from buildings and other objects. This causes the signal strength to vary within small distances. The signal received at the receivers can be strengthened by the use of a longer antenna. However, a longer antenna will also increase the cell reuse distance of the cellular system and therefore reduces the number of users that can be supported by the system. In order to overcome this dilemma, cellular systems are designed so that the power output of the mobile telephone and that of the transmitter of the cell-site are continuously and automatically adjusted from a maximum of several watts to a minimum of a few milliwatts. This ensures that the transmitted power is the minimum required to maintain quality transmission. An additional advantage of power control is the preservation of battery power. Transmission power control is an important feature of a modern cellular system. The power of the transmitted signal is controlled by

the MTSO. The MTSO exchanges power control information with the mobile unit using part of the bandwidth of the voice channel.

If the mobile unit travels too far from the cell-site and reaches the boundary of the cell or enters a pocket where the quality of the communication falls below "toll quality" (not good enough to charge toll rates for it). The cell-site informs the MTSO of its inability to maintain quality communication with the mobile unit. The MTSO instructs the neighboring cell-sites to monitor the signal quality. The MTSO then picks a cell and a channel that are most likely to provide the best quality of signal and asks the mobile unit and the cell-site to undertake a change over. When the mobile unit begins to use a new channel, the previous channel is released for other callers to use. This process is called *handoff* and usually occurs without the intervention of the user. Often the users do not even notice it. We will talk about handoffs in more detail in the next chapter.

1.4 RESEARCH FOCUS OF THE DISSERTATION

Our version of the cellular system design process consists of three steps, at each of which a subproblem is solved and the results are combined. In the first step, we solve the *economic problem.* This involves the determination of an appropriate price for a cellular telephone call, and an expected number of call arrivals given this price. In the second step, we take this estimate of the call arrival process and determine where the cells should be located and how many channels would be required in each cell. This process will be referred to as *cell design.* Finally, in the third step, channels are assigned to cells. This is known as the *channel assignment problem.*

Our focus in this research is on the second step of the cellular system design. This dissertation presents analytical tools that can be used in cell design. In particular, it presents a comprehensive stochastic model of the schemes of handling calls and allocating frequency channels. This model captures the behavior of the cell in a broader sense than ever before. We also present a stochastic optimization of the cell with respect to its frequency resource. The research presented in this dissertation will reduce the reliance on simulation techniques to undertake cellular system design.

In Chapter 2, we present a literature survey. In Chapter 3, we formally define the problem

that this dissertation addresses. In Chapter 4, we present a queueing model of a cell using a Fixed Channel Allocation strategy. In Chapter 5, we develop a queueing system with reserved and borrowable servers to model a cell using Hybrid Channel Allocation. In Chapter 6, the issue of cell design is addressed and finally in Chapter 7, we present a summary and conclusions.

Chapter 2.

Literature Survey

2.1 INTRODUCTION

In the last chapter we described how a call is set up and maintained. Call set up required an allocation of a channel. Call maintenance required handoffs and power control. We will now investigate channel allocation and handoff in more detail.

The channel allocation schemes and handoff management strategies are designed to reduce the probability of a new or a handoff call being denied a channel. The channel allocation scheme has been studied using simulation. In these studies, different schemes for allocation of channels were tried and the probability of a call being denied a channel was determined. Handoff management strategies have been studied using simulation and analytical modeling. Here, the effect of different handoff strategies on the probability of a handoff call being denied a channel was determined.

In this chapter, we will present the available literature on channel allocation schemes and handoff management strategies. During the course of our literature survey we found that in the channel allocation literature, the systems studied, assumed elementary handoff management strategy. In most of these studies, the handoff and the new call traffic were treated alike and calls that were not allocated a channel were removed from the system. Similarly, in the case of the systems studied for handoff management strategy, elementary channel allocation scheme was assumed. Though the stochastic models build using these assumptions provided useful results for some cellular systems, they are not useful for systems that use other more sophisticated channel allocation schemes and handoff management strategies. For these systems, using the above assumptions results in overstating the number of channels required by a cell to provide an acceptable level of call blocking (a new call being denied a channel) and dropping probabilities (a handoff call being denied a channel). An overview of the available literature is provided in Figure 2.1.1.

We divide the chapter into six sections. Section 2.2 presents the literature available on channel

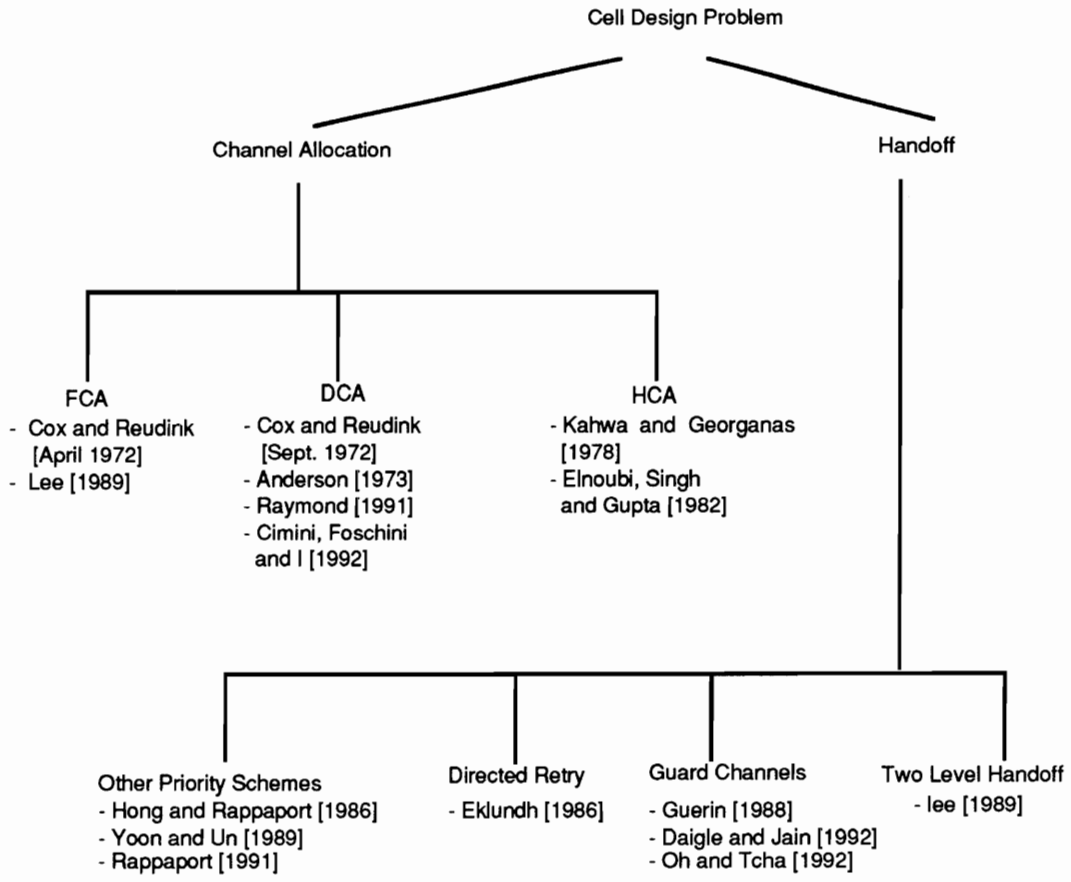

Figure 2.1.1 Overview of the available literature

allocation schemes. In Section 2.3, we discuss handoff. An optimization algorithm for a cell that assumes fixed channel allocation strategy is discussed in Section 2.4, and finally in Section 2.5, we conclude this chapter.

2.2 CHANNEL ALLOCATION SCHEMES

The FCC has allocated limited bandwidth for mobile telephones in the 800 MHz region. This frequency allocation, supports 832 voice channels that are divided equally between two competing cellular telephone companies. Therefore each cellular telephone system has a total of 416 channels available for its customers. To make a cellular call, a voice channel is made available to the mobile

unit. This process is called *channel allocation*. The allocated channel must be free from interference and have acceptable signal to noise characteristics. The channel allocation algorithm's objective is to ensure this.

It should be noted that in the literature, both "channel assignment" and "channel allocation" have been used to describe what was referred to as channel allocation above. In this dissertation, channel assignment is used to describe how channels are permanently assigned to cells or a common pool, and channel allocation is used to describe how the assigned channels are provided to a mobile unit requiring a channel. The channel assignment problem is discussed in Chapter 3.

Several channel allocation schemes have been discussed in the literature. The three important ones are Fixed Channel Allocation (FCA), Dynamic Channel Allocation (DCA) and Hybrid Channel Allocation (HCA). We now give a brief description of each of these.

In FCA, a fixed number of channels are assigned to a cell. These channels are called *nominal channels* and belong exclusively to the cell. The cell uses these channels to service calls that originate within it or are transferred in as a result of a handoff. In Cox and Reudink [1972] and Lee [1989], the FCA scheme is studied using simulation and results pertaining to call blocking and dropping probabilities are presented.

The FCA scheme has an obvious drawback: it does not handle random variations of the traffic load well. For example, if one cell has a large number of callers and an adjacent cell has no callers then there are no channels free in one cell verses a large number of channels free in the other. If the cell had the ability to borrow channels from other cells then it could have borrowed a channel and serviced a temporary increase in the demand. This idea led to the development of the DCA.

In DCA, no channels are assigned to any cell. Instead, all the free channels are in a common pool. When a cell receives a request to either complete a new call or accept a handoff call, it requests a channel from the MTSO. The MTSO then picks an appropriate channel from the common pool and allocates it to the cell. After the cell finishes using a channel, it is returned to the pool for other cells to use. The cell does not have any nominal channels as in the FCA scheme.

Several strategies for borrowing channels have been suggested by Cox and Reudink [1972]. Some of these are listed below:

- **First Available**: This strategy chooses the first available voice channel encountered during a channel search.
- **Nearest Neighbor (NN)**: This strategy chooses a channel that is "in use" at a cell-site nearest to the requesting cell-site, but still far away to avoid interference.

DCA works well under low traffic with high variations. It does not perform well under high uniform load [Cox and Reudink 1972].

Each of the allocation schemes mentioned above has its shortcomings and advantages. A compromise, hybrid channel allocation, has been suggested in Kahwa and Georganas [1978]. In HCA a certain number of channels, n_i, are assigned to the cell. We will refer to these as *nominal channels*. In addition, the system maintains a central pool of channels to be shared among cells. A cell can borrow channels from the pool if all of its nominal channels are being used. For ease of explanation, lets assume that there are a total of N_T channels available for full duplex communication. In this strategy, we divide these channels into two sets of channels: A having n_A and set B having n_B channels. The set A consists of the channels that are nominally assigned to cells and the set B is kept in the common pool. In other words, the set B is allocated using dynamic channel allocation only.

In Kahwa and Georganas [1978], the impact of changing the ratio $\frac{n_A}{n_B}$ on call blocking and call dropping probability is considered. The hybrid channel allocation scheme has been further modified in Elnoubi, Singh and Gupta [1982] to make it more efficient and have better performance at high traffic.

In the strategy described in Elnoubi, Singh and Gupta[1982], each cell is assigned nominal channels. These channels are then arranged in a list. The cell uses the channels at the top of the list first to service calls in the cell. If the cell runs out of the nominal channels, then it can borrow a channel from one of its neighboring cells. The channels are loaned from the back of the list. That

is, the last member of the list is loaned first. If a cell needs to borrow a channel, it borrows it from the neighboring cell that has the largest number of free channels. After the channel is borrowed, it is locked in all the cells that are within the reuse distance. The channel cannot be used in those cells or lent out from them. Call completion causes channel reassignment. First of all, if a nominal channel becomes free in the cell that is using a borrowed channel the call gets transferred from the borrowed channel to the nominal channel in the cell. The borrowed channel is then freed. Further, if a nominal channel is freed then calls are rearranged to ensure that the priority of channel use corresponds with the order of the list. This scheme not only makes bookkeeping easier but also makes the allocation resemble FCA at high traffic intensity and DCA at low traffic intensity.

These algorithms have been modeled using simulation. In one particular study, Elnoubi, Singh and Gupta [1982], the performance measure used was the sum of call blocking and dropping probabilities. The general trends of the probabilities as a function of the load are given in Figure 2.2.1.

In Figure 2.2.1, the curves describe how the sum of call blocking and dropping probability vary with the load. Each curve corresponds to a system with cells having the ratio of nominal channels to the number of channels in the common pool given by n_A:n_B. It can be seen that at low loads DCA performs better. Under high loads the FCA strategy has a better performance measure. The HCA is a good compromise for systems that have variable loads. The allocation scheme described in Elnoubi, Singh and Gupta [1982] performs well under different kinds of loads.

2.3 HANDOFF

A mobile unit usually travels during the course of a telephone call and may move outside the boundary of the cell-site that services it. If this happens, the control of the call is transferred to the new cell into which it moves. This process is called a handoff. The cell-site initiates a handoff whenever it determines that it is not able to continue quality communication with the mobile unit. For the sake of convenience, we will define *origin* to mean the cell from which the mobile moved and the *target* to mean the cell into which the call moved. As described in Chapter 1, the gain of the mobile unit and the cell-site is continually varied to maintain quality communication. When

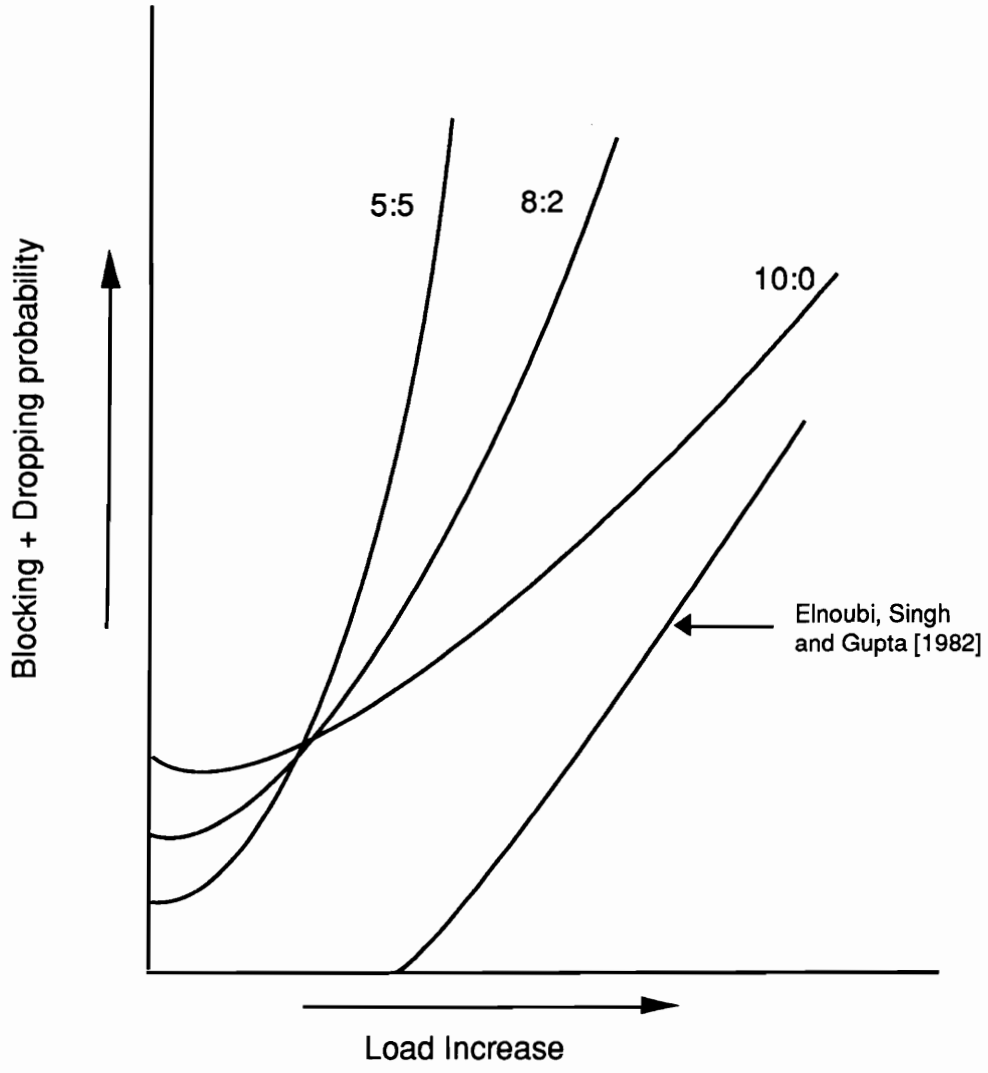

Figure 2.2.1 Comparison of channel allocation schemes

the cell-site's gain control is at the maximum and the quality of communication is deteriorating then the cell-site initiates a request to the MTSO for a handoff. The MTSO first pools all the cells neighboring the origin cell to determine the best target cell. It then tries to allocate a communication channel in the target cell. If it does not have a free channel then the call may be dropped unless the system has some other ad hoc means of managing the handoff call until a channel becomes available

in the target cell. Handoff, if successful, occurs automatically without either the intervention or the knowledge of the user.

Handoff is a very frequent occurrence in the cellular telephone system. As the number of cellular telephone subscribers increases, the size of the must be reduced to enable channels to be reused more frequently. As the cell size decreases, the probability that a given call requires at least one handoff increases. In a typical cellular telephone network, the following has been ascertained [Lee 1989]:

- 0.2 handoffs per cell in a 16-24 km cell,
- 1-2 handoffs per call in a 3.2-8 km cell, and
- 3-4 handoffs per call in a 1.6-3.2 km cell.

Furthermore, as the length of the call increases the handoff probability increases:

- Handoff probability 11.3 % in a 1.76 minute call,
- Handoff probability 18 % in a 3 minute call,
- Handoff probability 42.6 % in a 6 minute call,
- Handoff probability 59.3 % in a 9 minute call.

Handoff in the cellular system is initiated by the MTSO. The intensity of the received signal is observed and the MTSO uses this observation to predict how soon the mobile unit will be out of its present cell-site's area. This prediction is used to generate a handoff request.

Let us assume that a request for a handoff is made. The target cell must now have a channel available to service the call. If it has none, the call is dropped, resulting in lost revenue. The cellular system therefore handles handoff calls differently from the ones that originate in the cell. It tries to reduce the probability of call dropping at the expense of service to new calls.

Several schemes to manage handoff have been suggested in the literature. These include a two-level handoff algorithm, the use of guard channels, queueing of handoff calls, queueing of newly originating calls, and directed retry. These are discussed in more detail below.

2.3.1 Two-Level Handoff Algorithm

During a cellular telephone call, the voice quality is monitored both by the mobile unit and

the cell-site. When the quality of voice falls below an acceptable level, the system measures the intensity of the received signal at the adjacent cells and determines if any of them would service the call better. If it finds a better cell, a handoff may be initiated.

The system uses a two-threshold-level algorithm. These thresholds are based on the ratio of the transmitted power to the received power. For the sake of convenience we will call this the attenuation factor. The system identifies two levels of attenuation factor. When the received signal intensity falls below the first level (U) the system initiates a handoff. The call continues to be carried by the origin cell. Meanwhile, the system tries to identify the target cell by measuring the signal intensity at all neighboring cells. If it identifies a target cell, an idle channel is allocated to the call and the handoff is completed. If a channel cannot be found then the call continues to be carried by the origin cell until such time as a channel becomes free or the received signal intensity crosses a second threshold (L), at which point the call is dropped.

Consider Figure 2.3.1. Let A, B, and C be three neighboring cells. The contours U_i and L_i for $i = A, B, C$ represent equi-attenuation factors. That is, all the mobile units that are at the contour U_A will have the same attenuation factor with respect to cell A. Furthermore, the attenuation factor at U_i is lower than at L_i for $i = A, B, C$.

The handoff process works as follows. Let us assume that the mobile unit is being serviced by the cell-site in cell A and is moving towards cell B. When the caller crosses the contour corresponding to level U_A, the MTSO requests cell B to allocate a channel to the call. At this stage, it may turn out that cell B does not have a free channel. In this event, the call will continue to be carried by cell A until it crosses beyond contour L_A. If there is still no channel available in the cell B, the call is dropped.

The region between levels U and L can be viewed as the time during which the call is queued waiting for a channel to be allocated to it. In this sense, this represents the queueing of handoff calls with balking after a certain waiting time. The amount of time spent in this queue would be a useful design parameter.

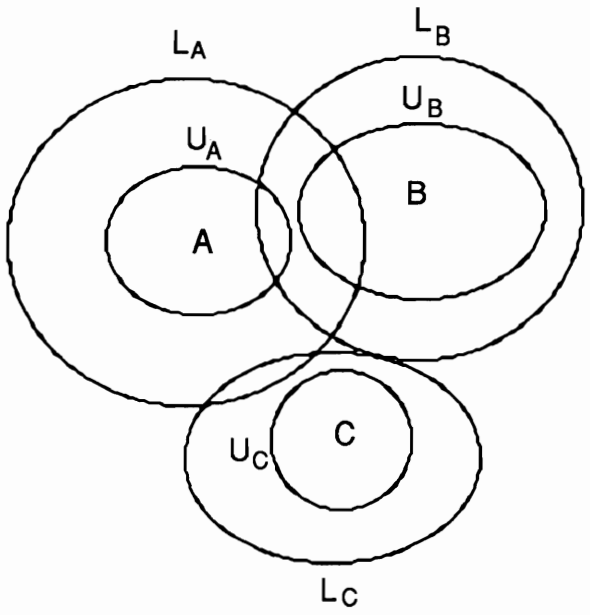

Figure 2.3.1 Two level handoff.

2.3.2 Directed Retry

This is another scheme where a handoff call is queued and waits to acquire a channel. For the ease of explanation, we will describe directed retry with the aid of an example. Consider Figure 2.3.1. Let a mobile unit be presently serviced by the cell-site A, also let it be moving into the area covered by cell-site B. As the mobile unit moves away from cell-site A, its communication quality falls. At some point, the system initiates a handoff to cell-site B. Now, let's assume that cell B does not have a channel free to allocate to the call. The system will then query the adjoining cells and if it finds a channel that would work better than the present channel, the call is then switched to that channel. As soon as the cell B has a free channel, and assuming that cell-site B is still the best choice to service the call, the call is handed off to cell B. This temporarily servicing the call by a less-than-optimal channel represents a queueing of handoff calls. This protocol is discussed in Eklundh [1986] and a FCA was assumed in the paper.

2.3.3 Guard Channels

Guard channels are used to decrease the call dropping probability at the expense of service to new calls. The idea is to set aside a certain number of channels that can only be used by handoff calls, or limiting the number of channels that can be used by originating calls in that cell. This makes some of the channels available exclusively for the calls that are handed over. If a new call arrives when there are g or less channels available then it may either be blocked or queued. Models in which new calls are blocked have been studied (Oh and Tcha [1992] and Lee [1989]) and it has been reported that the use of guard channels reduces the probability of a handoff call being dropped.

The idea of queueing of new calls has been proposed in Guérin [1988]. It has been found that if new calls are queued then the throughput of the system increases while keeping the call dropping probability low. This is desirable from a cellular telephone company's point of view because it represents an increase in revenue.

The systems using guard channels with new call queueing can be modeled using Quasi-Birth and Death Processes. In Jain[1991], an eigenanalysis-based solution approach to Quasi-Birth and Death processes has been suggested. As an illustration of the method the queueing system described in Guérin[1988] was used. The first complete and numerically stable technique to solve the handoff model of the cell using guard channels and new call queueing was presented in Daigle and Jain [1992]. In this study, a modified M/G/1 queueing system with exceptional first service was used to calculate the call dropping probability of a handoff call and the expected delay of a new call. In all of the above mentioned studies on guard channels, FCA was assumed.

As the cellular environment changes, the cell has to dynamically determine the optimal number of channels it would need to provide acceptable performance. In the next section, we discuss a strategy to optimize a cell with respect to the number of nominal channels required.

2.4 CELL OPTIMIZATION

The cellular system designer tries to design cells with low call dropping and blocking probability using a minimum number of channels. Though there are numerous discussions of different handoff

management schemes we found only one, Oh and Tcha [1992], that presents an algorithm to minimize the number of channels required. This study uses guard channels and blocked calls are cleared. It assumes FCA and there is no queueing of newly originating calls. In the following discussion, we briefly present the paper.

Consider a system having M cells. For $i = 1 \ldots M$, denote by n_i the number of channels allocated to cell i. New and handoff calls originate at Poisson rate λ_i and γ_i respectively. The channel holding time of both types of calls is assumed to be exponentially distributed with mean $\frac{1}{\mu}$. Out of the n_i channels, $g_i = n_i - x_i$ are the guard channels. Whenever fewer than x_i channels are occupied in cell i, originating and handoff call attempts are assigned a free channel. Otherwise, if x_i or more channels are occupied in a cell, new call attempts are blocked while handoff call attempts are assigned a free channel if one is available. Handoff calls are dropped only if the n_i channels are all busy. There is no queueing of calls.

The state of the cell at time t is defined as the number of busy channels and is represented by $B_i(t)$. Then $B_i(t)$ for $t \geq 0$ is a birth and death process. Define $P_i(j) = \lim_{t\to\infty} P\{B_i(t) = j\}$. From elementary queueing theory we can obtain the probabilities, $P_i(j)$ for $0 \leq j \leq n_i$. The blocking probability will be given by

$$BN_i(x_i, g_i) = \sum_{j=x_i}^{x_i+g_i} P_i(j) \tag{2.5.1}$$

and the call dropping probability is given by

$$BH_i(x_i, g_i) = P_i(n_i) = P_i(x_i + g_i). \tag{2.5.2}$$

According to Oh and Tcha[1992], the above probabilities have the following three properties:

1. For any channel assignment $BH_i(x_i, g_i) \leq BN_i(x_i, g_i)$.
2. $BH_i(x_i, g_i + 1) < BH_i(x_i + 1, g_i) < BH_i(x_i, g_i)$.
3. $BN_i(x_i + 1, g_i) < BN_i(x_i, g_i) < BN_i(x_i, g_i + 1)$.

Let B^n_{max} represent the maximum blocking probability desired for new calls and B^h_{max} represent the maximum call dropping probability for handoff calls. Then, for a given value of n_i, the optimal

values of x_i, and g_i are determined by solving the following integer programming problem (IP1):

$$\text{Problem IP1}: \quad \text{Minimize } BH(x,g) \tag{2.5.3}$$

Subject to

$$BN(x,g) \leq B_{max} \tag{2.5.4}$$

$$x+g \leq n \tag{2.5.5}$$

$$x,g \geq 0 \; x,g : \text{ integer.}$$

The optimal value of the total number of channels required, can be obtained by solving the following integer programming problem (IP2):

$$\text{Problem IP2}: \text{ Minimize } x_i + g_i \tag{2.5.6}$$

$$\text{Subject to } BN_i(x_i,g_i) \leq B^n_{max} \tag{2.5.7}$$

$$BH(x_i,g_i) \leq B^h_{max} \tag{2.5.8}$$

$$x_i,g_i \geq 0 \; x_i,g_i : \text{ integer.}$$

Using properties 1, 2, and 3 above, Oh and Tcha [1992] present algorithms to solve problems (IP1) and (IP2). Consider the problem IP1. The algorithm begins by setting the number of guard channels equal to 0, and finding $BN_i(x^0,0)$ such that $BN_i(x^0,0) \leq B^n_{max}$. This can be accomplished because the function $BN_i(x_i,0)$ is convex and strictly decreasing (Oh and Tcha [1992]). The value of x_k and g_k is set equal to x^0 and 1 respectively, i.e., increase the previous value of g_i by 1. Increasing g_i may violate the constraint (2.5.4) (see inequality 3 above). If (2.5.4) is violated then we can go back into the feasible region by increasing x_i. Thus by increasing g_i until we leave the feasible region and then increasing x_i to get back to the feasible region we can reach an optimum.

To the best of our knowledge, this is the only algorithm that minimizes the number of channels required in a cell. However, the model of the cell presented in this paper does not model any

queueing of calls. Several researchers have pointed out that queueing of new calls (Guérin [1986]) and a limited queueing of handoff calls (Lee [1989]) increases throughput and reduces call dropping probability. It is therefore desirable to build a modeling framework for such systems.

2.5 CONCLUSIONS

In this chapter we have presented literature relating to channel allocation algorithms and handoff management schemes. It was pointed out that simplifying assumptions about channel allocation and handoff management prevent accurate modeling of the cell. In this dissertation, we develop an analytical model that will improve upon the models presented here, because it will incorporate hybrid channel allocation scheme and queueing of new calls in the model of handoff.

As the environment of the cell changes, the cell may require a larger or smaller number of channels assigned to it. Therefore, an algorithm that determines the optimal number of channels required by a cell would be useful. Developing such an algorithm is also an objective of this dissertation. In the next chapter, we will look at how a cellular system is designed.

Chapter 3

Cellular System Design

3.1 INTRODUCTION

In this chapter, our objective is to present our understanding of the process of cellular system design. During this discussion we will identify the problem that this dissertation addresses and its relationship to the cellular system design.

In the cellular system design process, the frequency channels available for cellular service are used to produce a system, that can be used to provide cellular service. During this design process, channel allocation scheme, handoff management strategy, location, size, and number of channels in the cells, are determined. Through the system design, the cellular company tries to fulfill its business objective. We will assume that the cellular telephone company's business objective is to maximize its profit.

We will also assume that the type of system used to service calls uses guard channels, new call queueing, and HCA. Each cell is allocated n_i *nominal channels.* There is also a common pool of channels to help cells service calls. Cells that run out of channels can temporarily borrow more from the common pool and use them to service handoff calls. Some of the n_i nominal channels, are kept aside as guard channels. These are denoted by g_i.

The definition of "acceptable" service will now be investigated. The three important parameters of service are listed below:

1 The average delay before a user can make a new call (D_A). This can also be stated as the time between the time the user presses the "send" button and the time the user gets a channel allocated to commence the call.

2 The call dropping probability (P_D) is the probability that a handoff call is dropped due to unavailability of a voice channel.

3 The extent of interference (I) is measured as the signal to noise ratio. In general, the extent

of interference that can be tolerated determines how far away a given frequency channel can be reused.

Let C_c be the per second average price of a cellular telephone call, and N_c, the average number of calls in progress. Let the average per second cost of completing a cellular call to the cellular telephone company be given by C_t. The number of frequency channels available to the cellular telephone company is a constant N_T. In the subsequent analysis, when these variables have additional subscripts, they refer to a particular cell. For example, N_{c_i} refers to the average number of calls in progress in cell i.

We will assume that as the per second price for the cellular telephone calls decreases the average number of calls in progress increases. This assumption is made because cellular service is like any other good for which the demand increases as the price is reduced.

We also assume that, as the average number of calls in progress increases the cost of providing service increases. This is due to two reasons. First, since the number of voice channels available to each cell are limited, a larger number of calls in progress can be only supported by increasing the number of cells. As the number of cells increase, more channels will have to be set aside to handle handoff. Additionally, due to interference considerations, the reuse of channels will be further constrained. Thus, as the number of cells increases, the efficiency of the cellular telephone system decreases. In other words, the maximum number of people served by a cell decreases. This increases the value of C_t. Second source of increase in the value of C_t as the number of cells increase, is the increase in the infrastructure cost associated with an increase in the hardware required. Thus the cost of providing cellular service to the customers increase as the number of calls in progress increases.

In order for the system to service N_c calls, let the number of cells required be given by M (assuming that $N_c > N_T$) and the number of channels required in each cell i be given by n_i such that $n_i \leq N_T$ for $i = 0 \ldots M$. Let L represent the total number of channels used in the assignment over all the cells in a cellular system. Furthermore, let $D_A(n_i, g_i)$, $P_D(n_i, g_i)$ and $I(n_i)$ represent expected

delay, call dropping probability and extent of interference, respectively, with n_i total channels and g_i guard channels. Also C_{t_i} is a function of N_{c_i}. Then, under the assumption of maximizing profits, the cellular company tries to solve a problem similar to the following optimization problem (IP3.1.1). The actual problem varies from company to company and is usually more complicated than (IP3.1.1), however, (IP3.1.1) captures the essential elements of the issues involved and is stated below:

$$\text{Max} \sum_{i=0}^{M} \{N_{c_i} C_{c_i} - N_{c_i} C_{t_i}\}$$

Subject to :

$$D_A(n_i, g_i) \leq M_D \text{ for } i = 0 \ldots M$$

$$P_d(n_i, g_i) \leq P_D \text{ for } i = 0 \ldots M$$

$$I(n_i) \leq I \text{ for } i = 0 \ldots M$$

$$L \leq N_T$$

$$0 \leq g_i \leq n_i \text{ for } i = 0 \ldots M$$

$$n_i, g_i, N_{c_i}, M \ : \text{ integers.}$$

If the above problem is solved, it will yield a price to charge, the number of cells required, their locations, and the list of their nominal and guard channels.

In general, the solution of problem (IP3.1.1) is accomplished in three steps, at each of which a subproblem is solved. In the first step, we solve the *economic problem.* This involves the determination of an appropriate price for a cellular telephone call, and an expected number of call arrivals given this price. In the second step, we take this estimate of the call arrival process and determine where the cells should be located and how many channels would be required in each cell. This process will be referred to as *cell design.* Finally, in the third step, channels are assigned to cells. This is known as the *channel assignment problem.* This idea is described in Figure 3.1.1, and each of the subproblems are discussed below.

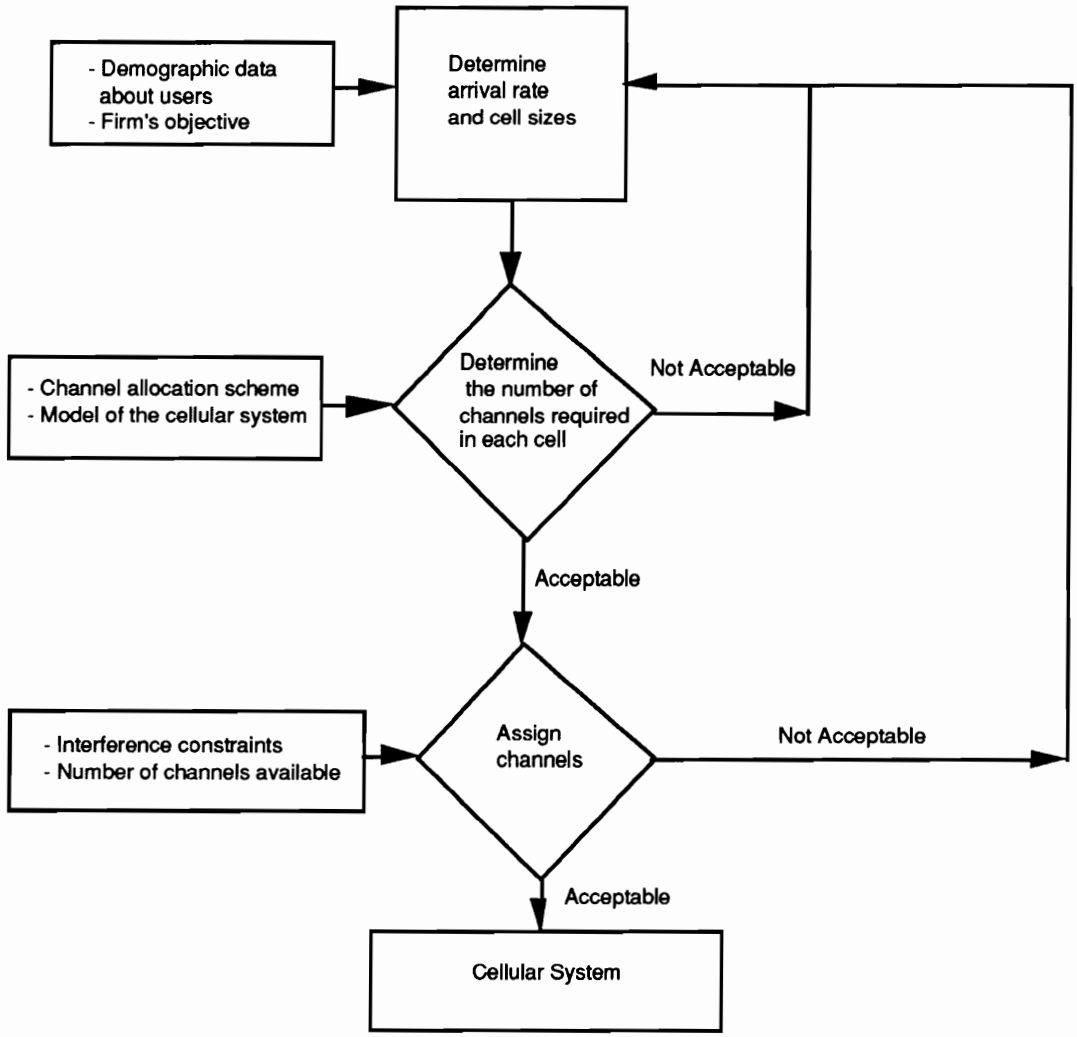

Figure 3.1.1 Cellular system design algorithm

3.2 ECONOMIC PROBLEM

The price of cellular telephone service has long term and short term implications on its demand. Its price today determines the demand for it now. Its price today also has an impact on cellular telephone technology's development in the future. This development affects the demand for the cellular telephone services in the future. Thus setting the price now has long term and short

term profitability implication for the telephone company. This makes setting prices for the cellular telephone service tricky. In general, the company first formulates a business strategy. Then, using an analyst's prediction of the consumer behavior over some planning horizon and estimates of the costs of providing cellular telephone services, it sets prices. The decided price is used to determine the call arrival rate. The technical aspects of the system design are based on this predicted call arrival rate.

3.3. CELL DESIGN PROBLEM

This problem uses the prediction of the call arrival rate to determine the optimal cell size, location, and number of channels required to fulfill operational goals for call dropping probability, and average delays. This problem is not adequately studied in the literature. The difficulty in developing usable solution techniques has stemmed from three sources. First, until recently, call handling has not been modeled in a manner conducive to optimization. The models did not provide a basic understanding of the dynamics of the cell. This prevented the designing of optimization algorithms with a predictable convergence behavior. Second, numerically efficient solution techniques for the queueing model of the cells were unavailable. The first numerically efficient solution technique for solving the queueing model for a cell with guard channels and new call queueing, appears in Daigle and Jain[1992]. Third, the queueing models of the cells did not use Hybrid Channel Allocation to model its behavior. As discussed in Chapter 2, studies on channel allocation schemes have reported better performance by using a Hybrid Channel Allocation strategy. The existing literature assumes FCA in the stochastic models of new and handoff call handling. This causes the number of channels required to be overestimated.

The cell design problem is presently solved in the following manner. The area that the cellular telephone company expects to service is divided into cells. Based on the area of the cell the arrival rates of new calls and handoff calls are determined. An M/M/k model using guard channels with no originating call queueing, (i.e. new and handoff calls balk if no channels are found) and FCA, is

used to determine the number of channels required, n. If this number is more than the number of channels available, then the cell size is reduced and the process is repeated.

In this dissertation, we develop methods that model cells more realistically than a simple M/M/k model and take into account gains due to Hybrid Channel Allocations. We also develop an optimization algorithm that determines the minimum number of channels required for a cell using guard channels, queueing of new calls, and HCA. Since this is the main focus of the cell design problem, the dissertation provides analytical tools to do important steps in cell design.

The issue of location of the cell-site is not addressed in this dissertation because it is dominated by real estate considerations. For example, a cell-site can only be located at a certain place if there is an area available to install the hardware required. We will assume in this dissertation that the location of the cell is given.

3.4. CHANNEL ASSIGNMENT PROBLEM

Having divided the geographical area into cells, and determined the number of channels required in each cell, channels are now assigned to the cells. This assignment is done ensuring that the total number of channels assigned equals the number required and the frequencies of the channels assigned provide interference free communication. This problem has been widely studied in the literature. In Funabiki, Nobuo and Yoshiyasu Takefuji [1992], a neural network parallel algorithm for the channel assignment problems in cellular radio networks is presented. In Duque-Antón, Kunz, and Rüber [1993], channel assignment for cellular radio using simulated annealing is presented.

Using the format presented in Gamst and Rave [1982], this problem can be formally described as follows. Consider an n-cell network. Let the voice channels be labeled by integers such that the voice channels using adjacent frequencies have adjacent integers for labels. The compatibility constraints are described by an $n \times n$ symmetric matrix called C. Each non-diagonal element c_{ij} represents the minimum separation distance between a frequency assigned to cell i and one assigned to cell j. Thus a co-channel constraint is represented by $c_{ij} = 1$. The adjacent channel constraints are given by $c_{ij} = 2$, and $c_{ij} = 0$ implies that cells i and j are allowed to use the same frequency. The

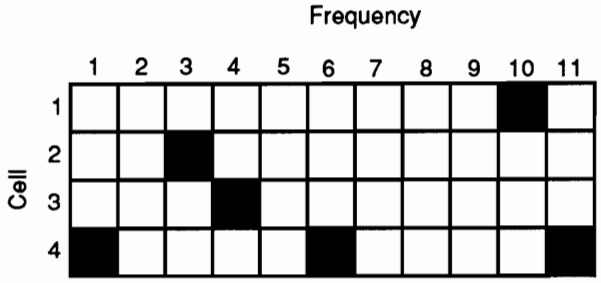

Figure 3.4.1 Optimal assignment for the four cell example

diagonal elements in the matrix C represent the minimum separation between any two frequencies assigned to the same cell. Thus $c_{ii} \geq 1$ is always satisfied, which is also called the *co-site constraint.* The channel requirement for each cell in an n-cell network is described by an n-element vector which is called the *demand vector* D. Each element, d_i, in the demand vector represents the number of frequencies to be assigned to cell i. Let f_{ik} indicate the integer label of the kth frequency assigned to cell i. The electromagnetic constraints can then be represented by:

$$|f_{ik} - f_{jl}| \geq c_{ij} \text{ for } i = 1, \ldots, n;\ j = 1, \ldots, n$$

$$\text{and } k = 1, \ldots, d_i;\ l = 1, \ldots, d_j.$$

The channel assignment problem in the cellular telephone system is to find a conflict-free frequency assignment with the minimum number of total frequencies, when C and D are given.

To illustrate the use of C and D we present a simple four cell example of a cellular system. The matrix C is given by

$$C = \begin{pmatrix} 5 & 4 & 0 & 0 \\ 4 & 5 & 0 & 1 \\ 0 & 0 & 5 & 2 \\ 0 & 1 & 2 & 5 \end{pmatrix},$$

and the vector $D = (1 \quad 1 \quad 1 \quad 3)$. An optimal assignment is given in Figure 3.4.1.

3.5 DISSERTATION PROBLEM STATEMENT

In this dissertation, we develop analytical tools to solve the cell design problem. The problem, that is addressed here is stated next.

Given a cell and its call arrival rate, transfer rate, and its channel allocation scheme, determine the minimum value of nominal channels, n_i, and an optimal value of the number of guard channels, g_i, that will provide acceptable values for the dropping probability of transferred calls, and the average delay of originating calls.

Let $D_A(n_i, g_i)$ represent the average delay of originating calls and $P_D(n_i, g_i)$ represent the dropping probability of transferred calls in a cell having a total of n_i channels of which g_i are guard channels. Furthermore, let M_D and M_P represent the maximum allowable delay and the dropping probability, respectively. The above problem can then be written as the following optimization problem (P3.5.1)

$$\text{Minimize } n_i$$

$$\text{Subject to}$$

$$P_D(n_i, g_i) \leq M_P$$

$$D_A(n_i, g_i) \leq M_D$$

$$0 \leq g_i \leq n_i; \quad n_i, g_i : \text{ integer.}$$

It should be noted that for practical systems n_i cannot be larger than the total number of channels available to the cellular telephone company.

3.6 METHODOLOGY

Our system services two kinds of calls. The new calls that originate in the cell, and the handoff calls that arrive into the cell due to the callers moving into the cell. If a new call arrives, it is allocated a channel if the number of channels available is greater than the number of guard channels. Otherwise the call is queued in a first-come-first-served infinite queue. If a handoff call arrives, it is allocated a channel if one is available else it is dropped.

We will assume that the newly originating calls in the cell arrive at a Poisson rate λ, and calls transferred into the cell from other adjoining cells arrive at a Poisson rate γ. A channel is held

by a call from the time it is allocated to the call until the time the call either completes, or is handed over to another cell. We assume that all calls in progress are alike, and the channel holding time is exponential with mean $\frac{1}{\mu}$. These assumptions are in line with those presented in Chapter 2 (Guérin[1988], Oh and Tcha[1992], Lee[1989], Rappaport[1991], Yoon and Un[1989]). The above assumptions are used in the cellular industry and have been found to provide useful results. However, it should be cautioned that, before using the queueing models developed in this dissertation, these assumptions must be verified in the context of the situation under consideration.

In Chapters 4 we develop techniques to determine the quantities $P_D(n_i, g_i)$ and $D_A(n_i, g_i)$. In Chapter 5 we incorporate HCA and improve the techniques developed in Chapter 4 to calculate $P_D(n_i, g_i)$ and $D_A(n_i, g_i)$. In Chapter 6, methods to solve the optimization problem (P3.5.1) are developed .

Chapter 4.

Queueing Model of a Cell with Fixed Channel Allocation

4.1 INTRODUCTION

In this chapter and in the next one we will present queueing models of a cell in a cellular system. In this chapter we will assume that the cell has a fixed number of channels n, i.e., uses FCA. The cell has g guard channels and will queue newly originating calls if fewer than $g+1$ channels are free. This model and its analysis are presented in Daigle and Jain [1992].

We will begin this chapter by describing the model and the notation in detail. In Section 4.3, we present the matrix geometric approach to solve the queueing system. In Section 4.4, we provide a solution technique using an M/G/1 queueing system with exceptional first service (Daigle[1989]). Finally, results and conclusions are presented in Section 4.5.

4.2. MODEL DESCRIPTION

We begin by a brief description of the model and the notation used. As discussed in Section 3.6, calls originating in the cell area arrive at Poisson rate λ, and calls transferred into the cell from other adjoining cells arrive at Poisson rate γ. All calls are serviced at Poisson rate μ. The state of the system is represented by the ordered pair (i,j), where i is the number of call requests in the queue, which we shall refer to as the *level* of the process, and j is the number of calls in progress, which we will also refer to as the *phase* of the process. The probability that the process is in state (i,j) at time t will be denoted by $P_{ij}(t)$, and the corresponding equilibrium probabilities will be denoted by P_{ij}. Additionally, the row vector of equilibrium probabilities for level i will be denoted by P_i. Thus, P_0 is a row vector of dimension $n+1$, and P_i, for $i>0$, is a row vector of dimension $g+1$. For reasons that will become obvious later, we partition the row vector P_0 into two row vectors, P_{0A} and P_{0B}, where the latter has the same dimension as P_i for $i>0$. The infinite row vectors of equilibrium probabilities, $[P_{0A} \quad P_{0B} \quad P_1 \quad P_2 \quad \ldots]$, will be denoted by P. Time dependent probabilities corresponding to the equilibrium probabilities will be denoted by appending the suffix

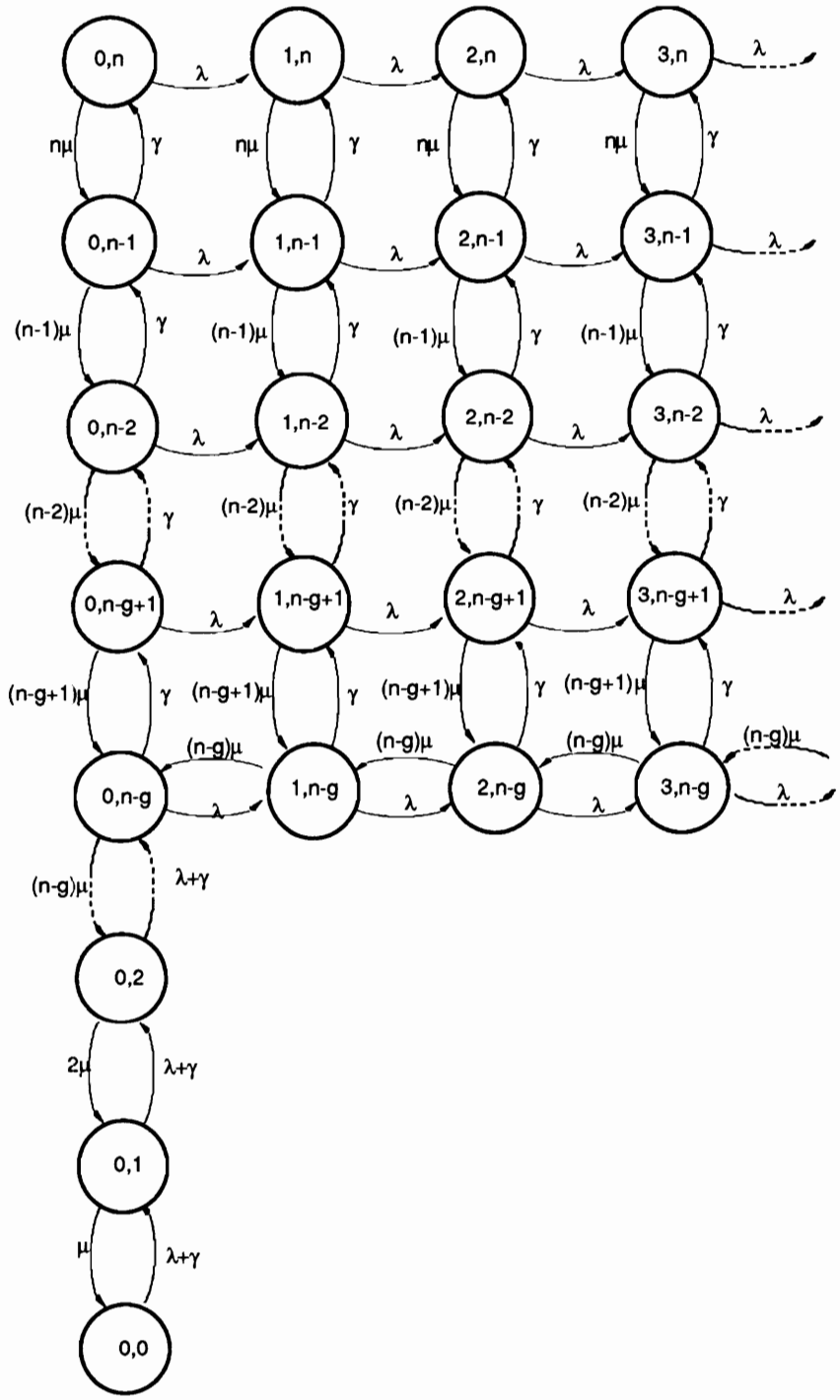

Figure 4.2.1 State diagram for the cellular telephone access system

(t) as in $P(t)$. The differential equations for the probabilities are, as usual, $\frac{d}{dt}P(t) = P(t)Q$, where Q is called the infinitesimal generator, or simply the generator. The dynamics of the system are depicted in Figure 4.2.1. An inspection of Figure 4.2.1 shows that the generator for the continuous-time Markov chain shown in the state diagram has the form

$$Q = \begin{pmatrix} B_{AA} & B_{AB} & \dots & \dots & \dots & \dots & \dots \\ B_{BA} & B_{BB} & A_0 & 0 & \dots & \dots & \dots \\ 0 & B_{1B} & A_1 & A_0 & 0 & \dots & \dots \\ 0 & 0 & A_2 & A_1 & A_0 & 0 & \dots \\ 0 & 0 & 0 & A_2 & A_1 & A_0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}.$$

The matrices B_{AA} and B_{BB} represent transitions among the states in the A and B partitions, respectively, of level zero and are $(n-g)$-square and $(g+1)$-square matrices, respectively. From the state diagram, we find

$$B_{AA} = \begin{bmatrix} -[\lambda+\gamma] & \lambda+\gamma & \dots & 0 & 0 \\ \mu & -[\lambda+\gamma+\mu] & \dots & 0 & 0 \\ & & & \dots & \\ 0 & \dots & \dots & -[\lambda+\gamma+(n-g-2)\mu] & \lambda+\gamma \\ 0 & \dots & \dots & (n-g-1)\mu & -[\lambda+\gamma+(n-g-1)\mu] \end{bmatrix}$$

and B_{BB} is given by

$$\begin{bmatrix} -[\lambda+\gamma+(n-g)\mu] & \gamma & \dots & 0 & 0 \\ (n-g+1)\mu & -[\lambda+\gamma+(n-g+1)\mu] & \dots & 0 & 0 \\ & & & \dots & \\ 0 & \dots & \dots & -[\lambda+\gamma+(n-1)\mu] & \gamma \\ 0 & \dots & \dots & n\mu & -[\lambda+\gamma+n\mu] \end{bmatrix}.$$

The matrices B_{AB} and B_{BA} represent transitions from the states in the A partition to the states in the B partition of level zero and from the states in the B partition to the states in the A partition of level zero, respectively. They have dimension $(n-g)\times(g+1)$ and $(g+1)\times(n-g)$, respectively. From the state diagram, we find

$$B_{AB} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ & & \dots & & \\ 0 & 0 & 0 & 0 & 0 \\ \lambda+\gamma & 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad B_{BA} = \begin{bmatrix} 0 & 0 & 0 & 0 & (n-g)\mu \\ 0 & 0 & 0 & 0 & 0 \\ & & \dots & & \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

The matrix B_{1B} represents transitions between the states of level 1 and the states of the B partition of level 0 and is a $(g+1)$-square matrix. Since transitions in this class occur only from the state $(1, n-g)$,

$B_{1B} = \text{diag}\,((n-g)\mu, 0, \ldots, 0)$. The matrix A_0 is a $(g+1)$-square diagonal matrix that represents transitions from level $i-1$ to level i for $i > 0$, and from the state diagram, $A_0 = \text{diag}\,(\lambda, \lambda, \ldots, \lambda)$. The remainder of the matrices are $A_1 = B_{BB}$ and $A_2 = B_{1B}$.

4.3 THE MATRIX GEOMETRIC APPROACH

This generator has the form of equation (1.5.1) of Neuts [1981], the form for a Markov chain of the G/M/1 type with complex boundary behavior. Neuts [1981] has shown that a Markov chain of this type has a matrix geometric solution. Namely,

$$P_i = P_{0B} R^i, \tag{4.3.1}$$

where R is the rate matrix. In turn, R satisfies the equation

$$R^2 A_2 + R A_1 + A_0 = 0,$$

which can be solved for R using the recursion

$$R_j = A_0 A_1^{-1} + R_{j-1}^2 A_2 A_1^{-1} \quad \text{for} \quad j \geq 1 \tag{4.3.2}$$

with $R_0 = 0$. The determination of the probability masses of the states on the zero level is described in Section 1.5 of Neuts [1981]. Corresponding to the matrix (1.5.2) in Neuts [2], we define the stochastic matrix

$$B(R) = \begin{bmatrix} B_{AA} & B_{AB} \\ B_{BA} & B_{BB} + R B_{1B} \end{bmatrix}.$$

We define $\boldsymbol{x}_0 = [\, x_{0A} \quad x_{0B} \,]$ to be the left eigenvector of $B(R)$ corresponding to its zero eigenvalue. Then, $P_0 = k\boldsymbol{x}_0$. By definition, $\boldsymbol{x}_0 \neq 0$ satisfies

$$\boldsymbol{x}_0 B(R) = 0.$$

The constant of proportionality, k, may be computed by recognizing that

$$P_{0A} + P_{0B} + \sum_{l=1}^{\infty} P_k = 1. \tag{4.3.3}$$

Substituting equation (2.4.1) into (2.4.3) leads to

$$P_{0A} + \sum_{i=0}^{\infty} P_{0B} R^i = 1,$$

so that

$$k[x_{0A} + x_{0B}(I - R)^{-1} e] = 1,$$

where k is the required constant of proportionality, and e is a column vector in which each element is unity. Although the above procedure may appear formidable at first glance, we have found this algorithmic approach very easy to program and our computational experience has been very positive. On the other hand, for the special case under consideration here, we will see in the next section that results may be conveniently obtained in closed form.

4.4 MODIFIED M/G/1 WITH EXCEPTIONAL FIRST SERVICE

In this section, we use the terms *queueing system* and *queue length* to refer to the queue of originating calls for a cell. There is either a queue or there is not. The periods during which the queue length is zero will be referred to as *idle periods,* and the periods of positive queue length will be referred to as the *busy periods.* A careful examination of Figure 4.2.1 reveals that the queue length decreases only if the system is in phase $n - g$. The length of the idle period is then the length of time between a transition from state $(1, n - g)$ to $(0, n - g)$ and the first transition to positive queue length, and the length of the busy period is the length of time between entry to a state at level one and the next transition from state $(1, n - g)$ to $(0, n - g)$. We may define the first service of the busy period as being the time between a transition from level 0 to level 1 and the next downward transition. Denote all other service times (also called ordinary service times) of the busy period as being the time between successive decreases in queue length. In this case, an ordinary service time always begins and ends in phase $n - g$ so that ordinary service times are independently and identically distributed. The first service is exceptional in that the starting phase for the first service is a random variable whose value depends upon the dynamics of the system during the idle period.

That is, the lengths of first services are drawn from a common distribution, but the distribution is not the same as that of the ordinary service times. We will look at the service times in more detail in Section 4.4.2.

In addition, the interarrival times to the queue during the busy period are independent and identically distributed exponential random variables with parameter λ. Thus, during the busy period, the dynamics of the queueing system are exactly those of the M/G/1 system with exceptional first service. However, the interarrival time for the first call in the queue is exactly the length of the idle period, which is not exponentially distributed. On the other hand, the process that counts the successive entries into level zero from the positive occupancy levels is an alternating renewal process. Our approach is to first analyze the system as an M/G/1 system with exceptional service Daigle [1992] and then to scale the resulting probabilities using results from renewal theory to obtain the ergodic probabilities for the actual system.

The remainder of this section is comprised of five subsections. In Subsections 4.4.1 and 4.4.2, we present expressions for the distribution of the lengths of the idle and service times, respectively. The ergodic distribution for the occupancy of the M/G/1 system with exceptional first service is discussed in terms of both classical and matrix analytic approaches in Subsection 4.4.3. In Subsection 4.4.4, we discuss the scaling factor for converting from the M/G/1 with exceptional first service to the actual system. Finally, in Subsection 4.4.5, we present the ergodic probabilities and mean value for the actual system.

4.4.1. Idle Period Distribution

The length of the idle period can be modeled as the time to absorption of the continuous-time Markov chain shown in Figure 4.4.1. As previously mentioned, each entry to level zero from states with positive occupancy occurs at state $(0, n-g)$. The system transitions among the level zero states until it is injected in one of the level one states. Denote the vector of absorption probabilities of the chain shown in Figure 4.4.1 by $\alpha = [\alpha_0\ \alpha_1\ \ldots \alpha_g]$, where α_i is the probability of leaving

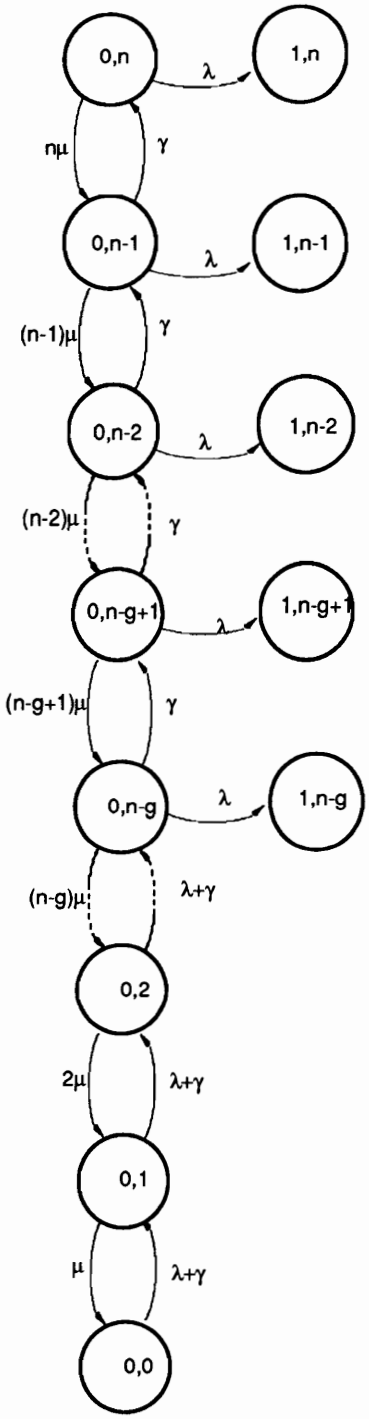

Figure 4.4.1 Dynamics of the idle period

the zero level by entering state $(1, i+n-g)$. Let $P_{\tilde{i}}(t)$ be the vector of probability masses for the states shown in Figure 4.4.1 at a time t. Then $P_{\tilde{i}}(t) = [P_{\tilde{i}_T}(t)\ P_{\tilde{i}_A}(t)]$, where $P_{\tilde{i}_T}(t)$ represents the masses for the transient (zero level) states and $P_{\tilde{i}_A}(t)$ represents the masses for the absorbing states $(1, n-g)$ to $(1, i)$ for $n-g < i \leq n$. We then have the differential equation

$$\frac{d}{dt}[P_{\tilde{i}_T}(t)\ P_{\tilde{i}_A}(t)] = [P_{\tilde{i}_T}(t)\ P_{\tilde{i}_A}(t)] \begin{bmatrix} T & T^0 \\ 0 & 0 \end{bmatrix}, \tag{4.4.1}$$

where $\begin{bmatrix} T & T^0 \\ 0 & 0 \end{bmatrix}$ is the generator for the chain. In the notation of the previous section,

$$T = \begin{bmatrix} B_{AA} & B_{AB} \\ B_{BA} & B_{BB} \end{bmatrix} \quad \text{and} \quad T^0 = \begin{bmatrix} 0 \\ A_0 \end{bmatrix},$$

and $Te + T^0 = 0$. Since $P(t)$ is a vector of probability masses we also have

$$P_{\tilde{i}_T}(t)e + P_{\tilde{i}_A}(t)e = 1, \quad \text{for} \quad t \geq 0. \tag{4.4.2}$$

The system of equations (4.4.1) and (4.4.2) can be solved to yield

$$P_{\tilde{i}_T}(t) = P_{T_0}e^{Tt} \quad \text{and} \quad P_{\tilde{i}_A}(t) = P_{T_0}(e^{Tt} - I)T^{-1}T^0, \tag{4.4.3}$$

where P_{T_0} is the vector of probability masses at time zero. Since all entries to zero level states occur in phase $(n-g)$, P_{T_0} is a vector of zeros except for a one in the $(n-g+1)^{st}$ position. Also, the zero level states are transient so that $\lim_{t\to\infty} e^{Tt} = 0$, and from the definition of α, we find $\alpha = P_{\tilde{i}_A}(\infty)$. Therefore,

$$\alpha = \lim_{t\to\infty} P_{\tilde{i}_A}(t) = -P_{T_0}T^{-1}T^0. \tag{4.4.4}$$

From Neuts [1981], the distribution of the length of the idle period is now found to be

$$F_{\tilde{i}}(x) = 1 - P_{T_0}e^{Tx}\mathbf{e}, \tag{4.4.5}$$

and the i-th noncentral moment, for $i \geq 0$, of $\tilde{i}$ is given by

$$E[\tilde{i}^i] = (-1)^i i! P_{T_0}T^{-i}\mathbf{e}. \tag{4.4.6}$$

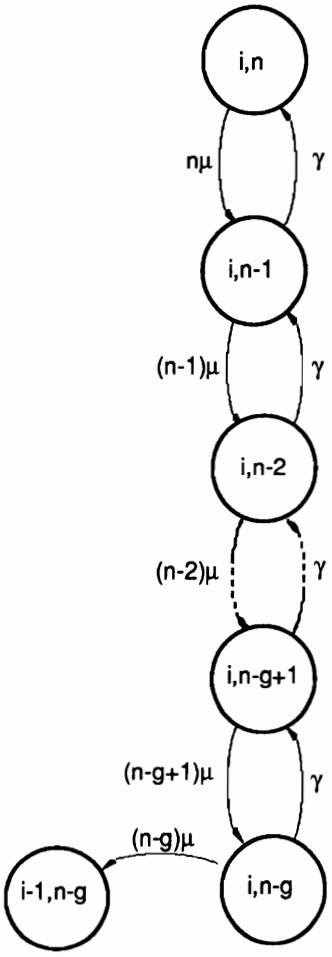

Figure 4.4.2 State diagram for the service times

4.4.2 Service Time Distributions

The service times, both exceptional and ordinary, are also of phase type. In particular, the service times are distributed as the time to absorption in the state $(i-1, n-g)$ for $i > 0$ of the Markov chain whose state transition diagram is shown in Figure 4.4.2. In case of the exceptional first service, the initial state is one of the states $(1, n-g)$ to $(1, j)$ for $n-g < j \leq n$ with the specific phase selected according to the probability vector α, which is defined above. All other services start in phase $(n-g)$. This is because, except for the final service of the busy period, all services end in

the phase $(n-g)$ and a new service begins immediately. The generator of the chain shown in Figure 4.4.2 can be written as

$$\begin{bmatrix} N & N^0 \\ 0 & 0 \end{bmatrix},$$

where N is a $(g+1)$-square matrix, and N^0 is an $(n+1)$ column vector. In particular,

$$N = \begin{bmatrix} -[(n-g)\mu+\gamma] & \gamma & 0 & 0 & \dots & 0 \\ (n-g+1)\mu & -[(n-g+1)\mu+\gamma] & \gamma & 0 & \dots & 0 \\ & & \dots & & & \\ 0 & & \dots & 0 & n\mu & -n\mu \end{bmatrix} \quad \textit{and} \quad N^0 = \begin{bmatrix} \mu(n-g) \\ 0 \\ \dots \\ 0 \end{bmatrix}.$$

Let $\tilde{x}$ denote the length of the ordinary service time and $\tilde{x}_e$ denote the length of the exceptional first service. Then, based on our earlier discussion of distributions of the phase type, we find that

$$F_{\tilde{x}}(x) = 1 - \beta e^{Nx}\mathbf{e}, \tag{4.4.7}$$

and the i-th noncentral moment, for $i \geq 0$, of $\tilde{x}$ is given by

$$E[\tilde{x}^i] = (-1)^i i!\beta N^{-i}\mathbf{e} \tag{4.4.8}.$$

Similarly,

$$F_{\tilde{x}_e}(x) = 1 - \alpha e^{Nx}\mathbf{e}, \tag{4.4.9}$$

and the i-th noncentral moment, for $i \geq 0$, of $\tilde{x}$ is given by

$$E[\tilde{x}_e^i] = (-1)^i i!\alpha N^{-i}\mathbf{e} \tag{4.4.10},$$

where $\beta = [1 \quad 0 \quad \dots \quad 0]$ is a row vector of dimension $g+1$.

4.4.3. Ergodic Probabilities for M/G/1 with Exceptional First Service

To obtain the probability masses for the system of Figure 4.2.1, we first obtain the probabilities of the M/G/1 queueing system with exceptional first service as shown in Figure 4.4.3. The level zero states are replaced by a state 0. The rates at which the process goes from state 0 to the level 1 states is given by the vector $\lambda\alpha$, and the time the process remains in state 0 before making a transition is exponentially distributed with parameter λ. The remainder of the state diagram is the same as

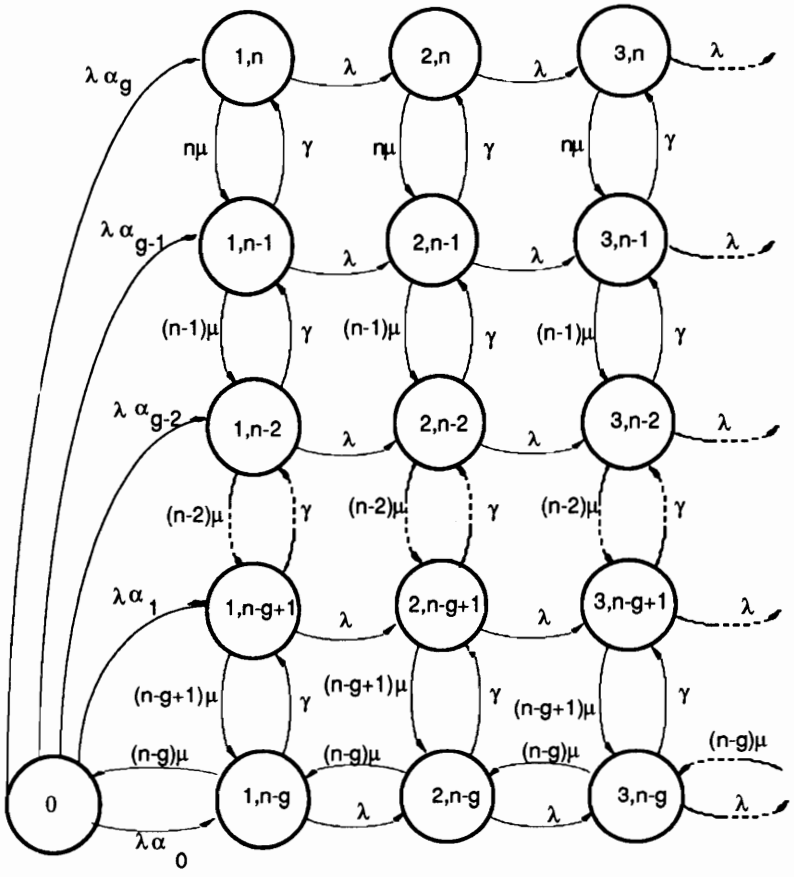

Figure 4.4.3 The M/G/1 queueing system

in Figure 4.2.1. We can now use the Pollaczek-Khintchine (PK) equation for the M/G/1 queueing system with exceptional first service to determine the probability generating function for the number in the queue. In particular, if we define $\tilde{q}_e$ to be the occupancy for the M/G/1 queueing system with exceptional first service, then we find from Daigle [1992] that the PK transform equation for the queue occupancy is given by

$$\mathcal{F}_{\tilde{q}_e}(z) = \frac{1-\rho}{1-\rho+\rho_e} \frac{zF^*_{\tilde{x}_e}(\lambda[1-z]) - F^*_{\tilde{x}}(\lambda[1-z])}{z - F^*_{\tilde{x}}(\lambda[1-z])}. \tag{4.4.11}$$

where $\mathcal{F}_{\tilde{n}}(z)$ denotes the probability generating function for a generic nonnegative, integer-valued random variable $\tilde{n}$, $F^*_{\tilde{x}}(s)$ denotes the Laplace-Stieltjes transform of a nonnegative random variable

$\tilde{x}$, $\rho = \lambda E[\tilde{x}]$, $\rho_e = \lambda E[\tilde{x}_e]$, and

$$P\{\tilde{q}_e = 0\} = \frac{1-\rho}{1-\rho+\rho_e}. \tag{4.4.12}$$

In the particular cases of $\tilde{x}$ and $\tilde{x}_e$, we find from Neuts that

$$F_{\tilde{x}}(s) = \alpha(sI - N)^{-1}N^0 \quad \text{and} \quad F_{\tilde{x}_e}(s) = \beta(sI - N)^{-1}N^0,$$

so that

$$\mathcal{F}_{\tilde{q}_e}(z) = \frac{1-\rho}{1-\rho+\rho_e}\frac{(z\alpha - \beta)(\lambda[1-z]I - N)^{-1}N^0}{z - \beta(\lambda[1-z]I - N)^{-1}N^0}. \tag{4.4.13}$$

Equation (4.4.13) may now be readily inverted to obtain the queue length distribution using discrete Fourier transforms as described in Daigle [1989]. A formula for the mean queue length in terms of the first two moments of the service time distributions defined above is also readily derived from (4.4.11) or (4.4.13) by using standard properties of probability generating functions. In particular, we simply evaluate the first derivative of $\mathcal{F}_{\tilde{q}_e}(z)$ at $z = 1$. The resulting equation is

$$E[\tilde{q}_e] = \frac{1-\rho}{1-\rho+\rho_e}\left[1 + \frac{\lambda E[\tilde{x}_e^2]}{2E[\tilde{x}_e]} + \frac{\lambda\rho E[\tilde{x}^2]}{2(1-\rho)E[\tilde{x}]}\right], \tag{4.4.14}$$

where $E[\tilde{x}^i]$ and $E[\tilde{x}_e^i]$ are specified by (4.4.8) and (4.4.10), respectively. We shall now see that we need not resort to Fourier transform analysis to invert the above PK equation. In particular, if we follow along the lines of Neuts [1981], pp. 84 and 85, we find with minimal algebra that for the case of exceptional first service,

$$P\{\tilde{q}_e = 0\} = \frac{1-\rho}{1-\rho+\rho_e}$$

as above, and

$$P\{\tilde{q}_e = i\} = \frac{1-\rho}{1-\rho+\rho_e}\alpha R^i \mathbf{e}, \tag{4.4.15}$$

where

$$R = \lambda(\lambda I - \lambda \mathbf{e}\beta - N)^{-1}. \tag{4.4.16}$$

We note that the above solution is entirely analytic and that (4.4.16) provides a check on the results obtained in (4.3.2). Additionally, we have the following alternate closed form expression for the

mean queue length which is obtained by simply summing the tail probabilities:

$$E[\tilde{q}_e] = \frac{1-\rho}{1-\rho+\rho_e}\alpha R(I-R)^{-2}\mathbf{e}. \tag{4.4.17}$$

This gives an additional check on (4.4.14).

4.4.4. Scaling Factor

The queue length probabilities of the actual system are different from those of the modified system by a scaling factor δ. We now present a simple argument to explain why this is so and also derive a formula for the factor δ. To aid in describing the theory that leads to the scaling factor, we will refer to all variables for the modified state diagram of Figure 4.4.3 with a subscript M and call it the process M. We will refer to the quantities in the actual state diagram of Figure 4.2.1 by a subscript A and call it the process A. The processes M and A are alternating renewal processes. Each process alternates between busy and idle periods, and the time between successive entries to either a busy or idle period is called a cycle. In each process, the lengths of the busy and idle periods are each statistically independent in successive cycles. For any alternating renewal process, the ergodic probability that the process is in any state (or any set of states) may be computed by taking the ratio of the total expected amount of time spent in the state (or set of states) during a cycle to the expected cycle length. For $r = A, M$, denote the length of the busy period by $\tilde{y}_r$, the length of the idle period by $\tilde{\imath}_r$, the total amount of time spent in state (i,j) during a cycle by $\tilde{s}_{ij_r}$, and the ergodic probability that the system is in state (i,j) by P_{ij_r}. Then, for $r = A, M$,

$$P_{ij_r} = \frac{E[\tilde{s}_{ij_r}]}{E[\tilde{\imath}_r] + E[\tilde{y}_r]}.$$

Thus,

$$\frac{P_{ij_A}}{P_{ij_M}} = \frac{E[\tilde{s}_{ij_A}]}{E[\tilde{\imath}_A] + E[\tilde{y}_A]} \frac{E[\tilde{\imath}_M] + E[\tilde{y}_M]}{E[\tilde{s}_{ij_M}]}.$$

Now, the dynamics of the actual and modified processes are statistically identical during busy periods so that $\tilde{y}_A$ and $\tilde{y}_M$ are identically distributed as are $\tilde{s}_{ij_A}$ and $\tilde{s}_{ij_M}$. We then find that

$$P_{ij_A} = \frac{E[\tilde{\imath}_M] + E[\tilde{y}]}{E[\tilde{\imath}_A] + E[\tilde{y}]} P_{ij_M}, \tag{4.4.18}$$

where $\tilde{y}$ is defined to be the random variable whose distribution is identical to that of $\tilde{y}_M$ and $\tilde{y}_A$. Thus, $P_{ij\,A} = \delta P_{ij\,M}$, where

$$\delta = \frac{E[\tilde{i}_M] + E[\tilde{y}]}{E[\tilde{i}_A] + E[\tilde{y}]}. \tag{4.4.19}$$

Computational formulae for the elements of (4.4.19) are now given. The expected length of the busy period for the M/G/1 system with exceptional first service found in Daigle [1992] and is given by

$$E[\tilde{y}] = \frac{E[\tilde{x}_e]}{1 - \lambda E[\tilde{x}]} = \frac{-\lambda\alpha N^{-1}\mathbf{e}}{1 + \lambda\beta N^{-1}\mathbf{e}}, \tag{4.4.20}$$

where $E[\tilde{x}_e]$ and $E[\tilde{x}]$ were specified in (4.4.8) and (4.4.10), respectively. The length of the idle period for system M is exponentially distributed with parameter λ so that $E[\tilde{i}_M] = 1/\lambda$. Finally, the expected length of the idle period for the actual system is calculated using (4.4.6) with $i = 1$; that is,

$$E[\tilde{i}_A] = -P_{T_0}T^{-1}e. \tag{4.4.21}$$

All quantities needed to specify δ have now been specified in terms of the model parameters. In particular, on substituting (4.4.20) and (4.4.21) into (4.4.19) and simplifying, we find

$$\delta = \frac{1 + \lambda(\beta - \alpha)N^{-1}\mathbf{e}}{(1 + \lambda\beta N^{-1}\mathbf{e})\lambda P_{T_0}T^{-1}\mathbf{e} + \lambda\alpha N^{-1}\mathbf{e}}. \tag{4.4.22}$$

4.4.5. Ergodic Probabilities for the Actual System

We have argued above that the positive level joint ergodic probabilities for the M/G/1 system with exceptional first service scale to the positive level probabilities for the actual system. The scaling factor is specified by (4.4.22). Similar arguments reveal that the marginal positive occupancy probabilities and the positive level probability vectors also scale in exactly the same manner. Therefore, the positive ergodic occupancy probabilities are given by

$$P\{\tilde{q}_A = i\} = \delta P\{\tilde{q}_e = i\} \quad \text{for} \quad i > 0. \tag{4.4.23}$$

The remaining occupancy probability, $P\{\tilde{q}_a = 0\}$, is readily found using a standard result from alternating renewal theory. That is, the ratio of the expected length of idle period to the expected

cycle length yields this result i.e.;

$$P\{\tilde{q}_a = 0\} = \frac{E[\tilde{i}_A]}{E[\tilde{i}_A] + E[\tilde{y}_A]} = \frac{P_{T_0}T^{-1}\mathbf{e}}{(1+\lambda\beta N^{-1}\mathbf{e})P_{T_0}T^{-1}\mathbf{e} + \lambda\alpha N^{-1}\mathbf{e}}. \tag{4.4.24}$$

For the vector quantities that represent level probabilities, we find by scaling (4.4.15) that

$$P_i = \delta\frac{1-\rho}{1-\rho+\rho_e}\alpha R^i \quad \text{for} \quad i > 0. \tag{4.4.25}$$

Further, we know $P_1 = P_{0B}R$ and in this particular case R is nonsingular. Thus,

$$P_{0B} = P_1 R^{-1} = \delta\frac{1-\rho}{1-\rho+\rho_e}\alpha. \tag{4.4.26}$$

The first element of P_{0B} is $P_{0,n-g}$ from which $P_{0,j}$ for $j = 0, 1, \ldots, n-g-1$ can be determined. For example, one can compute these probabilities as follows:

$$\begin{aligned} P_{0,n-g-1} &= \frac{(n-g)\mu}{\lambda+\gamma}P_{0,n-g} \\ P_{0,n-g-2} &= \frac{(n-g-1)\mu}{\lambda+\gamma}P_{0,n-g-1} \\ &\cdots \\ P_{0,0} &= \frac{\mu}{\lambda+\gamma}P_{0,1}. \end{aligned} \tag{4.4.27}$$

All level probabilities are now completely defined. An additional alternate expression for the mean queue length is also obtained by scaling the corresponding expressions for the M/G/1 with exceptional first service. Thus, for (4.4.14) and (4.4.17), we have

$$E[\tilde{q}_e] = \delta\frac{1-\rho}{1-\rho+\rho_e}\left[1 + \frac{\lambda E[\tilde{x}_e^2]}{2E[\tilde{x}_e]} + \frac{\lambda\rho E[\tilde{x}^2]}{2(1-\rho)E[\tilde{x}]}\right], \tag{4.4.28}$$

and

$$E[\tilde{q}_A] = \delta\frac{1-\rho}{1-\rho+\rho_e}\alpha R(I-R)^{-2}. \tag{4.4.29}$$

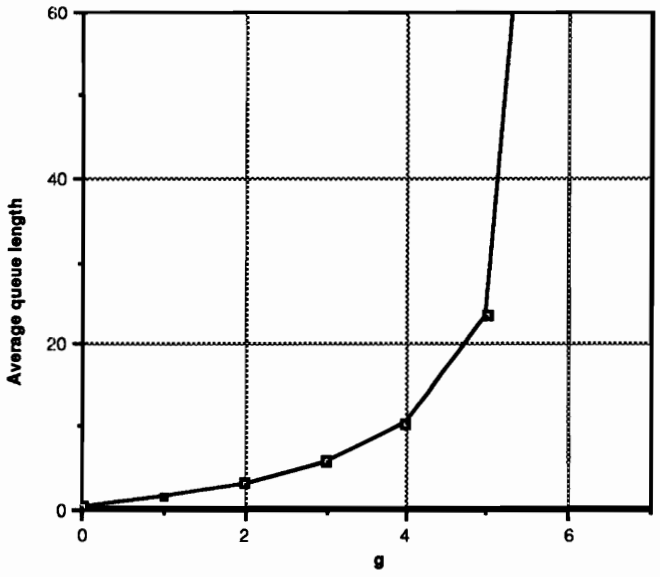

Figure 4.5.1 Average queue length for new customers with n=44, λ=30 and γ=8

4.5 RESULTS AND CONCLUSIONS

In Figures 4.5.1 to 4.5.3, we plot the effect on the performance of the cell with varying values of the number of guard channels. In general it is observed that as the number of guard channels are increased, the queue length, delay and throughput increase. The call dropping probability decreases with an increase in the value of g. These results suggest that if guard channels are used with new call queueing then the cellular company gains in two ways; reduced call dropping probability and increased throughput.

In the analysis in this chapter we assumed a Fixed Channel Allocation. In the following chapter we incorporate ideas of Hybrid Channel Allocation schemes.

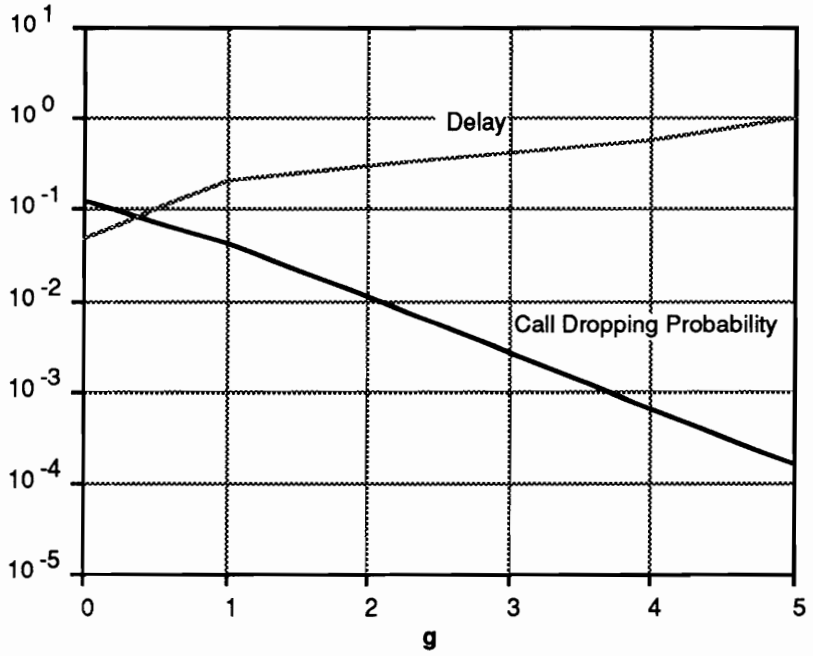

Figure 4.5.2 Expected delay and blocking probabilities for n=44, λ=30 and γ=8

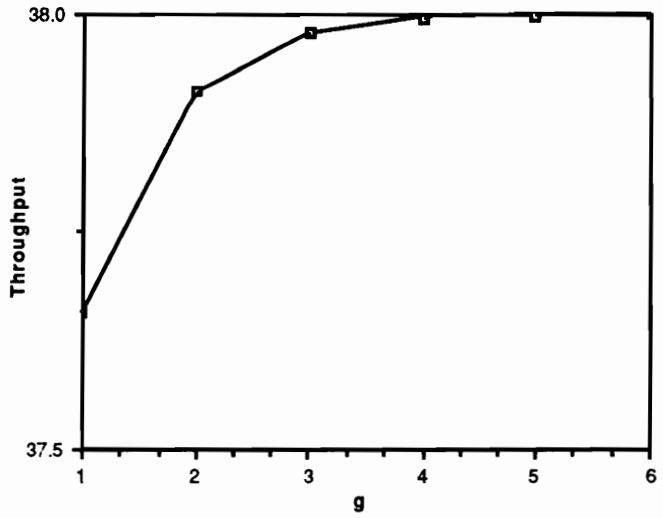

Figure 4.5.3 Throughput verses g for n=44, λ=30 and γ=8

Chapter 5.

Queueing System with Reserved and Borrowable Servers

5.1. INTRODUCTION

In this chapter we describe a queueing system that is used to model the cell behavior in a cellular system. This system has two Poisson arrival streams of customers. Customers with higher priority arrive at a Poisson rate γ and the ones with a lower priority arrive at a Poisson rate λ. It takes an exponentially distributed amount of time with mean $\frac{1}{\mu}$ to complete service. There are a total of n servers available to service customers. If the number of free servers reduces to g or less, the lower priority customers are queued in a first-come-first-served queue. If a high priority customer arrives and does not find a free server then a server can be borrowed from a common pool. At the most, b servers can be borrowed. Let ϕ_1 be the probability of making a successful bid for the first borrowed server, ϕ_2 be the probability of acquiring a second server if a second server is requested, and so on. If a server cannot be found either from the cell, or by borrowing for the high priority customer, then the customer balks. If a server becomes free, and if the cell is using a borrowed server, then the borrowed server is returned to the common pool immediately. This system will be used to model the cell dynamics of a cellular telephone system.

Other applications may include modeling the passenger traffic in an airport for a particular flight. The two arrival streams would correspond to the originating traffic and transfer traffic. Transfer traffic would correspond to passengers that are only changing planes. The service time would model the cancellation process and borrowed servers would represent using seats on other airlines.

In this chapter, we analyze the above queueing model in terms of its application to cell design in a cellular system. The higher priority arrivals correspond to handoff traffic and the lower priority arrivals correspond to new calls originating from within the cell. Each call takes an exponential amount of time with mean $\frac{1}{\mu}$ to complete. If a handoff call arrives and there is no channel available to service it, then one may be borrowed from a common pool of channels. The cell can borrow a

maximum of b channels and the request for the i^{th} borrowed channel is fulfilled with a Bernoulli probability ϕ_i. In the latter part of this chapter, we will discuss possible ways to determine ϕ_i. The analysis presented in this chapter runs parallel to the one presented in Chapter 4. However, the queueing model and the solution technique are different. The solution technique is numerically efficient because its numerical complexity is of order g.

This chapter is organized into five sections. In Section 5.2 we present the notation. In Section 5.3 we present a solution technique. In Section 5.4 we develop an approximation for the channel borrowing probabilities. In Section 5.5 we present some results, and finally, conclusions are presented in Section 5.6.

5.2. MODEL DESCRIPTION

As discussed in Section 3.6, calls originate in a cell area according to a Poisson process at rate λ, and calls are transferred into the cell from other adjoining cells according to a Poisson process at rate γ. It takes an exponentially distributed amount of time with mean $\frac{1}{\mu}$ to service calls. The cell can borrow upto b channels to service transferred calls. The channel allocation strategy is incorporated by the concept of channel borrowing probability ϕ_i for $i = 1, 2, \ldots$, where ϕ_1 is the probability that the cell will be able to borrow one channel and ϕ_2 is the probability that given that the cell has borrowed a channel already, it can borrow another channel, and so on. A borrowed channel is released as soon as a call completes in the cell that borrowed the channel. This may involve reassignment of channels in the cell but a borrowed channel is released as soon as the next channel becomes free in the cell.

The state of the system is represented by the ordered pair (i, j); where i is the number of call requests in the queue, which we will refer to as the *level* of the process, and j is the number of calls in progress, which we will refer to as the *phase* of the process. The probability that the process is in state (i, j) at time t will be denoted by $P_{ij}(t)$, and the corresponding equilibrium probabilities will be denoted by P_{ij}. Additionally, the row vector of equilibrium probabilities for level i will be denoted by P_i. Thus, P_0 is a row vector of dimension $b + n + 1$, and P_i, for $i > 0$, is a row vector of

dimension $g+b+1$. For reasons that will become obvious later, we partition the row vector P_0 into two row vectors, P_{0A} and P_{0B}, where the latter has the same dimension as P_i for $i > 0$. The infinite row vector of equilibrium probabilities, $[P_{0A} \quad P_{0B} \quad P_1 \quad P_2 \quad \ldots]$, will be denoted by P. Time dependent probabilities corresponding to the equilibrium probabilities will be denoted by suffixing (t) as in $P(t)$. The dynamics of a cellular telephone system (S_1) are depicted in Figure 5.2.1. The diagram shows a situation that allows only one channel to be borrowed. The states (i, B) for $i \geq 0$, represent states when one channel is borrowed from the common pool. In this state diagram value of b is 1.

5.3. SOLUTION METHODOLOGY

In the remainder of this chapter, we use the terms queueing system and queue length to refer to the queue of originating calls for a cell. There is either a queue or not. The periods during which the queue length is zero are referred to as idle periods, and the periods of positive queue length are referred to as busy periods.

A careful examination of Figure 5.2.1 reveals that the queue length decreases only if the system is in phase $n-g$. The length of the idle period is then the length of time between a transition from state $(1, n-g)$ to $(0, n-g)$ and the first transition to positive queue length, and the length of the busy period is the length of time between entry to a state at level one and the next transition from state $(1, n-g)$ to $(0, n-g)$.

We may define the first service of the busy period as being the time between a transition from level 0 to level 1 and the first decrease in queue length, and all other service times of the busy period as being the time between successive decreases in queue length. In this case, an ordinary service time always begins and ends in phase $n-g$ so that ordinary service times are independently and identically distributed. The first service is exceptional in that the starting phase for the first service is a random variable whose value depends upon the dynamics of the system during the idle period. That is, the lengths of first services are drawn from a common distribution, but the distribution is not the same as that of the ordinary service times.

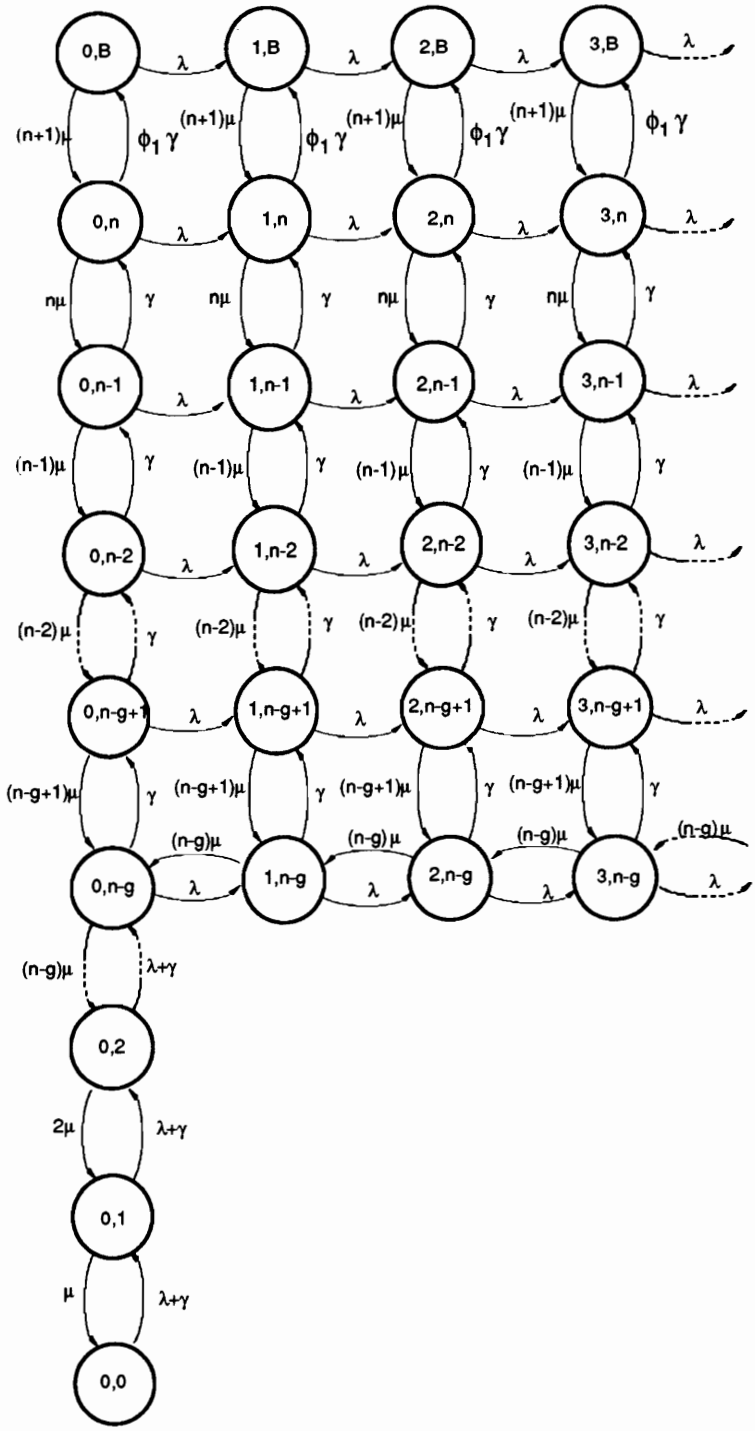

Figure 5.2.1. State diagram for the cellular telephone access system

In addition, the interarrival times to the queue during the busy period are independent and identically distributed exponential random variables with parameter λ. Thus, during the busy period, the dynamics of the queueing system are exactly those of the M/G/1 system with exceptional first service. But, the interarrival time for the first call of the busy period is exactly the length of the idle period, which is not exponentially distributed. On the other hand, the process that counts the successive entries into level zero from the positive occupancy levels is a delayed alternating renewal process. Our approach is to first analyze the system as an M/G/1 system with exceptional first service, as described in Figure 5.3.1 (S_2), and then to scale the resulting probabilities, using results from renewal theory to obtain the ergodic probabilities for the actual system.

In subsection 5.3.1, we describe the procedure for calculating the ergodic distribution of system S_2. In subsection 5.3.2, we deal with the process of determining the probability distribution of the actual system S_1, using the solution to system S_2.

5.3.1 Analysis of the M/G/1 system (S_2)

Using the same format presented in Neuts[1981], we define $(b+g+1)$ row vectors $B^0 = [1 \quad 0 \ldots \quad 0]$, $S^{0T} = [(n-g)\mu \quad 0 \quad \ldots \quad 0]$ and $e^T = [1 \quad 1 \ldots \quad 1]$. We further define the following four $(b+g+1)$-square matrices:

$$A_2 = S^0 B^0,$$

$$B^{00} = eB^0,$$

$$B_0 = \begin{pmatrix} -\gamma & \gamma & \ldots & 0 & 0 & 0 \\ (n-g+1)\mu & -((n-g+1)\mu+\gamma) & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ 0 & 0 & \ldots & -((n-1)\mu+\gamma) & \gamma & 0 \\ 0 & 0 & \ldots & n\mu & -(\Phi_1\gamma+n\mu) & \Phi_1\gamma \\ 0 & 0 & \ldots & 0 & (n+1)\mu & -(n+1)\mu \end{pmatrix},$$

and

$$A_1 = \begin{pmatrix} -(\gamma+(n-g)\mu) & \gamma & 0 & \ldots & 0 & 0 & 0 \\ (n-g+1)\mu & -((n-g+1)\mu+\gamma) & \gamma & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ 0 & 0 & 0 & \ldots & n\mu & -(n\mu+\Phi_1\gamma) & \Phi_1\gamma \\ 0 & 0 & 0 & \ldots & 0 & (n+1)\mu & -(n+1)\mu \end{pmatrix}.$$

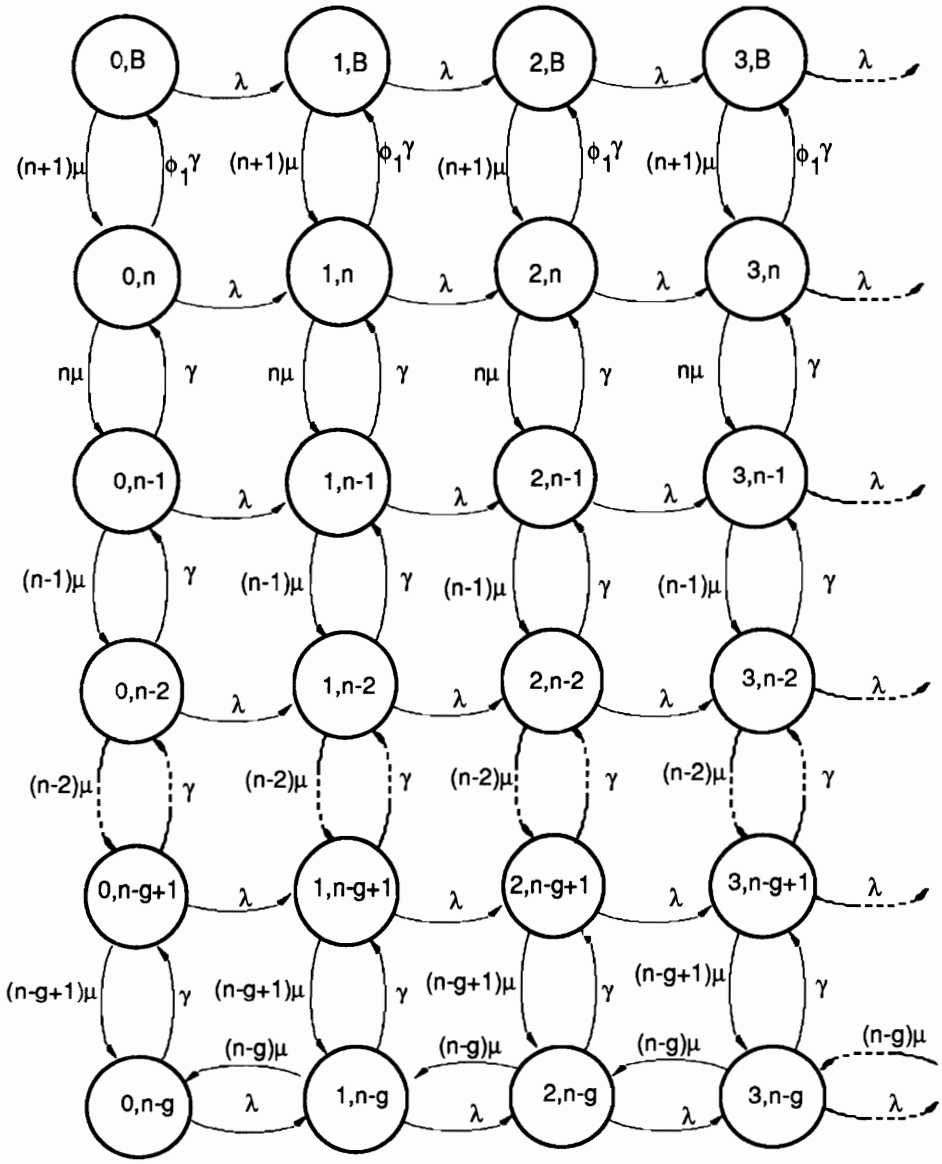

Figure 5.3.1. M/G/1 Queueing model

Using the above notation, the generator for system S_2 is given by

$$\tilde{Q}_2 = \begin{pmatrix} B_0 - \lambda I & \lambda I & 0 & 0 & 0 & 0 & \dots \\ A_2 & A_1 - \lambda I & \lambda I & 0 & 0 & 0 & \dots \\ 0 & A_2 & A_1 - \lambda I & \lambda I & 0 & \dots & \\ 0 & 0 & A_2 & A_1 - \lambda I & \lambda I & \dots & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}.$$

At this point, it should be noted that $A_1 e + A_2 e = 0$.

Let P_i', $i \geq 0$ denote the row vector of equilibrium probabilities associated with level i of system S_2. The infinite vector of equilibrium probabilities for system S_2 may then be defined as $P' = [P_0' \quad P_1' \quad P_2' \ldots]$.

Neuts [1981] has shown that a Markov chain of this type has a matrix geometric solution. Namely,

$$P_i' = P_{i-1}' R, \tag{5.3.1}$$

where R is the rate matrix.

At steady state $P'\tilde{Q}_2 = 0$, which, yields the following set of equations

$$P_0'(B_0 - \lambda I) + P_1' S^0 B^0 = 0, \tag{5.3.2}$$

$$P_0'\lambda I + P_1'(A_1 - \lambda I) + P_2' A_2 = 0 \tag{5.3.3}$$

and

$$P_{i-1}'\lambda I + P_i'(A_1 - \lambda I) + P_{i+1}' A_2 = 0 \quad for \quad i \geq 2. \tag{5.3.4}$$

Multiplying equations (5.3.2) and (5.3.3) by e yields

$$-P_0'\lambda e + P_1' S^0 = 0 \tag{5.3.5}$$

and

$$P_0'\lambda I e + P_1' A_1 e - P_1'\lambda I e + P_2' S^0 B^0 e = 0. \tag{5.3.6}$$

Substituting (5.3.5) into (5.3.6) gives

$$P_1' S^0 - P_1' S^0 B^0 e - P_1'\lambda I e + P_2' S^0 = 0,$$

which simplifies to

$$-P_1'\lambda I e + P_2' S^0 = 0. \tag{5.3.7}$$

Similarly for a general i, we establish the following expression:

$$-P_i'\lambda I e + P_{(i+1)}' S^0 = 0 \quad for \quad i \geq 1. \tag{5.3.8}$$

Now consider equation (5.3.4). Using $A_2 = S^0 B^0$, $P'_{i+1} S^0 = P'_i \lambda e$ from (5.3.8) and $e.B^0 = B^{00}$ we obtain

$$P'_{i-1} \lambda I + P'_i (A_1 - \lambda I) + P'_i \lambda I B^{00} = 0. \tag{5.3.9}$$

Using (5.3.1) in (5.3.9) results in

$$P'_{i-1} \lambda I - P'_{i-1} R[\lambda I - A_1 - \lambda B^{00}] = 0,$$

which implies

$$R = [\lambda][\lambda I - A_1 - \lambda B^{00}]^{-1}. \tag{5.3.10}$$

It is argued in Neuts[1981] that $[\lambda I - A_1 - \lambda B^{00}]$ is invertible for all irreducible stable systems.

The determination of the vector P'_0 is accomplished by using (5.3.2)

$$P'_0 (B_0 - \lambda I) + P'_1 A_2 = 0,$$

which can be written using (5.3.1) as

$$P'_0 (B_0 - \lambda I + R A_2) = 0. \tag{5.3.11}$$

The matrix $(B_0 - \lambda I + R A_2)$ is an infinitesimal generator and the vector P'_0 is proportional to its stationary probability vector. Let $\underline{P}_0$ denote the stationary vector for $(B_0 - \lambda I + R A_2)$. Then

$$P'_0 = k \underline{P}_0.$$

From (5.3.1) we find

$$\sum_{i=0}^{\infty} P'_i = P'_0 \sum_{i=0}^{\infty} R^i$$

Solving (5.3.11) we obtain a vector $\underline{P}_0$ which has the following relationship to P'_0:

$$P'_0 = k \underline{P}_0.$$

Here k is a scalar given by

$$k = \frac{1}{\underline{P}_0 [I - R]^{-1} e}.$$

The ergodic distribution of the system S_2 is then given by

$$P'_i = P'^i_0 R^i \quad for \quad i \geq 0.$$

5.3.2 Determination of a Scaling Factor

The stationary distribution of states corresponding to the busy period of S_1 are different from that of S_2 by a scaling factor δ. We now present a simple argument to explain why this is so and also derive a formula for the factor δ. To aid in describing the theory that leads to the scaling factor, we will refer to all variables for the S_1 with the subscript 1 and those of S_2 with the subscript 2. The processes S_1 and S_2 are alternating renewal processes. Each process alternates between busy and idle periods, and the time between successive entries to either a busy or idle period is called a cycle. In each process, the lengths of the busy and idle periods are each statistically independent in successive cycles. For any alternating renewal process, the ergodic probability that the process is in any state (or any set of states) may be computed by taking the ratio of the total expected amount of time spent in the state (or set of states) during a cycle to the expected cycle length. For $r = 1, 2$, denote the length of the busy period by $\tilde{t}_{r_B}$, the length of the idle period by $\tilde{t}_{r_I}$, the total amount of time spent in state (i, j) during a cycle by $\tilde{t}_{r_{i,j}}$, and the ergodic probability that the system is in state (i, j) by P_{ij} for system S_1 and $P_{ij}{}'$ for S_2. Then, for $j \geq (n - g)$:

$$P_{ij} = \frac{E[\tilde{t}_{1_{i,j}}]}{E[\tilde{t}_{1_B}] + E[\tilde{t}_{1_I}]}$$

and

$$P'_{ij} = \frac{E[\tilde{t}_{2_{i,j}}]}{E[\tilde{t}_{2_B}] + E[\tilde{t}_{2_I}]}.$$

Thus,

$$\frac{P_{ij}}{P'_{ij}} = \frac{E[\tilde{t}_{1_{i,j}}]}{E[\tilde{t}_{1_B}] + E[\tilde{t}_{1_I}]} \frac{E[\tilde{t}_{2_B}] + E[\tilde{t}_{2_I}]}{E[\tilde{t}_{2_{i,j}}]}.$$

Now, the dynamics of the processes S_1 and S_2 are statistically identical during busy periods so that $\tilde{t}_{1_B}$ and $\tilde{t}_{2_B}$ are identically distributed as are $\tilde{t}_{1_{i,j}}$ and $\tilde{t}_{2_{i,j}}$ for $j \geq (n - g)$. We then find that

$$P_{ij} = \frac{E[\tilde{t}_{2_I}] + E[\tilde{t}_{2_B}]}{E[\tilde{t}_{1_I}] + E[\tilde{t}_{2_B}]} P'_{ij} \quad \text{for} \quad j \geq (n - g). \tag{5.3.12}$$

Thus, $P_{ij} = \delta P'_{ij}$, where

$$\delta = \frac{E[\tilde{t}_{2_I}] + E[\tilde{t}_{2_B}]}{E[\tilde{t}_{1_I}] + E[\tilde{t}_{2_B}]}. \tag{5.3.13}$$

Computational formulae for the elements of (5.3.13) are now obtained. Unlike the treatment in Daigle and Jain[1992] the theory presented here does not require the inversion of a matrix of dimension n. Thus the approach presented here is computationally efficient for systems with large n.

In order to evaluate (5.3.13), we determine the expected length of cycle of the system S_1 and S_2. The idle period of the system S_1 can be found by inverting a matrix of the order n. However, n maybe large in a typical cellular system. In the following paragraphs we present a theory that will enable the determination of the idle time by inverting a matrix of order g.

The distributions of the idle time of the systems S_1 and S_2 are the time to absorption of the transient Markov chain described in figures 5.3.2 and 5.3.3, respectively (system S_3 and S_4). The distribution of the idle time of S_1 is the time to absorption of the into level 1 of the process S_3, given that the process S_3 started in state $(0, n-g)$. Similarly, the idle time of S_2 is the time to absorption of process S_4 into level 1 given that the process S_4 started in state $(0, n-g)$. Furthermore, we state the following theorem that will make it possible to calculate the expected length of the idle time of S_1 using the process S_2.

Theorem 1. *The distribution of time spent in state $(0, n-g)$ in S_3 and S_4 in a cycle is identical.*

Proof : Let S_3^+ refer to the set of states (h, j) for $j \geq n-g$ and $h \geq 0$ in S_3. Let S_3^- refer to the states $(0, j)$ for $i < n-q$ in S_3. Let the process S_3 make $\tilde{n}$ transitions to the states S_3^- before making the first transition to S_3^+. Let $\tilde{x}_i$ be the time spent in state $(0, n-g)$. Then the total time spent in $(0, n-g)$ before the first transition to S_3^+ will be given by

$$\tilde{x} = \sum_{k=1}^{\tilde{n}} \tilde{x}_k.$$

The random variables $\tilde{x}_i$ for $i = 1, 2, \ldots \tilde{n}$ are independent, identically identically distributed exponential random variables with parameters $(\lambda + \gamma + (n-g)\mu)$, and their Laplace-Stieltjes transform is given by

$$F^*_{\tilde{x}_i}(s) = \frac{(\lambda + \gamma + (n-g)\mu)}{s - (\lambda + \gamma + (n-g)\mu)}.$$

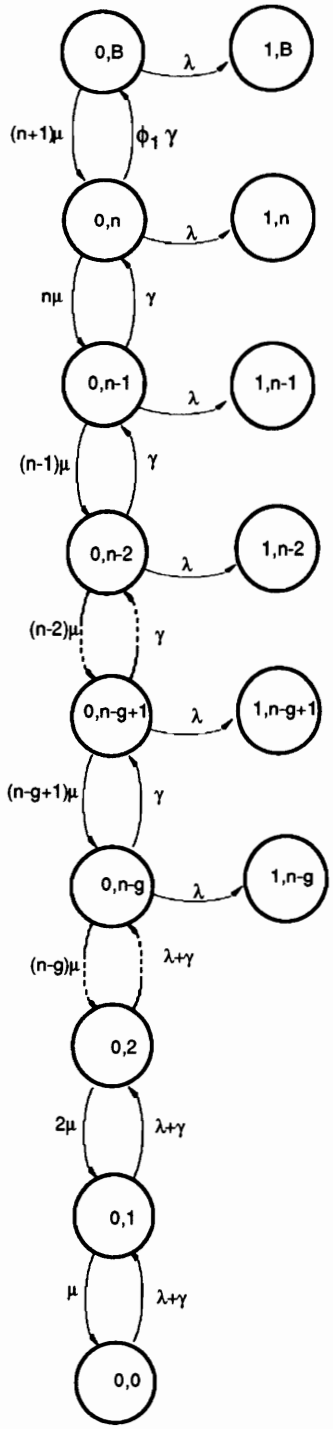

Figure 5.3.2. State diagram for S_3, the system used to model the idle period dynamics of S_1

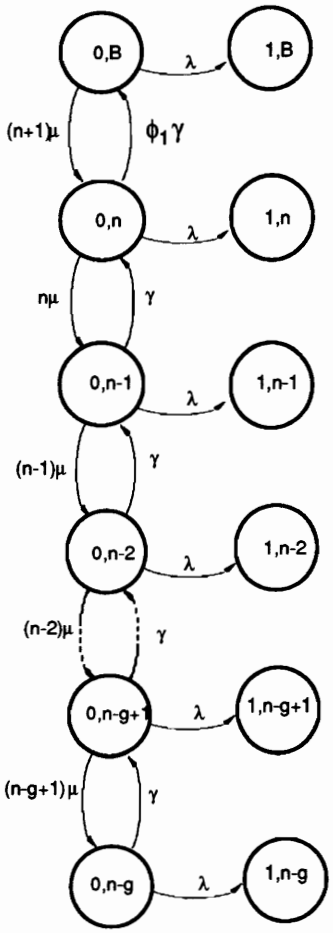

Figure 5.3.3. State Diagram for S_4, the system used to model the dynamics of the idle period of S_2

The probability of a transition to states S_3^- and $S_3^+$from $(0, n-g)$ is given by $p = \frac{(n-g)\mu}{\lambda+\gamma+(n-g)\mu}$ and let $q = 1-p = \frac{\lambda+\gamma}{\lambda+\gamma+(n-g)\mu}$. Then

$$F_x^*(s) = \sum_{n=1}^{\infty} (F_x^*(s))^n p^{(n-1)} q,$$

which when simplified yields

$$F_X^*(s) = \frac{\lambda+\gamma}{s-\lambda+\gamma}. \tag{5.14}$$

Equation (5.14) is a Laplace-Stieltjes Transform for the exponential random variable with rate $(\lambda+\gamma)$. This implies that the distribution of the time spent in state $(0, n-g)$ before making the

first transition to S_3^+ is exponential with rate $(\lambda+\gamma)$ and is the same as the time spent in $(0, n-g)$ in S_4 before making a transition to S_4^+.

Corollary 1. *The distribution of the time spent in phases j for $j \geq (n-g)$ is identical in S_3 and S_4.*

We now set up the expressions to determine the expected amount of time spent in the transient states of the processes S_3 and S_4.

Define $(g+b+1)$ square matrices T_4 and T_4^0 as

$$T_4 = \begin{pmatrix} -(\gamma+\lambda) & \gamma & 0 & \dots & 0 & 0 & 0 \\ (n-g+1)\mu & -((n-g+1)\mu+\lambda+\gamma) & \gamma & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & n\mu & -(n\mu+\phi_1\gamma) & \phi_1\gamma \\ 0 & 0 & 0 & \dots & 0 & (n+1)\mu & -((n+1)\mu+\lambda) \end{pmatrix}$$

and

$$T_4^0 = diag(\lambda, \lambda, \dots, \lambda).$$

The infinitesimal generator of the chain S_4 is given by

$$\tilde{Q}_4 = \begin{pmatrix} T_4 & T_4^0 \\ 0 & 0 \end{pmatrix}.$$

Let $P_{i_4}(t) = [P_{i_{4T}}(t), \quad P_{i_{4A}}(t)]$, be the vector of the probability masses for the states shown in Figure 5.3.3. The subscript T refers to the transient states and the subscript A refers to the absorbing states. The differential equation for the Markov chain described in S_4 can be written as

$$\frac{d}{dt}P_{i_4}(t) = [P_{i_{4T}}(t), \quad P_{i_{4A}}(t)] \begin{pmatrix} T_4 & T_4^0 \\ 0 & 0 \end{pmatrix}. \tag{5.3.15}$$

Define P_{T_0} as the $(g+1)$ vector $[1, 0, \dots, 0]$. Then, the solution to the system (5.3.15) is given by

$$P_{i_{4T}}(t) = P_{T_0} e^{T_4 t}, \tag{5.3.16}$$

and

$$P_{i_{4A}}(t) = P_{T_0}(e^{T_4 t} - I)T_4^{-1}T_4^0. \tag{5.3.17}$$

It can be verified that (5.3.16) and (5.3.17) are a solution by substituting them into (5.3.15). Let $\tilde{t}_{4i}$ represent the time spent in states $(0,i)$ for $i \geq (n-g)$. The expected total time spent in each transient state before absorption is given by the vector

$$[\bar{t}_{4(n-g)}, \bar{t}_{4(n-g+1)}, \ldots, \bar{t}_{4n}] = -1P_{T_0}T_4^{-1}. \tag{5.3.18}$$

The average time to absorption is given by

$$E[\tilde{t}_{4_I}] = E[\tilde{t}_{4T}] = -1P_{T_0}T_4^{-1}e \tag{5.3.19}$$

We now turn to the task of finding the total time spent in the transient states of S_3 before absorption. To this end it is observed that the ergodic probability of being in state $(0, n-g)$ is related to the probability of being in the states of S_3^-. The first element of P_{0B} is $P_{0,n-g}$ from which $P_{0,j}$ for $j = 0, 1, \ldots, n-g-1$ can be determined. For example, one can compute these probabilities as follows:

$$\begin{aligned} P_{0,n-g-1} &= \frac{(n-g)\mu}{\lambda+\gamma} P_{0,n-g} \\ P_{0,n-g-2} &= \frac{(n-g-1)\mu}{\lambda+\gamma} P_{0,n-g-1} \\ &\ldots \\ P_{0,0} &= \frac{\mu}{\lambda+\gamma} P_{0,1}. \end{aligned} \tag{5.3.20}$$

Since the ergodic probabilities are directly proportional to the expected amount of time spent in the state in a cycle, the expected time spent in the states $(0,j)$ for $j = 0, 1, \ldots, n-g-1$ can be obtained from the time spent in the state $(0, n-g)$. For example one can compute these times as follows:

$$\begin{aligned} E[\tilde{t}_{3_{0,n-g-1}}] &= \frac{(n-g)\mu}{\lambda+\gamma} E[\tilde{t}_{4_{0,n-g}}] \\ E[\tilde{t}_{3_{0,n-g-2}}] &= \frac{(n-g-1)\mu}{\lambda+\gamma} E[\tilde{t}_{3_{0,n-g-1}}] \\ &\ldots \\ E[\tilde{t}_{3_{0,0}}] &= \frac{\mu}{\lambda+\gamma} E[\tilde{t}_{3_{0,1}}]. \end{aligned} \tag{5.3.21}$$

Let the total idle time be given by $\tilde{t}_{3_I}$. Then the expected idle time per cycle is given by

$$\begin{aligned} E[\tilde{t}_{3_I}] &= E[\tilde{t}_{3_0}] + E[\tilde{t}_{3_1}] + \ldots + E[\tilde{t}_{4_{n-g}}] + \ldots + E[\tilde{t}_{4_n}] \\ &= E[\tilde{t}_{3_0}] + E[\tilde{t}_{3_1}] + \ldots + E[\tilde{t}_{4_I}]. \end{aligned}$$

As pointed out above, the expected lengths of the busy periods in processes S_1 and S_2 are equal. The expected length of the busy period ($\tilde{t}_{4_B}$)in chain S_2 can now be calculated using the relation

$$P_0'.\mathbf{e} \quad = \quad \frac{\text{idle period}}{\text{idle period} + \text{busy period}},$$

which yields

$$E[\tilde{t}_{4B}] = \frac{(1 - P_0 e)E[\tilde{t}_{4_I}]}{P_0' e}.$$

We now have all the quantities need to estimate the multiplier δ. There is a simple physical interpretation of δ. It represent the probability of not being in one of the (i, j) states for $(n - g) < j$ of the chain S_1.

The probability distribution of states in the system S_1 enable useful performance measures of the cellular telephone system to be determined. For example, the expected queue length ($E[\tilde{q}_A]$) for the queued customers can be obtained by using the expression

$$E[\tilde{q}_A] = \delta P_0' R(I - R)^2. \tag{5.3.22}$$

Using Little's result the average weighting time in the queue is given by

$$D_A = \frac{E[\tilde{q}_A]}{\lambda}.$$

The probability that an arriving higher priority handoff call will be denied service can be calculated as follows. Let P_B be a $(g + b + 1)$ vector and P_{B_j} for $j = 0, 1, \ldots, g + b$ be the $(j + 1)^{th}$ value. Then

$$P_B = \delta P_0'(I - R)^{-1},$$

and the call dropping probability is given by

$$P_D = \sum_{i=g}^{g+b-1} P_{B_i}(1 - \phi_{i-g+1}) + P_{B_{g+b}}.$$

5.4 BORROWING PROBABILITIES

The common pool of channels sees only overflow traffic created when there are no available servers in the cells. This traffic is peaked i.e.; the arrival process has a higher variance than mean.

Overflow traffic has been studied extensively in the literature since the early seventies (Kuczura [1973], Kuczura [1977], Fredericks [1979], Akimaru and Takahashi [1983] and Meier-Hellstern [1989]). It has been found (Kuczura[1973]) that a good approximation for overflow processes is an Interrupted Poisson Process (IPP). An IPP is characterized by three parameters κ, χ and τ. This process turns on for an exponential amount of time with mean $\frac{1}{\tau}$ and turns off for another independent exponentially distributed time with mean $\frac{1}{\chi}$. When the process is on, arrivals occur at Possion rate κ. In our system of K identical cells the overflow process from each cell will be modeled as an IPP. We then combine these overflow streams from the K cells and model the dynamics of the common pool using a Markov Modulated Poisson Process. These ideas are explained below.

Consider a system with K cells. Let b_j, for $j = 1, 2, \ldots, K$, be the maximum number of channels that cell j can borrow. Let ϕ_{i_j} be the borrowing probability for the i^{th} channel in the cell j. Let n_c be the number of channels in the common pool therefore $b_j \leq n_c$ for $j = 1, 2, \ldots, K$. Let $\tilde{b}_j$ be the random number of channels that are borrowed by cell j at any time t, if every request for a borrowed channel that it makes is fulfilled. Further, let $\bar{b}_j = \mathrm{E}(\tilde{b}_j)$ and $\sigma^2 = \mathrm{var}(\tilde{b}_j)$, then

$$\bar{b}_j = \sum_{k=g+1}^{g+b_j} k P_{B_k},$$

and

$$\sigma^2 = \sum_{k=g+1}^{g+b_j} k^2 P_{B_k} - \bar{b}_j^2.$$

The peakedness ratio z, as defined in (Kuczura [1973]), is the ratio of the variance to the mean and is given as

$$z = \frac{\sigma^2}{\bar{b}_j}.$$

Using the techniques described in Kuczura [1972] on page 442, these values of $\bar{b}_j$ and z are used to find the parameters for the IPP, and κ, τ and χ are calculated as follows:

$$\kappa = \bar{b}_j z + 3z(z-1),$$

$$\chi = \frac{\bar{b}_j}{\kappa}\left(\frac{\kappa - \bar{b}_j}{z-1} - 1\right),$$

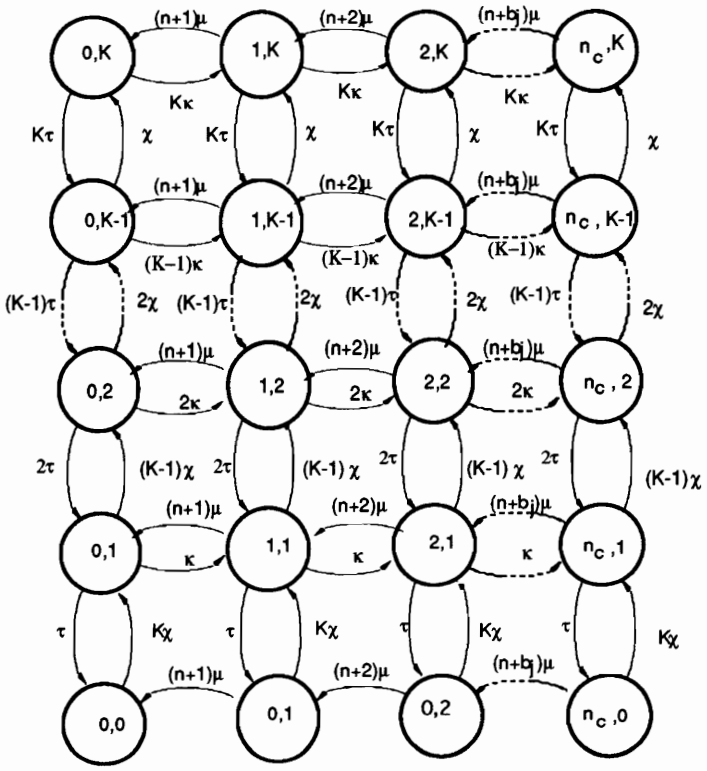

Figure 5.4.1. State diagram of the MMPP to model the common pool for $b_j = n_c$

and

$$\tau = (\frac{\kappa}{b_j} - 1)\chi.$$

Since we have a system with K cells, K of these IPP are applied to the common pool. The dynamics of the common pool can be modeled by a Markov Modulated Poisson Process (MMPP). The state diagram for the MMPP for the case when $b_j = n_c$ is shown in Figure 5.4.1.

Define the number of channels that are in the *on* state at time t as $\tilde{n}_{on}(t)$ and the number of channels of the common pool in use at time t as $\tilde{n}_{cp}(t)$. Then the system is defined by the process $\{\tilde{n}_{cp}(t), \tilde{n}_{on}(t), t \geq 0\}$. Let the the probabilities that the process in state i, j for $i = 0, 1, \ldots, n_c$ and $j = 0, 1, \ldots, K$, at time t be given by $P\{\tilde{n}_{cp}(t) = i, \tilde{n}_{on}(t) = j\}$ and the corresponding equilibrium probabilities are defined by $P\{n_{cp} = i, n_{on} = j\}$. We also define a $(K+1) \times (n_c+1)$-vector π such

that

$$\pi_{(K+1)i+j} = P\{n_{cp} = i, n_{on} = j\}.$$

We now determine the rate of transitions from the state of the type i, j to $i-1, j$, for $n_c \geq i > 0$ and $j = 0, 1, \ldots, K$. In our explanation of the channel borrowing process we had stated that the borrowed channel is returned to the common pool by the cell as soon as one becomes available. In order to do so the cell may have to reassign calls to a different set of channels to free the borrowed channel. Now, if a cell has a borrowed channels, it is in the state of the form k, l (Figure 5.2.1) for $k = 0, 1, \ldots$ and $l > n$, since the borrowed channel is returned to common pool as soon as the cell makes a transition to the state of the form k, l for $k \geq 0$ and $l = n$ from l, k for $l \geq 0$ and $k > n$. In the states k, l for $k \geq 0$ and $l > n$, channels are becoming free at a Possion rate greater than $(n+1)\mu$. In the worst case, the borrowed channels become free at the slowest rate when they are borrowed from a single cell. In the case g channels are borrowed from a single cell then they become free at Poisson rate given by $(n+g)\mu$ for $g = 1, 2, \ldots, b_j$. If $n_c > b_j$ then beyond b_j the channels have to be borrowed from more than one cell, hence the channels will become free at Possion rate $2(n+1)+g-1$ for $g = 1, 2, \ldots, b_j$, and $b_j + g \leq n_c$. Thus the $(hb_j + g)^{th}$ channel in the common pool becomes free at a minimum of Possion rate $(h(n+1)+g-1)\mu$ for $g = 1, 2, \ldots, b_j$, $hb_j + g \leq n_c$ and $h = 1, \ldots, \arg\min_h\{hb_j \geq n_c\}$.

The process that governs the transitions along j for $K+1 \geq j \geq 0$ is called a *phase process* (Daigle [1992]). We will write down the generator of the phase process, a $K+1$ square matrix $\tilde{Q}_{PH}$, as

$$\tilde{Q}_{PH} = \begin{bmatrix} -K\chi & K\chi & 0 & \ldots & 0 & 0 \\ \tau & -((K-1)\chi+\tau) & (K-1)\chi & \ldots & 0 & 0 \\ 0 & 2\tau & -((K-2)\chi+2\tau) & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ldots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & K\tau & -K\tau \end{bmatrix}.$$

Let Λ, a $K+1$-square matrix be defined as, $\Lambda =$ diag $(0, \kappa, 2\kappa, \ldots, K\kappa)$ and n_c, $K+1$-square matrices as $M_{(hb_j+g)} =$ diag $((h(n+1)+g-1)\mu, (h(n+1)+g-1)\mu, (h(n+1)+g-1)\mu, \ldots, (h(n+1)+g-1)\mu)$ for $g = 1, 2, \ldots, b_j$, $hb_j + g \leq n_c$ and $h = 1, \ldots, \arg\min_h\{hb_j \geq n_c\}$.

Using these initial definitions, we can define the $(K+1)$ by n_c infinitesimal generator, $\tilde{Q}_{MMPP}$, of the chain shown in Figure 5.4.1 as

$$\begin{bmatrix} -(\Lambda - Q_{PH}) & \Lambda & 0 & \dots & 0 & 0 \\ M_1 & -(\Lambda - Q_{PH} + M_1) & \Lambda & \dots & 0 & 0 \\ 0 & M_2 & -(\Lambda - Q_{PH} + M_2) & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & M_{n_c} & -(M_{n_c} - Q_{PH}) \end{bmatrix}.$$

The chain's equilibrium stochastic probabilities are calculated by solving the system of equations

$$\pi \tilde{Q}_{MMPP} = 0,$$

and

$$\pi e = 1.$$

The probability that a request for an extra channel will be denied $(P_b)$is given by the ratio of the expected number of requests that are blocked to the expected number of requests. Therefore, P_b

$$P_b = \frac{\sum_{j=0}^{j=K+1} j\kappa \ \pi_{(K+1)n_c+j}}{\sum_{i=0,j=0}^{i=n_c,j=K+1} j\kappa \ \pi_{(K+1)i+j}}.$$

A conservative estimate for ϕ_{i_j}, for $i = 1, \dots, b_j$ and $j = 1, \dots, K$, is $1 - P_b$.

Consider a cluster of seven similar cells with $n = 38$, $g = 1$ and $\lambda = 22$, $\gamma = 15$, $\mu = 1$. Let this cluster have a common pool containing three channels. Let each cell be allowed to borrow all three channels i.e.; $b_j = 3$. Then the parameters of the IPP are calculated to be $\kappa = 2.8463$, $\chi = 0.235824$ and $\tau = 3.597313$. Using these values and modeling the common pool as an MMPP the value of ϕ is calculated to be equal to 0.9989. Using this ϕ, the call dropping probability is found to be 0.01141 and the average delay equal to 1.695 seconds. If there was no borrowing of channels then these values are 0.149 and 0.139 respectively.

5.5. OTHER RESULTS

The probability distribution of states in the system of Figure 5.2.1 enable performance measures of interest to be determined and plotted. These results are plotted in Figures 5.5.1 and 5.5.2. In

Figure 5.5.1, we plot the average queue length as a function of the number of guard channels. It is found that as the number of guard channels increases, the average number of new calls in the queue increases. This is because an increase in the number of guard channels increases the service time of the system shown in Figure 4.4.2, while the arrival rate to the M/G/1 queue is still λ. Therefore, from the M/G/1 queue's perspective, the service rate increases, keeping the arrival rate constant and hence, the average number in the queue increases. In Figure 5.5.2 we plot the value of the average delays and call dropping probabilities as a function of the number of guard channels. It is found that as the number of guard channels is increased, the average delay increases and the call dropping probability falls.

5.6. CONCLUSIONS

In this chapter we have developed a queueing model with reserved and borrowable servers. This model will be used in Chapter 6 to efficiently construct designs of a cell in a cellular telephone system. The analytical model is computationally stable and is easy to implement for large systems. This will be of value in the algorithm to optimize a cell because, we will be dealing with large values of n and a repetitive solution of the model described in this chapter is necessary.

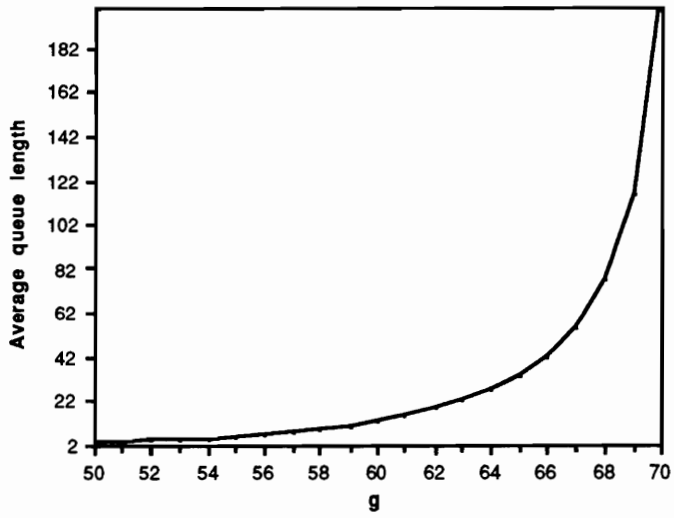

Figure 5.5.1. Expected queue length for $n = 416$, $\lambda = 50$ and $\gamma = 300$

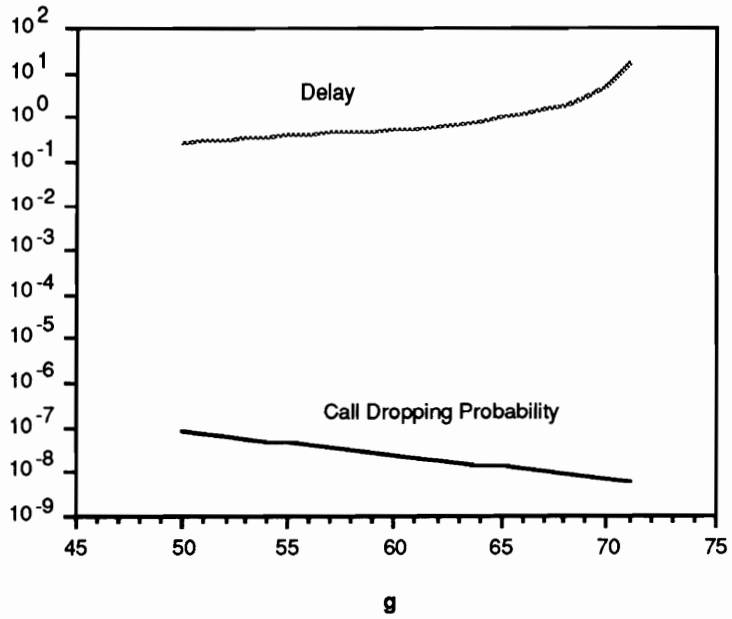

Figure 5.5.2. Delay and call dropping probability for $n = 416$, $\lambda = 50$ and $\gamma = 300$

Chapter 6.

Cell Design

6.1. INTRODUCTION

The cell design problem is identified and described in Chapter 3. In this chapter using the analytical tools developed in Chapters 4 and 5, we present a technique for solving this problem. Our objective in this chapter is two fold: to understand the operational behavior of a cell under the use of guard channels and Hybrid Channel Allocation and to use this information to develop an algorithm to determine the minimum number of nominal channels required by the cell.

We begin by discussing some important properties of a cell in Section 6.2. In Section 6.3 we present results from a SLAM II simulation of a cluster with three identical cells. In Section 6.4 we present cell optimization, in Section 6.5, a numerical example, and in Section 6.6 conclusions are presented.

6.2 PECULIARITIES OF CELL BEHAVIOR

The development of the queueing model for the cell and the subsequent analysis has provided a basic understanding of the cell. For example, the idea that the service time is the time to absorption of a Markov chain, provides valuable insights into the queueing behavior of the cell. Two very useful properties of the average delays and the blocking probabilities are described below.

In Figures 6.2.1 and 6.2.2, we plot the behavior of call dropping probabilities and expected delay as functions of guard channels, g, and the total number of channels, n. Based on the figures and the following discussion we state properties 6.2.1 and 6.2.2.

Property 6.2.1. *If g and all other parameters except n are kept constant then the expected delay and the call dropping probabilities both decrease when n is increased.*

This property can be established by the following argument. Assume that a cell has n channels and g guard channels. Now, suppose we have an extra channel called the *special channel*, that we assign at our discretion to service customers. Let us consider the following two cases.

Case I

We only use the special channel to service calls in the queue. As soon as the queue length is reduced to zero this server is withdrawn. This implies that whenever there is a queue this server reduces the queue. If we assume that nothing in the original system has changed except the availability of the new server, then the average waiting time will decrease. This is because the waiting time of the customers served by the special server decreases. Thus, if the system has any queueing at all, then the average waiting time will decrease as a result of providing the special server. Now consider case II.

Case II

Assume the original system with n total channels and g guard channels remains unchanged (like at the beginning of Case I). Now assume that the special channel is only used to service those calls that are dropped by our original system. In this case, the number of calls that will be dropped will decrease. Since some of the calls dropped by our system will be serviced by the special server. Given that the total number of calls that arrive to the system has not changed, the call dropping probability will decrease.

The above two cases represent extremes. In general, the permanent addition of our new channel to the cell will cause a channel from among the $n + 1$ that the cell has, to service calls sometimes according to Case I and other times according to Case II. We can therefore conclude that both the call dropping probability and average delay will decrease when an extra channel is added to the cell.

Property 6.2.2. *If n and all other parameters except g are held constant then as g is increased, the dropping probability decreases, and expected delay increases.*

The effect of increasing g while keeping n constant is equivalent to transferring a channel that was previously shared between new and handoff calls to being used exclusively by handoff calls only. The amount of work brought in by the new calls is constant because they are all being served in either case. However, there are fewer channels for the new calls. Since the same amount of work

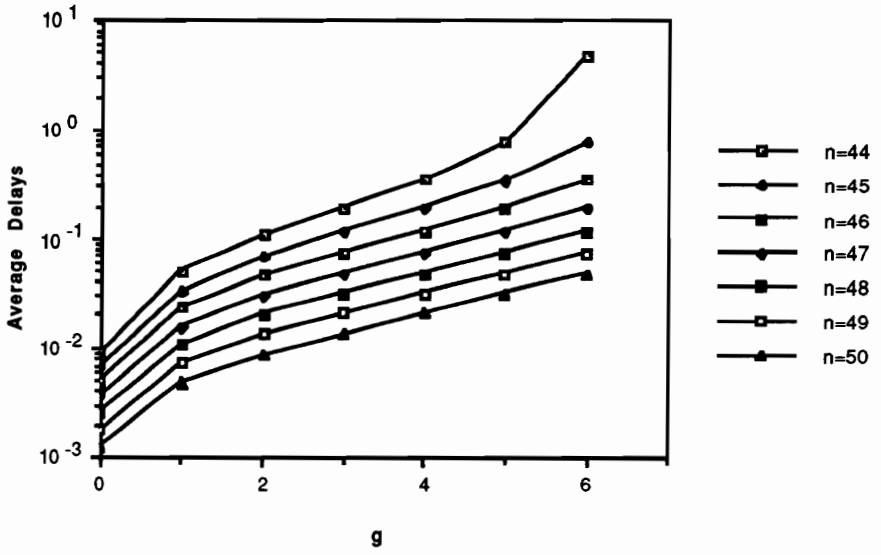

Figure 6.2.1 Expected delays for $\lambda = 30$, $\gamma = 8$, $\mu = 1$

must be done by fewer servers, longer average queue lengths and, therefore, longer delays result. From the point of view of handoff calls, all the channels are available as before, except this time more channels are exclusively dedicated to them. Thus the probability that a handoff call arrives and finds all the channels occupied decreases.

Another explanation for the average delay increase is as follows. Since the number of channels available to serve the new calls decreases, the probability of forming a queue increases. Furthermore, increasing g has the effect of adding another transient state to the chain shown in Figure 2.4.3. The addition of this new state increases the total time spent in the transient states by the amount of time spent in the new state. This translates to increased service time and therefore longer queues. Longer queues lead to longer delays, hence the property.

These trends in the expected delays and dropping probabilities are shown in Figures 6.2.1 and 6.2.2.

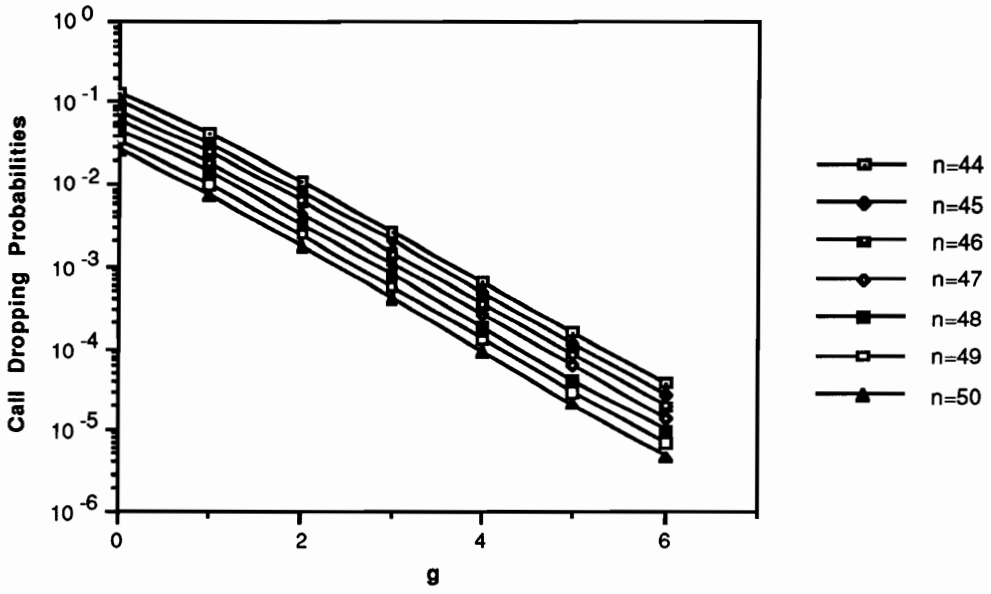

Figure 6.2.2 Call Dropping Probabilities for $\lambda = 30$, $\gamma = 8$, $\mu = 1$

The above two properties are important in explaining the dynamics of a cell when its operational environment changes. Most of these changes can be thought of as a change in the number n or g, or both, and the result of the change can be predicted.

Consider a cell operating in a system that uses Hybrid Channel Assignment. In such a system, a common pool is shared by a large number of cells. A cell may borrow a channel from the pool to service a handoff call, if it does not have an available channel. The borrowed channel is returned to the pool as soon as a call completes in the cell. In order to return the borrowed channel, the cell may have to reassign calls to make the borrowed channel free. Let the maximum number of cells that a cell can borrow be given by b, and the probability of borrowing the first one be given by ϕ_1, the second one by ϕ_2, and so on. Consider a case when $b = 2$. Figure 6.2.3 describes the effect of changes in the borrowing probability on blocking probability and throughput. These trends can be

explained using properties 6.2.1 and 6.2.2.

Borrowing a channel represents both an increase in the number of n and g. Therefore, blocking probability and delay both decrease due to an increase in n. However, due to a concurrent increase in g, delays may increase and the blocking probabilities may decrease more than they would from a simple increase in n.

Providing the cell with an ability to borrow channels increases its throughput. This represents added revenue for the cellular telephone company. Also, being able to borrow channels decreases the call dropping probabilities, which is also desired. The throughput behavior of the cell is shown in Figure 6.2.4.

6.3 CELL BEHAVIOR IN A CLUSTER WITH HYBRID CHANNEL ALLOCATION

In the previous section and in Chapter 5, we have modeled the behavior of the cell when the borrowing probabilities are known. In this section, we will use a simulation experiment to study a cluster of cells with a common pool. Our objective is to understand the relationship between the number of channels in the common pool and performance of each cell in the cluster. The design of the simulation is described by Figure 6.3.1.

The system consists of three identical cells having a common pool of channels from which any cell can borrow a channel. The common pool contains a total of n_c channels. Each cell has n nominal channels and g guard channels. New calls arrive at all three cells at a Poisson rate λ and the handoff calls arrive at Poisson rate γ. Channels are held for an exponentially distributed time with mean 1 sec. Our objective is to study the effect of changing the values of n_c while keeping n and g constant, on call dropping probabilities and average delays.

The design of the simulation experiments was done by first running short experimental simulation runs. These were used to determine the length of the run that would be sufficient to provide stable results. Then 16 simulation runs using 8 pairs of synchronized antithetic seeds were made and the responses obtained were averaged. The each simulation run ran for 500,000 entities.

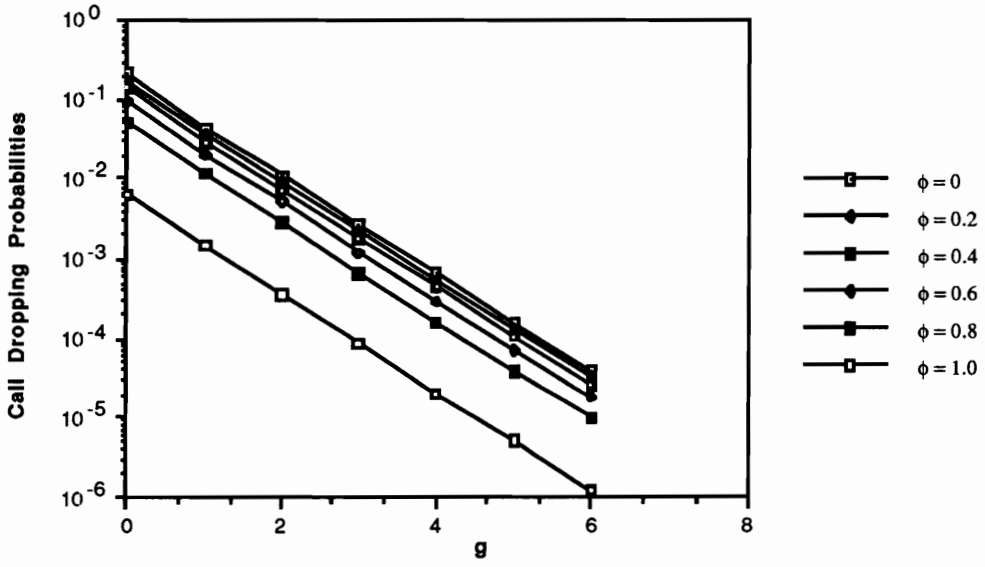

Figure 6.2.3 Call dropping probability curves for $n = 44$, $\lambda = 30$ and $\gamma = 8$ with channel borrowing with $b = 2$ and $\phi_1 = \phi_2$

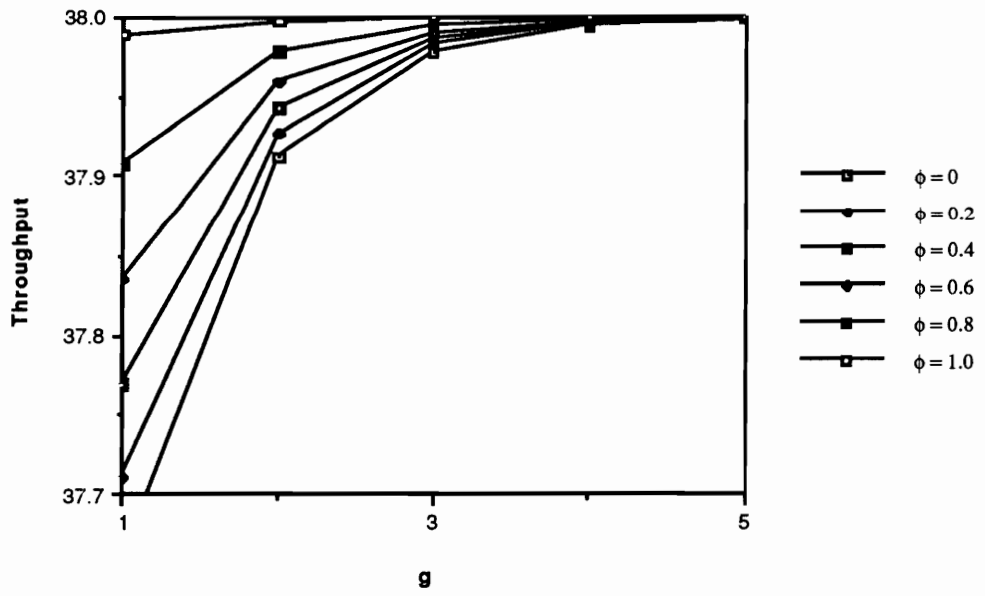

Figure 6.2.4 Call throughput curves for $n = 44$, $\lambda = 30$ and $\gamma = 8$ with channel borrowing

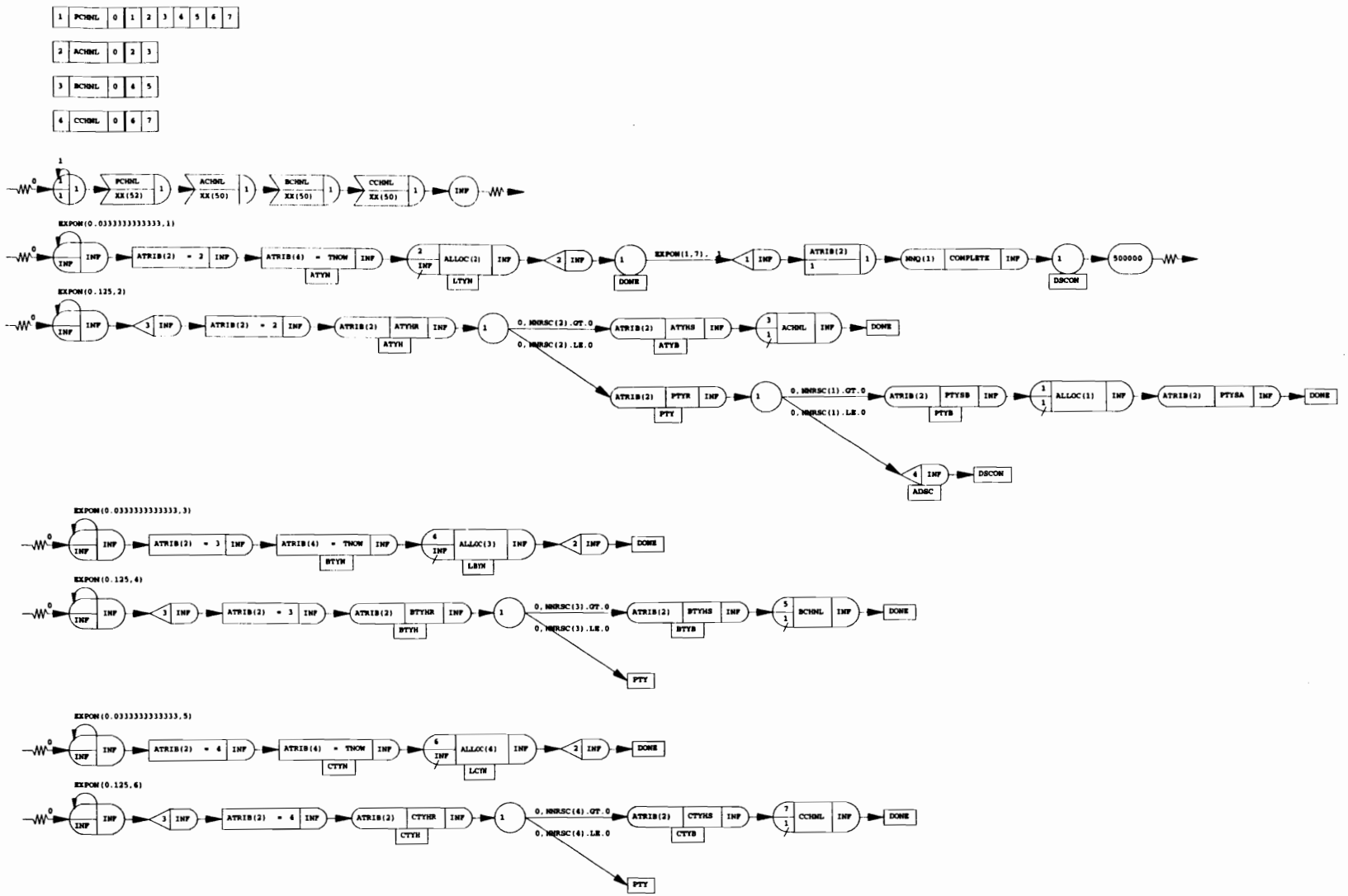

Figure 6.3.1 SLAM II Network diagram of the simulation for a three cell system.

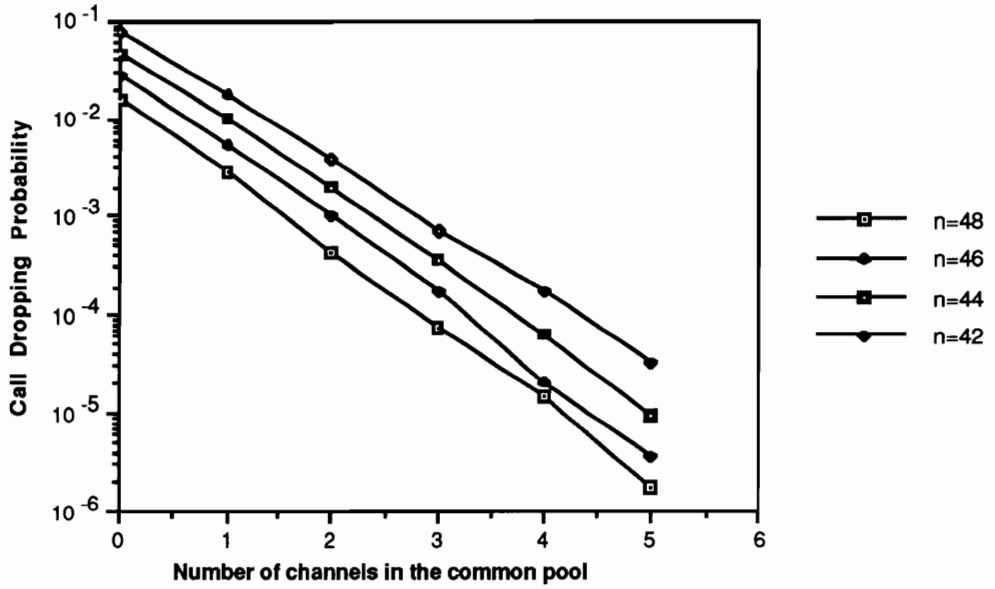

Figure 6.3.2 Cluster performance with $\lambda = 30$, $\gamma = 8$, $\mu = 1$ and $g = 1$

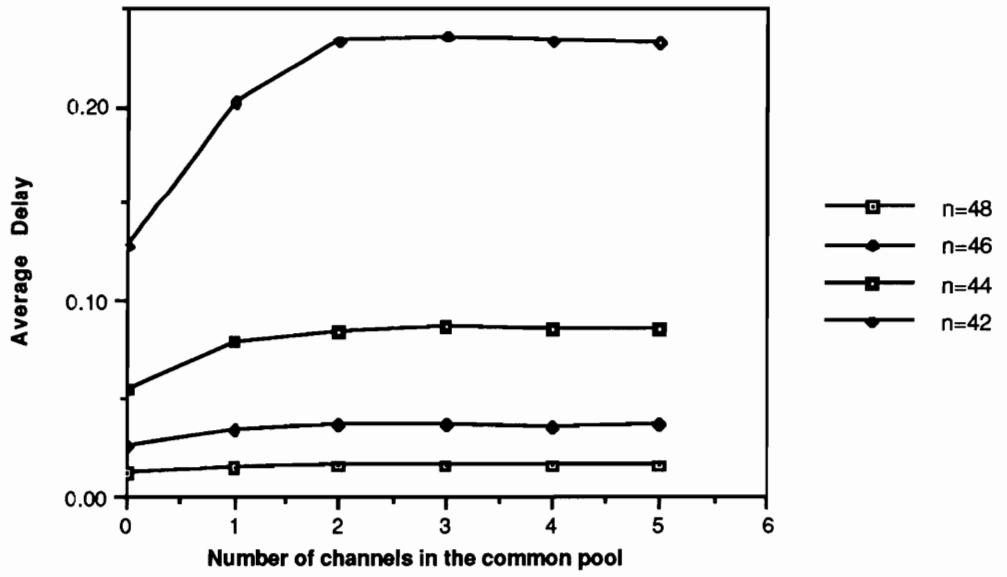

Figure 6.3.3 Cluster performance with $\lambda = 30$, $\gamma = 8$, $\mu = 1$ and $g = 1$

The results are plotted in Figures 6.3.2 and 6.3.3. It was found that as the number of channels in the common pool were increased, the call dropping probability decreases. However, the average delays increased rapidly initially and then the rate of increase decreased. A possible explanation is that as the number of channels in the common pool are increased, more of the calls that would have been dropped find a channel. Thus all the channels that were free because of dropped calls now begin to be utilized. This takes away channel resources from new calls and therefore increases their delay. It can be conjectured that the net effect of having an extra channel in the common pool amounts to simultaneously increasing the number of guard channels and the total channels of a cell. The call dropping probability decreases due to an increase in n and an increase in g. The average delay decreases due to an increase in n and increases due to an increase in g.

6.4 OPTIMIZATION OF CELLS

We began this chapter by stating that the ultimate objective of the study of the dynamics of a cell is to be able to promote spectrally-efficient cell designs. The objective of this section is to present an optimization algorithm to minimize the number of channels required by a cell. Our aim is to design a scheme that optimizes a cell using minimal computational effort.

6.4.1. Definition of the optimization problem

Let a cell consist of some n nominal channels and g guard channels. The cell also has access to a common pool of channels from which it can borrow channels. The maximum number of channels that a cell can borrow is denoted by b. The probability with which it can borrow these channels is defined in Chapter 5 and is denoted by ϕ_i for $i = 1, 2 \ldots, b$. As discussed earlier, the call dropping probability (P_D) and the average delay (D_A) are functions of the total number of channels (n), the number of guard channels (g), and the probabilities ϕ_i for $i = 1, \ldots, b$. Let the maximum permissible call dropping probability be given by M_P and the maximum allowed average delay be given by M_D.

Then the cell optimization problem can be written as:

$$\text{Minimize } n \qquad OPT6.1$$

$$\text{Subject to :}$$

$$P_D(n, g) \leq M_P$$

$$D_A(n, g) \leq M_D$$

$$0 \leq g \leq n; \quad n, g : \text{ integer.}$$

6.4.4 Solution Technique

The basic idea behind the solution technique is to first find a feasible solution to the problem (OPT 6.1). This constitutes Phase I of the proposed algorithm, and its role is to quickly provide a good upper bound on the value of n. In Phase II of the algorithm, we find a true optimal solution by decreasing n and using the monotonicity properties 6.2.1 and 6.2.2. Note that the algorithm could have used Phase II itself by first incrementing n to a sufficiently large value, and adjusting g accordingly, until a feasible solution was found, and then decreasing n to an optimal value as recommended by Phase II. However, the use of Phase I essentially achieves this by providing a quick and good upper bound on the value of n.

The Phase I algorithm works by first initializing the values of n and g such that the delay constraint is not violated. The algorithm then proceeds to the main part of Phase I. Here the value of g is increased using a logarithmic extrapolation such that the delay constraint is not violated, then the value of n is increased using a linear approximation. If the resulting value of n, along with g, satisfies the call dropping probability constraint, then by property 6.2.1, a feasible solution to

Problem OPT6.2.1 has been found, and the algorithm proceeds to Phase II. Otherwise, the main part of Phase I is repeated. The convergence of Phase I is guaranteed by the property 6.2.1.

To aid in the discussion we will define the following two functions $h_1(n,g)$ and $h_2(n,g)$ as

$$h_1(n,g) = \text{maximum } \{0, P_D(n,g) - M_P\}$$

and

$$h_2(n,g) = \text{maximum}\{0, D_A(n,g) - M_D\}.$$

The queueing system is deemed to be stable if the expected rate of arrival to the M/G/1 queueing system discussed in Chapters 4 and 5, is less than the expected rate of service.

Phase I

Initialization

Set iteration counter, $k = 0$, $g = g_0 = 0$ and $b = 0$ and select n_a by solving an M/M/n_a model and finding n_a such that the delay requirements are fulfilled. Increase b to its required value. Holding $g = 0$, increase n_a if required until the delay requirements are fulfilled. Let $n_0 = n_a$ and proceed to the main part.

Main Part

1. Calculate $h_2(n_k, 1)$. If the queueing system is unstable, or if $h_2(n_k, 1) > 0$, then set $g_{k+1} = 0$ and go to step 3. Else, calculate $h_2(n_k, 2)$. If the queueing system is unstable, or if $h_2(n_k, 2) > 0$, then set $g_{k+1} = 1$ and go to step 3. Else, $g_{k+1} = \lfloor 2 + \frac{log(M_D) - log(D_A(n_k,2))}{log(D_A(n_k,2)) - log(D_A(n_k,1))} \rfloor$ and go to step 2.
2. If $h_2(n_k, g_{k+1}) > 0$, decrease g_{k+1} until $h_2(n_k, g_{k+1}) = 0$.
3. Calculate $P_D(n_k, g_{k+1})$. If $P_D(n_k, g_{k+1}) \leq M_P$, then stop; (n_k, g_{k+1}) is a feasible solution to OPT6.1. Else, calculate $P_D(n_k+1, g_{k+1})$, and if $P_D(n_k+1, g_{k+1}) \leq M_P$, then stop; (n_k+1, g_{k+1}) is a feasible solution to OPT6.1. Failing this, compute an extrapolated value $n_{k+1} = \lfloor n_k + 1 + \frac{h_1(n_k+1, g_{k+1})}{h_1(n_k, g_{k+1}) - h_1(n_k+1, g_{k+1})} \rfloor$, replace k with $k+1$, and return to step 1.

Phase II

Due to the monotonicity properties 6.2.1 and 6.2.2 respectively of the average delay and call dropping probabilities of the cell with respect to n and g, we can design a simple algorithm that refines the solution obtained via Phase I to a true optimal solution for the problem OPT6.1.

The basic idea in the implementation is as follows. We begin with the given feasible solution from Phase I, and reduce the value of n in unit steps. If reducing the value of n violates the delay constraints under all admissible values of g, then the previous values of g and n were optimal. To verify this efficiently for the reduced value of n, we make the value of g the largest possible without violating the delay constraints. We then check if the new values of n and g violate the call dropping probability constraint. If they do not then we have found a better solution, and we retry this step by reducing n by 1, and then readjusting g if required to the largest possible value without violating the delay constraints and repeating. On the other hand, if the new values of n and g violate the call dropping probability constraint, then by properties 6.2.1 and 6.2.2, the previous values of n and g were optimal.

The algorithm to search for the optimal solution to OPT6.1 is described below. Finite convergence to an optimum is readily evident by properties 6.2.1 and 6.2.2.

1. Reduce n_k by 1, i.e., let $n_{k+1} = n_k - 1$.
2. Evaluate $h_2(n_{k+1}, g_k)$
   - If $h_2(n_{k+1}, g_k) = 0$, then find maximum $i \geq 0$ for $g_k + i < n_{k+1}$ such that $h_2(n_{k+1}, g_k + i) = 0$. Let $g_{k+1} = g_k + i$ and go to step 3.
   - If $h_2(n_{k+1}, g_k) > 0$ find minimum $i \geq 1$ such that $i \leq g_k$ and $h_2(n_{k+1}, g_k - i) = 0$. If no such i can be found then (n_k, g_k) solves OPT6.1; stop. Else set $g_{k+1} = g_k - i$ and go to step 2.
3. Evaluate $h_1(n_{k+1}, g_k)$.
   - If $h_1(n_{k+1}, g_k) = 0$, then set $g_{k+1} = g_k$, $k = k + 1$ and return to step 1.
   - If $h_1(n_{k+1}, g_k) > 0$, then (n_k, g_k) solves OPT6.1 stop.

6.5 NUMERICAL EXAMPLE

In this section, we consider several instances of this problem for different values of the maximum allowed delay (M_D) and the call dropping probabilities (M_P), and determine corresponding optimal values for n and g. In the tables given below, the third and fourth columns give the values of average delay and call dropping probabilities obtained at the suggested values of n and g. In column seven, the total number of functional evaluations h_1 or h_2 in Phase I of the algorithm are given. In column eight, the number of the functional evaluations of h_1 or h_2 in Phase II are given. Tables 6.1 and 6.2 summarize the results for the case with and without channel borrowing, respectively. In the case of Table 6.2 we used $b = 2$ and $\phi_1 = \phi_2 = 0.6$.

Table 6.1

Optimized n and g for $\lambda = 30$ and $\gamma = 8$.

M_D	M_P	Realized Delay	Realized drop Prob.	n	g	Func. eval. Phase I	Func. eval. Phase II
0.4	0.2	0.292405	0.107305	40	1	4	1
0.4	0.02	0.300733	0.018578	42	2	9	5
0.02	0.3	0.016307	0.195681	42	0	4	2
0.4	0.003	0.348604	0.000646	44	4	9	3
0.02	0.0003	0.013948	0.000061	51	4	14	1
0.02	0.00003	0.013966	0.000009	52	5	20	4

Table 6.2.

Optimized n and g for $\lambda = 30$, $\gamma = 8$,$b = 2$ and $\phi_1 = \phi_2 = 0.6$.

M_D	M_P	Realized Delay	Realized drop Prob.	n	g	Func. eval. Phase I	Func. eval. Phase II
0.4	0.2	0.248202	0.042963	41	1	4	3
0.4	0.02	0.326364	0.008618	42	2	4	1
0.02	0.3	0.055876	0.018054	46	0	8	2
0.4	0.0003	0.349676	0.000293	44	4	9	5
0.02	0.0003	0.013913	0.000175	50	3	13	1
0.02	0.00003	0.013960	0.000027	51	4	18	3

6.6 CONCLUSIONS

In this chapter, we have developed techniques to understand and predict the operational behavior of the cell. Due to the analytical tools developed in Chapter 5, we have been able to predict changes in average queue length and call dropping probabilities with changes in the operational environment of the cell. This knowledge was used to design an optimization technique to configure cells in a spectrally-efficient manner.

If a cluster of similar cells was to be optimized, the approximations developed for borrowing probabilities in Chapter 5 could be used to get a value of ϕ and the optimization techniques described in this chapter could be used.

If the cluster consisted of dissimilar cells, then the techniques described in Chapter 5 can also be used. In this case, we could assume that all the cells in the cluster are like the worst cell, from the point of view of borrowing. Assuming that the entire system consists of that type of cell, we could determine the corresponding optimal values of n and g. These could be used as a starting value for a response surface optimization using simulation. From Figures 6.3.2 and 6.3.3 it can be conjectured that the montonicity properties of average queue lengths and call dropping probability with respect to n_c continue to hold. If that is the case, then letting H denote the cost associated with each channel in the common pool, a cluster optimization problem of the form :

$$\text{Minimize } Kn + Hn_c \qquad OPT6.3$$

Subject to :

$$P_D(n, g, n_c) \leq M_P$$

$$D_A(n, g, n_c) \leq M_D$$

$$0 \leq g \leq n; \quad n_c \geq 0; \quad n, g, n_c : \text{ integer},$$

may be solved in the manner similar to that described in Section 6.4, using simulation to provide the values of P_D and D_A. Details of this will be undertaken as future work.

Chapter 7.

Summary and Conclusions

In this dissertation, we have explored several operational aspects of a cellular system. We have defined the problem faced by a typical system analyst in developing a design for a cellular system. A literature survey has been conducted and it was determined that a typical cellular system is designed by sequentially solving the economic problem, the cell design problem, and then the channel assignment problem. Our definition of the cell design problem considers a given prediction of the user population behavior within a certain geographical region, and seeks an optimal cell size, along with the location and number of channels required to fulfill the operational goals of dropping probability and average delays. The modern cellular systems use guard channels with call queueing and a Hybrid Channel Allocation scheme. The analytical techniques available, model handoff behavior without considering the effect of Hybrid Channel Allocation. Furthermore, the techniques available to model call queueing with guard channels do not have solution techniques that are easy to implement and numerically efficient. Usually, the effects of Hybrid Channel Allocations have been studied through simulation. These studies assumed simplified versions of handoff management strategies. It is also observed that, it is difficult to derive stable results from simulations because they involve simulation of call dropping, which is a rare event. In this dissertation, we present a comprehensive model that incorporates Hybrid Channel Allocation and guard channels with new call queueing into a single analytical model. The optimization algorithm developed is also unique to this problem. The proposed solution scheme is designed to be implemented on a low power cell-site processor.

More specifically, we have addressed the problem of determining the minimum number of channels that would be required to fulfill certain defined operational goals of a cell. A formal mathematical statement of the problem definition is presented next. Let $D_A(n_i, g_i)$ represent the average delay of newly originating calls, and let $P_D(n_i, g_i)$ represent the dropping probability of a transferred call in a cell having a total of n_i channels and g_i guard channels. Furthermore, let M_D and M_P represent the maximum allowable delay and the dropping probability, respectively. The cell design problem

can be written as the following stochastic optimization problem:

$$\text{Minimize } n_i$$

$$\text{Subject to}$$

$$P_D(n_i, g_i) \leq M_P$$

$$D_A(n_i, g_i) \leq M_D$$

$$0 \leq g_i \leq n_i; \quad n_i, g_i \quad : \text{integer.}$$

The queueing model of the cell behavior, the solution technique for this model and the optimization algorithm developed are all contributions of this dissertation.

A G/PH/1 queueing system is used to model the cell behavior in a cellular system. This system has two Poisson arrival streams of customers. Customers with a higher priority arrive at a Poisson rate γ, and the ones with a lower priority arrive at a Poisson rate λ. It takes an exponentially distributed amount of time with mean $\frac{1}{\mu}$ to complete service. There are a total of n servers available to service customers. If the number of free servers reduces to g, or less then the lower priority customers are queued in a first-come-first-served queue. If a high priority customer arrives and does not find a free server, then a server can be borrowed. At the most, b servers can be borrowed. Let ϕ_1 be the probability of making a successful bid for the first borrowed server, let ϕ_2 be the probability of acquiring a second server if a second server is requested, and so on. If a high priority customer does not find a free server, either in the cell or by borrowing, then the customer balks. When a server becomes free. it is returned to the common pool, if one has been previously borrowed by the cell. The use of this queueing model allows the cell behavior to be modeled in a broader sense than ever before. It incorporates the use of guard channels, queueing of new calls, and the Hybrid Channel Allocation scheme. A numerically stable and efficient solution to the queueing system has been developed. This solution technique allows us to easily determine performance measures of systems of practical dimensions. The numerical stability and efficiency of the queueing model is very valuable because it allows the optimization algorithm to execute quickly.

The analysis technique of the G/PH/1 queueing system has also provided valuable insights into the behavior of the cell. This led to the discovery of some important properties related to cell behavior. These are enumerated below.

1) Increasing the number of guard channels while keeping the shared channels and the nominal channels constant, increases the average delay and decreases the call dropping probabilities.
2) Increasing the number of nominal channels while keeping the number of guard channels and the number in the common pool constant, decreases the average delays and the call dropping probability.
3) An extra borrowed channel available is equivalent to increasing the number of nominal channels and guard channels simultaneously.

The above properties were used to design an algorithm that provides optimized values of n and g.

A simulation was implemented using the simulation language SLAM II. This simulation was used to validate analytical results. The simulation was designed using synchronized antithetic random numbers to reduce variances.

The discussion in this dissertation is based on a frequency multiplexed cellular system for which frequency channels are the voice medium. The same analysis can be applied to other systems. For example, in a TDMA system, the frequency channels would be replaced by time slots.

7.2. FUTURE RESEARCH

In Chapter 5 we have presented an approximation for the channel borrowing probabilities, assuming K similar cells. An extension of the method for dissimilar cells can be done. Consider K cells that may not be similar. Each of these cells will generate overflow traffic as described in Chapter 5, which can be modeled as interrupted Poisson Processes(IPP). The multiple IPPs can then be represented by a Markov Modulated Poisson Process having 2^K states, which can then be applied to the common pool. In order to further develop this idea, ideas from Meier-Hellsterm[1989] may be used.

In some places in the cellular network, a third type of calls occur, which represent retries of blocked calls. As the system gets overloaded, the calls that are dropped, retry and generate a peaky traffic. In some cases, this retry traffic can be many times the new call traffic. Furthermore, as the cell saturates, the retry traffic increases. A possible way to model the retry traffic may be through an IPP. A model of the cell that incorporates this third kind of traffic would be very useful.

Standards for the cellular telephone system (Telecommunications Industry Association, IS-54-B [1992]), require the new call to wait for only a finite amount of time. This can be approximated by using a finite queue length extension of the G/PH/1 queueing system described in Chapter 5.

Bibliography

- Akimaru, Haruo; Takahashi, Haruhisa . June 1983. "An Approximate Formula for Individual Call Losses in Overflow Systems." *IEEE Transactions in Communications,* Vol. COM-31, No. 6:808-811.

- Altiok, T. 1989. "(R, r) Production/Inventory Systems." *Operations Research,* Vol. 37, No. 2:266-276.

- Anderson, Lewis G. November 1973. "A Simulation Study of Some Dynamic Channel Assignment Algorithms in a High Capacity Mobile Telecommunication Systems." *IEEE Transactions in Communications* Vol. COM-21, No. 11:1294-1301.

- Bazaraa, Mokhtar S.; Sherali, Hanif D.; Shetty, C. M. 1993. *Nonlinear Programming Theory and Algorithms.* Second Edition. John Wiley & Sons, Inc., New York.

- Beuerman, S. L. and E. J. Coyle. 1988. "The Delay Characteristics of CSMA/CD Networks." *IEEE Transactions on Communications,* Vol. COM-36, No. 5:553-562.

- Cimini, Leonard J.; Foschini, Gerard J.; I, Chih-Lin. 1992. "Call Blocking Performance of Distributed Algorithms For Dynamic Channel Allocation in Microcells." *Proceedings of International Conference on Communications ICC '92*; June 14-18, 1992:1327-1332.

- Cooper, R. B. 1972. *Introduction to Queueing Theory.* Macmillan, New York.

- Cox, D. C. and D. O Reudink. July-August 1971."Dynamic Channel Assignment in High-Capacity Mobile Communications Systems." *The Bell System Tehnical Journal,* Vol. 50, No. 6:1833-1849.

- Cox, D. C. and D. O Reudink. April 1972. "A Comparison of Some Channel Assignment Strategies in Large-Scale Mobile Communications Systems." *IEEE Transactions on Communications,* Vol. COM-20, No. 2:190-195.

- Cox, D. C. and D. O Reudink. September 1972. "Dynamic Channel Assignment in Two-Dimensional Large-Scale Mobile Radio and Systems." *The Bell System Tehnical Journal,* Vol. 51, No. 7:1611-1629.

- Coyle, E. J. and B. Liu. 1985. "A Matrix Representation of CSMA/CD Networks." *IEEE Transaction on Communications,* Vol. COM-33, No. 1:53-64.

- Daigle, J. N. and J. D. Langford. 1985. "Queueing Analysis of a Packet Voice Communication System." *Proc. IEEE INFOCOM,* Washington, D. C.:18-26.

- Daigle, J. N. 1989. "Queuelength Distributions from Probability Generating Functions via Discrete Fourier Transforms." *Operations Research Letters*, Vol. 8:229-236.

- Daigle, J. N. and D. M. Lucantoni. 1990. "Queueing Systems Having Phase Dependent Arrival and Service rates." *Proceedings of First International Conference on the Numerical Solution of Markov Chains,* Raleigh, N.C.:375-395.

- Daigle, J. N. 1992. *Queueing Theory for Telecommunications.* Addison-Wesley, New York.

- Daigle, J. N.; Jain, N. May 1992. "Queueing System with Two Arrival Streams and Reserved Servers with Application to Cellular Telephone. " *Proceedings IEEE INFOCOM '92,* Florence, Italy :2161-2167.

- Duque-Antón, Manuel; Kunz, Dietmar; Rüber, Bernard. February 1993. "Channel Assignment for Cellular Radio Using Simulated Annealing." *IEEE Transactions on Vehicular Technology.* Vol. 42, No. 1:14-21

- Eklundh, B. April 1986. "Channel Utilization and Blocking Probability in a Cellular Mobile Telephone System with Directed Retry." *IEEE Transactions on Communications,* Vol. COM-34, No. 4:329-337.

- Elnoubi, S. M.; Singh, R.; Gupta, S. C. August 1982. "A New Frequency Channel Assignment Scheme in High Capacity Mobile Communication Systems." *IEEE Transactions Veh. Technology* VT-32 :125-131.

- Everitt, D. E.; Macfadyen, N. W. . October 1983. "Analysis of Multicellular Mobile Radiotelephone System with Loss." *Brit. Telecommun. Technol. Journal,* Vol. 1, No. 2:37-45.

- Everitt, D. E. and D. Manfield. October 1989. "Performance Analysis of Cellular Mobile Communication System with Dynamic Channel Assignment." *IEEE Journal of Selected Areas in Communication,* Vol. 7:1172-1180.

- Fredericks, A. A. 1980. " Congestion in Blocking Systems - A Simple Approximation Technique." *The Bell System Technical Journal,* Vol. 59:805-827.

- Funabiki, Nobuo; Takefuji, Yoshiyasu. November 1992. "A Neural Network Parallel Algorithm for Channel Assignment Problems in Cellular Radio networks." *IEEE Transactions on Vehicular Technology,* Vol. 41, No. 4:430-437.

- Gamst, A; Rave, W.1982. "On Frequency Assignment in Mobile Automatic Telephone Systems." *Proc. GLOBECOM '82*:309-315.

- Gamst, A; Beck, R; Simon, R; Zinn, E. -G. May 21-23, 1985. "An Integrated Approach To Cellular Radio Network Planning." *Proc. 35th IEEE Vehicular Technology Conference;* Boulder, Colorado:21-25.

- Gamst, Andreas. Febuary 1986. "Some Lower Bounds for a Class of Frequency Assignment Problems." *IEEE Transactions on Vehicular Technology,* Vol. VT-35, No. 1:8-14.

- Guérin, Roch. February 1988. "Queueing-Blocking System with Two Arrival Streams and Guard Channels." *IEEE Transactions in Communications,* Vol. COM-36, No. 2:153-163

- Gün, L. 1989. "Experimental Results on Matrix-Analytical Solution Techniques - Extensions and Comparisons." *Stochastic Models, Special Issue on Computer Experimental Methods.*

- Hong, Daehyoung and Stephen S. Rappaport. August 1986. "Traffic Model and Performance Analysis for Cellular Mobile Radio Telephone Systems with Prioritized and Nonprioritized Handoff Procedures." *IEEE transactions on Vehicular Technology,* Vol. VT-35, No. 3:77-92.

- Hunter, J. J. 1983. *Mathematical Techniques of Applied Probability Volume 1, Discrete Time Models: Basic Theory.* Academic Press, New York.

- IEE Telecommunications Series 14. 1985. *Land Mobile Radio Systems,* Peter Peregrinus Ltd., UK.

- Telecommunications Industry Association. April 1992. *EIA/TIA Interim Standard; IS-54-B; Cellular System Dual-Mode Mobile Station - Base Station Compatibility Standard.* Telecommunications Industry Association, Engineering Department, 2001 Pennsylvania Avenue N.W., Washington, D.C. 20006.

- Jain, Nikhil. 1991. *Eigenanalysis Solution for Quasi Birth and Death Process.* A thesis submitted to Virginia Polytechnic and State University, Blacksburg VA.

- Jakes, W. C., Jr. 1974. *Microwave Mobile Communications.* New-York: Wiley.

- Kahwa, T. J.; Georganas, N. D. April 1978. " A Hybrid Channel Assignment Scheme in Large-Scale, Cellular-Structured Mobile Communication Systems." *IEEE Transactions in Communications,* Vol. COM-26:432-438.

- Kelly
Statistical Society, Vol. 47, No. 3:379-395.

- Kelly, F. P. 1986. "Blocking Probability in Large Circuit Switched Network." *Advances in Applied Probability,* Vol. 18:473-505.

- Kuczura, A. March 1973. "The Interrupted Poisson Process as an Overflow Process." *Bell System Technical Journal,* Vol. 52:437-448.

- Kuczura, Anatol and Dinesh Bajaj. February 1977. "A Method of Moments for the Analysis of a Switched Communication Network's Performance." *IEEE Transactions on Communications,* Vol. COM-25, No. 2:185-193.

- Lee, William C. Y. 1989. *Mobile Cellular Telecommunications Systems.* McGraw-Hill Book Company, New York.

- Lee, William, C. Y. 1991. "In Cellular Telephone Complexity Works." *IEEE Circuits and Devices.* January 1991.

- MacDonald, V. H. January 1979. "Advanced mobile phone service: The Cellular Concept." *Bell Syst. Tech. Journal,* Vol. 58, No. 1:15-41.

- Matsumoto, Jun; Watanabe, Yu . January 1985. "Individual Traffic Characterstics of Queueing Systems with Multiple Possion and Overflow Inputs." *IEEE Transactions on Communications,* Vol. COM-33, No. 1:1-9.

- Meier-Hellstern, K. S. 1989. The Analysis of a Queue Arising in Overflow Models. *IEEE Transactions on Communications,* Vol. 37, No. 4:367-371.

- Neuts, M. F. 1981. *Matrix Geometric Solutions in Stochastic Models.* The John Hopkins University Press, Baltimore.

- Noble, B.; Daniel, J. W. 1977. *Applied Linear Algebra.* Prentice Hall, New Jersey.

- Oh, Se-Hyun; Tcha, Dong-Wong. July 1992. "Prioritized Channel Assignment in a Cellular Radio Network. " *IEEE Transactions on Communications,* Vol. 40, No. 7:1259-1269.

- Rappaport, Stephen. August 1991. "The Multiple-Call Hand-Off Problem in High-Capacity Cellular Communication Systems." *IEEE Transactions on Vehicular Technology,* Vol. 40, No. 3:546-557.

- Raymond, Paul-Andŕe. December 1991. "Performance Analysis of Cellular Networks." *IEEE Transactions on Communications,* Vol. 39, No. 12.

- Ross, S. M. 1989. *Introduction to Probability Models.* Academic Press, New York.

- Sherali, H. D; Tuncbilek, C. H. 1991. "A Global Optimization Algorithm for Polynomial Programming Problems Using Reformation-Linearization Technique." *Recent Advances in Global Optimization.* Editors Floudas C. and P. M. Pardalos.

- Yoon, C. H; Un, C. K. 1989. "Effecient Handoff Policy Without Guard Channels for Mobile Radio Telephone System." *Electronic Letters,* Vol. 25, No. 11.

- Zhang, M.; Yum, T. -S. 1989. "Comparison of Channel Assignment Strategies in Cellular Mobile Telephone Systems." *IEEE Int. Conf. Commun* :15.2.1-15.2.5.

Glossary of Frequently Used Symbols

C_c : Per second average price for a cellular call charged by the cellular telephone company

C_t : Average cost incurred per second by the cellular telephone company in setting up and maintaining a call

D : Channel reuse distance

$D_A(n, g)$: Expected delay in a cell with n nominal channels and g guard channels

g : Number of guard channels

I : Extent of interference, measured as the signal to noise ratio

K : Number of cells in a cluster

M : Total number of cells in a cellular system

M_D : Maximum average delay permitted for new calls

M_P : Maximum call dropping probability allowed

n : Number of nominal channels assigned to a cell

n_c : Number of channels in the common pool

N_c : Average number of calls per second

N_T : Total number of channels available for cellular service to a cellular telephone company

$P_D(n, g)$: Call dropping probability in a cell with n nominal channels and g guard channels

R : Rate matrix

$\tilde{x}$: Length of a service time

$\tilde{x}_e$: Length of the exceptional first service.

λ : Rate of arrival of new calls

γ : Rate of arrival of handoff calls

δ : Scale factor

μ : $\frac{1}{\mu}$ is the mean of an exponentially distributed time for which a channel is held. The channel is held from the time it is allocated to a call to the time the call either completes or is handed over to another cell.

Vita

Nikhil Jain graduated from St. Joseph's Academy high school as valedictorian in 1980 and earned the bachelor's degree in electrical engineering from the Indian Institute of Technology, Madras in 1985. On graduating, he received a grant from German Council for Technical Exchange (DAAD) to work on automation of high voltage measurements at Physikalische Technische Budestanstalt (PTB) in Braunschweig, Germany. He subsequently earned the MBA degree with specialization in Finance and Information Systems at University of Rochester, Rochester, New York, in 1987. In 1991 he received the M.S. degree in Operations Research at Virginia Polytechnic and State University, Blacksburg, Virginia. He has published articles on stochastic models of cellular telephone system and on designing and implementing library information systems. He has taught the junior-level random process course at Virginia Polytechnic Institute and State University, where he entered the Ph.D. program in Operations Research in 1989. He is a member of Alpha Pi Mu.