RF Engineering
Continuing Education and Training

Introduction to Traffic Planning

August 23, 2001

Ivica Kostanic
Target Audience
This class is intended for intermediate to experienced RF engineers.

Course Description
This one-day class introduces the importance of traffic analysis in the provisioning of optimum wireless networks. The course is divided into four general areas that cover terminology and background knowledge, traffic models for voice applications, traffic models for data applications, and engineering applications and techniques. The goal is to provide the trainee with sound engineering and statistical techniques useful for the appropriate allocation of traffic resources for wireless networks under a variety of operating conditions and growth scenarios. Class topics include Erlang B, Erlang C and data traffic models, traffic planning for a startup system vs. an established system, and strategies for relieving traffic congestion.

Objectives
- Define all units of measure in traffic engineering
- Describe the concept of Grade of Service (GOS) and it’s impact on the number of required voice channels
- Learn how to do traffic planning in circuit switched voice cellular networks
- Familiarize with implementation of data services in 3G cellular networks
- Learn traffic planning techniques for data traffic in wireless cellular networks
- Explain the traffic trending methodology
- Explain the most important methods for traffic congestion control
- Optionally – analyze traffic dimensioning in CDMA systems and packet data traffic modeling in GSM/GPRS systems.

Prerequisites
The student should be familiar with college level algebra and basic cellular engineering concepts.

Length
8 Hours
# TABLE OF CONTENTS

## 1 Introduction........................................................................................................................................5
  1.1 Traffic in Cellular Telecommunication Networks ........................................................................5
  1.2 Circuit and Packet Switching ........................................................................................................6
  1.3 Types of Traffic in Existing Cellular Networks .............................................................................8

## 2 Queuing Systems..................................................................................................................................9
  2.1 Description of a Queuing System ..................................................................................................9
  2.2 Poisson Process of Random Arrivals ..........................................................................................16
  2.3 Birth and Death Processes ........................................................................................................21
    2.3.1 State Diagram Representation of Birth and Death Process ................................................22
    2.3.2 Little’s Formula ....................................................................................................................26
  2.4 Kendall’s Notation .......................................................................................................................27
  2.5 Examples .........................................................................................................................................28

## 3 Traffic Planning for Circuit Switched Voice Services ....................................................................32
  3.1 Common Descriptors of the Circuit Switched Voice Service ......................................................32
    3.1.1 Busy Hour............................................................................................................................32
    3.1.2 QoS parameters ..................................................................................................................33
  3.2 Trunking Models for Circuit Switched Services ..........................................................................34
    3.2.1 Erlang B Formula................................................................................................................34
    3.2.2 Erlang C Formula................................................................................................................39
  3.3 Optional: Traffic Planning for Circuit Switched Services in CDMA Networks .........................45
    3.3.1 Reverse link.........................................................................................................................45
    3.3.2 Forward link .......................................................................................................................49
  3.4 Estimation of the Offered Traffic in Cellular Systems .................................................................52
  3.5 Case Study Example .....................................................................................................................55

## 4 Traffic Planning for Packet Data Services.......................................................................................59
  4.1 Implementation of Data Services in Cellular Networks ...............................................................61
  4.2 Packet Data Traffic Modeling .....................................................................................................63
    4.2.1 Description of Packet Data Traffic ......................................................................................64
    4.2.2 Self Similarity in Data Traffic ..............................................................................................65
    4.2.3 Pareto Distribution of the Service Times ..............................................................................68
  4.3 Data Services in Cellular Networks .............................................................................................72
    4.3.1 Low-End Data Service in Data Only Networks ..................................................................73
    4.3.2 Low-End Data Services in Mixed Traffic Networks ..........................................................75
    4.3.3 High-End Data Services .....................................................................................................76
  4.4 Optional: Traffic Planning for Packet Data in GSM/GPRS Networks ........................................79
    4.4.1 Estimation of the GPRS Data Capacity ...............................................................................81

## 5 Traffic Trending and Growth Planning ............................................................................................87
  5.1 Linear Regression ........................................................................................................................87
  5.2 Special Events and Seasonal Trends ...........................................................................................89
  5.3 Case Study: Growth Planning for a Site in a Cellular Network ................................................91

## 6 Methods for Traffic Congestion Control in Cellular Systems .......................................................94
  6.1 Congestion Control Through Addition of Resources ..................................................................95
    6.1.1 Addition of new radios/trunks .............................................................................................95
6.1.2  Sectorization ................................................................. 95
6.1.3  Cell Splitting and Microcell Deployment ................................ 96
6.2  CONGESTION CONTROL THROUGH CELL SITE COVERAGE BALANCING ................................................. 98
  6.2.1  Optimization of System Parameters ................................ 99
  6.2.2  Overlay and Underlay Configurations ................................ 99
  6.2.3  Hierarchical cell structure .................................................. 100

7  REFERENCES .............................................................................. 101
1 Introduction

Telecommunication traffic, or simply traffic, can be defined as the flow of information messages through a communication network. The traffic is generated as a result of telephone conversations, data exchange, audio and video delivery and various other types of communication services offered by the network. In general, networks are designed to provide communication services to many users. For example, a typical cellular network may have several millions of subscribers. However, not all of users access the network at the same time. In reality, at any given time, only a small fraction of eligible users are communicating. For that reason, the networks allow communication resources to be shared between different users, which helps reduce the cost of the network deployment and increases its overall efficiency. However, whenever there are multiple users sharing the same set of resources, there is a finite probability that a particular user may not be able to obtain service because all existing resources are already busy. If more equipment is provisioned, the probability of service denial becomes smaller, but at the same time, the cost of the network increases.

In this document, we address one of the most important areas of the communication network's design and operation: traffic planning. Traffic planning can be loosely defined as a set of engineering practices and procedures that balance the overall cost of the communication network and its availability. The subject of traffic planning is extremely wide and therefore we limit our focus to the area of cellular communications. However, although it may not be explicitly indicated, many of the results presented here have a general applicability and can be used for the design and analysis of other communication networks as well.

1.1 Traffic in Cellular Telecommunication Networks

An outline of a typical cellular communication network is presented in Fig. 1.1. As seen, the network consists of many interconnected elements. The traffic planning for the network shown in Fig. 1 has two aspects. First, each of the interconnected elements must have sufficient processing capability to provide service to the incoming traffic. Second, each of the communication links between the elements has to have sufficient capacity to carry the traffic generated at each end.

In general, analysis of the entire network presented in Fig. 1 is a complicated task. Typical engineering practice is to analyze each of the links individually and guarantee the meeting of certain performance requirements. If individual links are dimensioned properly, the behavior of the entire network is likely to be within a required quality margin as well. Furthermore, in the case of cellular communication networks, the most critical communication link is the radio link between mobile terminals and base stations. Due to limited availability of the radio spectrum, this link is usually the traffic bottleneck of the system. For that reason, the majority of material presented in this document focuses on the traffic dimensioning of the air interface.
1.2 Circuit and Packet Switching

The first and second generation cellular technologies provide connection oriented communication service for each user. A dedicated voice channel is allocated throughout the entire duration of the mobile call. It is common practice to refer to this mode of communication as *circuit switched* mode. The interpretation of the term *circuit* is a function of the access scheme in use and it can be a pair of radio frequencies in FDMA systems, a pair of frequencies and associated time slots in FDMA/TDMA systems, or, appropriate orthogonal codes in CDMA based systems. Table 1.1 presents the meaning of the term *circuit* as it is interpreted in different radio technologies of the first and the second generation.

**Table 1.1.** Interpretation of term *circuit* for various first and second generation technologies

<table>
<thead>
<tr>
<th>Technology</th>
<th>Circuit</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMPS/NAMPS</td>
<td>A pair of frequencies</td>
</tr>
<tr>
<td>TACS/NMT</td>
<td>A pair of frequencies</td>
</tr>
<tr>
<td>GSM</td>
<td>A pair of frequencies with associated time slots</td>
</tr>
<tr>
<td>IS-136 (NA-TDMA)</td>
<td>A pair of frequencies with associated pair of time slots</td>
</tr>
<tr>
<td>CdmaOne (IS-95)</td>
<td>A pair of frequencies with associated codes</td>
</tr>
</tbody>
</table>

No matter what the physical interpretation is, a circuit can be seen as a basic communication resource. Circuit switching assumes permanent allocation of the resource over the duration of the mobile call regardless of user activity. Within the voice centric networks, this may be considered adequate since the activity of the users is relatively high. For nominal planning purposes, it is assumed that during the voice call, each party exhibits approximately 50% activity. This is “only” 50% wasteful since the circuit is occupied even in the time interval when the party is silent.
On the other hand, data services are commonly characterized by short and bursty transmissions followed by long periods of user inactivity. For such modes of communication, circuit switching is highly inefficient and most of the data communication networks are designed to allocate the communication resources using *packet switching* principles. In packet switching networks, a user utilizes the network resources only during the periods when there is a need to transmit data. During periods of inactivity, resources are released and the network can perform their reallocation to support the communication needs of other users in the system. To allow for easier management of network resources, users segment their data streams into packets. This is where the name for the switching scheme is derived. Depending on how the packets are delivered from the source to the destination, we distinguish two kinds of packet switching: connection oriented (*virtual path*) and connectionless (*datagram*). Virtual path based packet switching delivers all packets using the same path through the network, while in datagram networks, packets follow their independent paths. The illustration of the two concepts is shown in Figs. 1.2 and 1.3.

![Virtual path packet switching](image1)

**Figure 1.2.** Virtual path packet switching

![Datagram path packet switching](image2)

**Figure 1.3.** Datagram path packet switching

Virtual path switching is commonly identified with the ATM transport while datagram is the default mode of transport for TCP/IP based networks.
1.3 Types of Traffic in Existing Cellular Networks

At present, the most prevalent type of traffic in the existing cellular network is circuit switched voice. With the enormous success of the land line data services, many cellular providers are attempting to enrich their service portfolio by providing some of the most popular data services to mobile users. As a result, many wireless cellular networks are capable of providing service to both voice and data traffic. Since most of the existing digital networks are second generation networks, the data services are still provided either in the form of a circuit switched service or through common control channels. Recently, a new group of cellular standards has been developed with a main goal of providing simultaneous support for both circuit switched voice and packet switched data. This group of standards is commonly referred to as the third generation and many cellular providers have announced their rapid implementation.
2 Queuing Systems

Figure 2.1 shows a schematic representation of a queuing system. This representation is a mathematical abstraction suitable for many different arrangements in which users compete for a shared set of resources (servers). In everyday life, such arrangements are very common, and their analysis provides useful results with wide range applicability. In this section we will address the queuing problem in its general form. This approach will allow us to treat various practical problems in traffic engineering through a unified mathematical framework. Despite the general approach, we will illustrate the underlying concepts by using examples that are relevant to the field of cellular system traffic engineering.

![Figure 2.1. A schematic representation of a queuing system]

2.1 Description of a Queuing System

As evident from Fig. 2.1, queuing systems are relatively complex. Before we make an effort to analyze them, we need to define some important terms and variables.

1. Source population (number of subscribers). The source population consists of all users that are eligible for service in a given queuing system. In general, the most important property of the source population is its size. From the standpoint of theoretical modeling, we make a distinction between finite and infinite source population. For the infinite population the average number of service requests does not depend on the number of users that are currently being served. On the other hand, for finite populations, the probability of a new service
request decreases every time a user enters the queuing system. From a mathematical standpoint, the infinite population is easier to describe and is frequently used for traffic analysis. In reality, every population is finite and which one of the two assumptions is used depends on the ratio between the number of potential users and the number of available servers. If this ratio is large, we routinely assume that the population is infinite.

2. **Arrival rate and interarrival time.** The arrival rate is one of the variables used to quantify the volume of generated traffic. Within the queuing system, the arrival rate is defined as a number of service requests made in some specified time interval. The ability of the queuing system to provide effective service depends not only on the mean arrival rate but also on its distribution. If the requests for service are evenly spaced in time, the queuing system can provide better service than if the call attempts are clustered. As an illustration, consider two graphs showing the number of call attempts for an imaginary cell site presented in Fig. 2.2. Both graphs have the same mean arrival rate of about 20 call attempts per minute. However, the statistical behavior of the number of call attempts in Fig. 2.2 (b) is much burstier. To assure that no calls are rejected, the number of resources allocated to the site shown in Fig. 2.2 (a) should be 31, while in Fig. 2.2 (b), we need to allocate 42 resources. This is a significant difference (more than 30%) and it underlines the importance of the arrival rate distribution.

![Figure 2.2](image.png)

**Figure 2.2.** Number of call attempts during one hour of cell site operation. Both graphs have average of 20 call attempts per hour.

The standard way to specify the arrival rate is through distribution of interarrival times. The interarrival time is defined as the time interval between two consecutive service requests. The arrival rate and interarrival time are inversely proportionate. In other words, as the arrival rate increases, the interarrival time becomes smaller.

3. **Servers.** The server is a part of the queuing system capable of performing a service task. The practical implementation of the server is determined by the type of service that the queuing system is intended to provide. Examples of servers are: a computer scheduling jobs that are sent to a shared printer; a cashier in the supermarket, a toll booth on the highway and so on. In cellular systems, the notions of the server and the circuit are essentially the same. Table 1.1 specifies what can be seen as a server in various first and second generation cellular technologies. The part of the queuing system hosting servers is usually referred to as
the service facility. If all servers at the service facility are busy when the call enters the system, the call must join the queue and wait for a server to become available.

4. Service time (Call holding time). The period of time over which a server is allocated to an individual user is called the service time or the call holding time. In general, the service time can also be seen as a random variable. As in the case of the interarrival times, performance of the queuing system depends fundamentally on the service time distribution. For example, in cellular networks carrying predominantly voice traffic, the exponential distribution is commonly used to describe distribution of the service times. Consider measurements of the service time illustrated in Fig. 2.3. The exponential character of the distribution is evident. The only significant deviation from the exponential distribution occurs for brief service time duration.

The measurements presented in Fig. 2.3 were collected in a cell servicing users with relatively low mobility. In cells where users are highly mobile, the distribution of holding time deviates from exponential for large call holding time values as well. The reason for deviation resides in the handoff process. Due to mobility, a user spends only a portion of the call holding time within the coverage area of a given cell. Therefore, the calls of extremely long duration become highly unlikely.

For exponential distribution of the call holding time we can write

\[
\Pr \{ \text{CHT} < t \} = 1 - \exp \left( -\frac{t}{T_s} \right)
\]  

(2.1)
where CHT is the call holding time and $T_s$ is the distribution parameter referred to as the *average call holding time*. The average call holding time in cellular networks varies as a function of service price, cultural differences, time of the day and number of other parameters. Typical values range from 120 to 180 seconds.

The quantity that is an inverse of the service time is the *service rate*. The service rate is defined as the number of users that can be provided with the service in a given unit time provided that the server is never idle. For example, for the distribution of the service times given in (2.1), the average service rate can be calculated as $\mu = 1/E[t] = 1/T_s$.

5. **Average resource occupancy – traffic in erlangs.** The unit used in traffic engineering as a measure for the server occupancy is called erlang (E). By definition, a single device occupied continuously or intermittently for a total time $t$ over some averaging time $T$ carries traffic of

$$A = \frac{t}{T} \ [E]$$

(2.2)

From (2.2) we see that the maximum traffic that can be carried by a single resource is 1 E. The traffic of 1 E corresponds to the case when the resource is occupied for the entire duration of the averaging time interval $T$. As an illustration, consider the graph in Fig. 2.4. The graph specifies the occupancy of a server over some interval $T$. It is important to note that at any given time, the resource is either occupied or not. However, for a stationary environment, the average occupancy of the resource remains constant.

![Average traffic graph](image)

$$A = \frac{t_1 + t_2 + t_3}{T} = \frac{1.5 + 2 + 1}{8} = \frac{4.5}{8} = 0.5635 \ E$$

**Figure 2.4.** Calculation of the resource occupancy

To assure a valid estimate of the average resource occupancy, the averaging time should be long enough. In cellular communication, the typical averaging time is 1 hour.

Since the maximum traffic that can be carried by a single resource has to be smaller than 1, the total traffic carried by a service facility cannot exceed the number of resources. Considering a group of servers in Fig. 2.1, let $t_n$ denote the sum of times during which
exactly $n$ out of $C$ servers are held simultaneously within the averaging period $T$. The total traffic carried by the group can be expressed as

$$A = 1\frac{T_1}{T} + 2\frac{T_2}{T} + \cdots + C\frac{T_n}{T} = \sum_{n=1}^{C} n \frac{T_n}{T}$$  \hspace{1cm} (2.3)$$

From (2.3) we derive a different interpretation of the average traffic for multi-server systems. The expression on the right hand side of (2.3) expresses the average number of servers held simultaneously during the averaging period $T$. This interpretation allows easier measurement of traffic carried by a group of servers. The measurement procedure involves regular polling of the service facility and logging the number of resources occupied at the measurement time.

6. **Offered, Carried and Lost Traffic.** The average offered traffic is defined as

$$A_{offered} = \frac{\lambda T_s}{T}$$  \hspace{1cm} (2.4)$$

where $\lambda$ is the average arrival rate, $T_s$ is the average call holding time, and $T$ is the averaging period. For example, if the rate of phone call attempts at a given cellular site is 100 calls/hour with an average call holding time of 90 sec, the offered traffic is given as

$$A_{offered} = \frac{\lambda T_s}{T} = \frac{100 \times 90}{3600} = 2.5E$$  \hspace{1cm} (2.5)$$

According to the alternative interpretation for traffic in erlangs, (2.5) can be seen as the average number of resources occupied at the service facility. Measurement of the offered traffic requires continuous resource availability. In other words, every service request should find an unoccupied resource and be able to hold it for a desired period of time. Due to a relatively large variability in the offered traffic, this would require a large over-provisioning of server resources. Although in some circumstances it may be justified, the resource over-provisioning is not regarded as a sound engineering practice. Most of the queuing systems are designed to operate with some probability that a particular service request will be denied. The probability of service denial is commonly referred to as the *blocking probability*. Figure 2.5 illustrates the resulting tradeoff in a case of a cellular system cell site. If the cell site is required to operate with no blocking, the number of assigned channels needs to be at least 22. However, it can be seen that with 18 assigned channels, the portion of time when the cell site is blocking is only 1 min during the entire 60 min of monitoring period. This portion of time corresponds to a blocking probability of $1/60 = 1.67\%$, which is assumed acceptable in most cellular systems. Therefore, in practice, only a portion of the offered traffic will be served. This portion, referred to as the *served traffic*, can be formally defined as

$$A_{served} = \sum_{n=1}^{C} n \frac{T_n}{T}$$  \hspace{1cm} (2.6)$$
where \( C \) is the total number of network resources, \( t_e \) is the period of time when exactly \( n \) resources are occupied, and \( T \) is the time period used for date collection and averaging.

![Figure 2.5. Relationship between offered, carried, and lost traffic](image)

The difference between offered and served traffic is commonly referred to as lost traffic. Real systems always operate with a certain level of lost traffic. The task of the traffic planning engineer is to carefully balance the volume of the lost traffic against the number of required resources and provide the most economical solution.

7. **Service discipline (lost calls disposition).** If at the time of service request arrival all resources are occupied, the request has to be placed in a queue. When one of the resources becomes available, it will be allocated to one of the requests in the queue. There are several different algorithms used in determining the order of the resource allocation for the requests that are in the queue. These algorithms are commonly referred to as the queuing discipline. The most common algorithm is the **First Come – First Serve (FCFS)**, which is sometime referred to as the **First In – First Out (FIFO)**. In this algorithm the queuing system keeps track of the order in which the requests are performed, and when the resource becomes available, the same order is used for the resource allocation. Examples of the FCFS queuing discipline are a queue formed in front of an airline ticket counter and a queue of printing jobs in the print server. Another common queuing discipline is the **Last Come – First Serve (LCFS)**, which is sometimes referred to as the **Last In – First Out (LIFO)**. According to this discipline the resources will be allocated in the opposite order of the order request arrivals. This queuing discipline accurately models behavior of the stack in computer systems. Some other queuing disciplines are possible. In systems where the resource access is based on a version of ALOHA protocol, the queuing discipline is commonly referred to as the **Random Selection Order (RSS)** or the **Service In Random Order (SIRO)**. According to this queuing discipline...
The queuing discipline has a significant impact on the performance of the queuing system. Parameters like the average delay time, the average number of users in the queue, the probability of excessive delay, and the probability of the user deflecting from the queue all depend on the enforced queuing discipline. For that reason, when a given queuing system is analyzed, the queuing discipline needs to be taken into account.

8. Maximum Queue Capacity. One of the main characteristics of the queuing system is the capacity of its queue. The capacity of the queue is defined as the number of service requests that it can hold. Based on the queue capacity, systems can be classified as either lossless or lossy. In lossless systems, the capacity of the queue is infinite and every service request is allowed to wait until a resource becomes available. In lossy systems, the queue has a limited capacity and only a limited number of user requests can be placed in the queue. If the number of requests exceeds the queue capacity, the request is denied or blocked. An extreme case of the lossy queuing system is the system with queue capacity equal to zero. This system is commonly referred to as the loss system.

Depending on the goals of traffic engineering the queuing system in Fig. 2.1 is analyzed for different aspects of its performance. Examples of some relevant performance measures that would result from such analysis are given as [1]:

- Expected number of the service request in the queuing system
- Expected number of requests in the queue
- Traffic carried by the servers
- Lost traffic
- Probability of request blocking
- Average waiting time
- Average time spend in the queuing system,
- Server utilization

Calculation of each of the above performance measures is not a trivial task since it requires a thorough queuing system description. In general, some assumptions need to be made regarding the behavior of the user population, and distribution of the interarrival and service times. The accuracy of the assumptions will limit the accuracy of the mathematical model and hence, the accuracy of the obtained results. Since the performance of the queue changes drastically as a function of adopted assumptions, analysis of a general queuing system is a challenging task. For that reason the queuing systems are divided into several classes and the analysis of each class is performed independently. A method for the queuing system classification will be described in section 2.4.
2.2 Poisson Process of Random Arrivals

As previously discussed one of the most important assumptions regarding the queuing system is the distribution of the service request interarrival times. The interarrival times are property of the user population, and in general, they depend on many factors. For example, in cellular systems the call origination process is a function of the habits of mobile phone users, their lifestyle, occupation, mobility pattern and so on. A similar situation arises in other queuing systems as well. However, extensive observation and measurements have revealed that in many systems the service requests assume behavior of a Poisson process. Having in mind large variability between different queuing systems, this is a remarkable result. For that reason, in this section we provide a brief description of the Poisson process. Many practical methods used in cellular system traffic engineering that are presented in the subsequent sections will be based on the assumption of Poisson service request arrivals.

Consider a stochastic process that provides a count of a certain random event in a given time interval starting from some conveniently chosen origin. Let this process be described as a function of time $N(t)$. For any particular realization the function $N(t)$ will be a "staircase" like function gradually stepping through the positive integers. A process of such nature is commonly referred to as the counting process and it can be formally defined as follows [1]:

**Definition 2.1.** A stochastic process $N(t)$ constitutes a counting process if the following conditions are satisfied:

1. $N(0) = 0$
2. $N(t)$ assumes only nonnegative integer values
3. $t_1 < t_2$ implies that $N(t_1) \leq N(t_2)$, i.e. $N(t)$ is non-decreasing integer function, and
4. $N(t_2) - N(t_1)$ is the number of random events that have occurred after $t_1$ but not later than $t_2$, that is in the interval $(t_1, t_2]$.

An example of the counting process realization is shown in Fig. 2.6. From the conditions that are given in Definition 2.1, and the graphical representation in Fig. 2.6, we see that the counting process can be used to model the service request arrivals in a queuing system. In other words, the graph of the function $N(t)$ shown in Fig. 2.6 may be seen as a count of the number of service requests that have arrived in the time interval $(0, t]$. 
The Poisson process is a counting process that satisfies some additional requirements. These requirements are given as:

1. For every two non-overlapping time intervals \((t_1, t_2]\) and \((t_3, t_4]\) the number of the events are independent random variables. In other words \(N(t_2) - N(t_1)\) is independent from \(N(t_4) - N(t_3)\). Therefore, the Poisson process is a counting process with independent increments.

2. Distribution of events in any given interval depends only on the length of the interval and is independent from the actual time of its beginning. In other words, the Poisson process has stationary increments.

3. The probability that exactly one event occurs in a time interval of length \(h\) is given by

\[
P[N(h) = 1] = \lambda h + o(h),
\]

where \(\lambda\) is a constant.

4. The probability that more than one event occurs within the time interval of duration \(h\) is given by

\[
P[N(h) > 1] = o(h)
\]
In the requirements 3 and 4, symbol \( o(h) \) indicates a function that tends towards zero faster than \( h \) itself. In other words, as \( h \) becomes smaller, the effects of \( o(h) \) can be neglected.

In summary, in a Poisson process, the events occur one-at-the-time and at a constant rate equal to \( \lambda \). In addition, the process stays independent of the beginning of the observation time. Finally, the Poisson process "has no memory". The distribution of events in a given interval does not depend on the distribution in any previous non-overlapping interval, nor will it impact the distribution of events in any future non-overlapping interval.

There are several important properties of the Poisson process that can be derived from its definition. The two most important ones are given as follows.

**Property 1.** Let \( N(t) \) be a Poisson process with the parameter \( \lambda \). The random variable describing the number of events in any given interval of length \( t \) is given as

\[
P(Y = k) = \exp(-\lambda t) (\frac{\lambda t}{k!})^k
\]  

(2.7)

**Property 2.** Let \( N(t) \) be a Poisson process with the parameter \( \lambda \). The interarrival time between events is an exponentially distributed random variable with mean given as \( 1/\lambda \). In other words the probability density function of the interarrival times is given as

\[
pdf(\tau) = \lambda \exp(-\lambda \tau)
\]  

(2.8)

Proof of the above two properties can be found in [1-3]. Here, we provide some examples that will illustrate the use of (2.7) and (2.8).

**Example 2.1.** Consider a Poisson process with the parameter \( \lambda \). Calculate the average time between two consecutive events and the average rate of the event occurrence.

The average time between events can be calculated as

\[
\bar{\tau} = E\{\tau\} = \int_0^{+\infty} \tau \lambda \exp(-\lambda \tau) d\tau = \frac{1}{\lambda}
\]

The average rate is given as \( \bar{\tau} = 1/E\{\tau\} = \lambda \). Therefore, the distribution parameter \( \lambda \) can be interpreted as the average rate of the event arrivals.

**Example 2.2.** Assume that the number of call arrivals in a given cell of a cellular system may be modeled as a Poisson process with an average rate of 10 calls per minute. What is the average interarrival time? What is the probability of receiving more than 15 calls per minute?

Using the results of the previous example, we have
The probability of receiving more than 15 calls can be found using

\[ \Pr\{Y > 15\} = \Pr\{Y = 16\} + \Pr\{Y = 17\} + \cdots \]

Using (2.7), we have

\[ \Pr\{Y > 15\} = 1 - \Pr\{Y \leq 15\} = 1 - \sum_{k=0}^{15} \frac{(\lambda t)^k}{k!} \exp(-\lambda t) \]

Substituting the numerical values

\[ \Pr\{Y > 15\} = 1 - \sum_{k=0}^{15} \frac{(10 \cdot 1)^k}{k!} \exp(-10) = 0.0487 \]

As can be seen, although the average number of calls per minute is 10, about 5% of the time, the actual number of calls placed within one minute will be more than 15. Therefore, to assure that most of the calls are served, the number of channels at the site has to be larger than 10.

**Example 2.3.** Consider the measurements in Table 2.1. The measurements report the number of jobs sent to a printer server on a minute by minute basis for a period of one hour. Determine if the process can be modeled as the Poisson process and if that is the case, estimate the average rate of service request arrivals.

**Table 2.1.** Measurements reporting the number of jobs serviced by a printer server

<table>
<thead>
<tr>
<th>time</th>
<th># jobs</th>
<th>time</th>
<th># jobs</th>
<th>time</th>
<th># jobs</th>
<th>time</th>
<th># jobs</th>
<th>time</th>
<th># jobs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>13</td>
<td>2</td>
<td>25</td>
<td>2</td>
<td>37</td>
<td>6</td>
<td>49</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>14</td>
<td>5</td>
<td>26</td>
<td>8</td>
<td>38</td>
<td>5</td>
<td>50</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>15</td>
<td>8</td>
<td>27</td>
<td>5</td>
<td>39</td>
<td>6</td>
<td>51</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>16</td>
<td>5</td>
<td>28</td>
<td>4</td>
<td>40</td>
<td>1</td>
<td>52</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>17</td>
<td>10</td>
<td>29</td>
<td>3</td>
<td>41</td>
<td>5</td>
<td>53</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>18</td>
<td>7</td>
<td>30</td>
<td>1</td>
<td>42</td>
<td>8</td>
<td>54</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>9</td>
<td>19</td>
<td>3</td>
<td>31</td>
<td>6</td>
<td>43</td>
<td>3</td>
<td>55</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
<td>20</td>
<td>6</td>
<td>32</td>
<td>4</td>
<td>44</td>
<td>7</td>
<td>56</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>7</td>
<td>21</td>
<td>6</td>
<td>33</td>
<td>4</td>
<td>45</td>
<td>10</td>
<td>57</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>22</td>
<td>6</td>
<td>34</td>
<td>3</td>
<td>46</td>
<td>8</td>
<td>58</td>
<td>12</td>
</tr>
<tr>
<td>11</td>
<td>7</td>
<td>23</td>
<td>5</td>
<td>35</td>
<td>4</td>
<td>47</td>
<td>2</td>
<td>59</td>
<td>6</td>
</tr>
<tr>
<td>12</td>
<td>5</td>
<td>24</td>
<td>4</td>
<td>36</td>
<td>5</td>
<td>48</td>
<td>2</td>
<td>60</td>
<td>4</td>
</tr>
</tbody>
</table>
First we will estimate the average rate of job arrival

\[ \bar{\lambda} = \frac{1}{60} (6 + 6 + 7 + 5 + \cdots + 6 + 4) = 5.2833 \text{ arrivals/min} \]

Using Table 2.1, we can calculate the normalized frequency of occurrence which can be used as an estimate of the probability mass function of a discrete process. The normalized frequency of occurrence is calculated in accordance with:

\[
\text{Normalized Frequency} = \frac{1}{60} \times \text{Number of Occurrences}
\]

For example, the normalized frequency of occurrence for three jobs within a minute is given by

\[ F_3 = \frac{5}{60} = 0.0833 \]

Figure 2.7 shows the plot of the relative frequency of occurrence derived from the measurements in Table 2.1. On the same plot we show the values for the probability mass function of an ideal Poisson process that has the same mean rate of arrivals. As evident, the difference is relatively small and for practical traffic dimensioning of this system we may assume that the process of service request arrivals is a Poisson process.

**Figure 2.7.** Comparison of the frequency of occurrence plot and the PMF of the ideal Poisson process for data in Example 2.3
2.3 Birth and Death Processes

In the previous section we considered the Poisson process and saw that it can be used to describe the arrivals of service requests in many cases of great practical interest. In a practical queuing system, the request arrivals result in resource allocation and eventually the users get served and leave the queue. It is customary to view this process as a member of a wider class of stochastic processes that are commonly referred to as the birth and death. Within this framework, every incoming request is regarded as a birth and every user that, after being served, leaves the system is regarded as a death. For the Poisson process the average birth rate is specified by the distribution parameter \( \lambda \). The birth rate can change as a function of the state of the queuing system. However, we can still say that in a short time interval \( h \), the probability of a single birth is equal to \( \lambda_n h + o(h) \), where subscript \( n \) indicates one of the system states. Likewise, it is reasonable to assume that in a short time interval \( h \), the number of users leaving the system is equal to \( \mu_n h + o(h) \), where \( \mu_n \) indicates the average death rate, and index \( n \) references the state of the queuing system.

The birth and death process is frequently used as a mathematical model of a queuing system and in this section we provide its description. The framework of the birth and death process will allow us to derive some results that describe the behavior of the queuing systems in general.

The formal definition of the birth and death process is given as [1]:

**Definition 2.2.** Consider a stochastic process \( N(t) \) that is continuous in time but has a discrete state space \( \Omega = \{0,1,2,\cdots\} \). Suppose that this process describes a physical system that is in state \( E_n, n = 0,1,2,\cdots \) at time \( t \), if and only if \( N(t) = n \). Then the system is described by the birth-and-death process if there exist nonnegative birth rates \( \lambda_n, n = 0,1,2,\cdots \), and nonnegative death rates \( \mu_n, n = 0,1,2,\cdots \), such that the following postulates (sometimes called nearest neighbor assumptions) are true:

1. State changes are only allowed between state \( E_n \) to state \( E_{n+1} \) or from state \( E_n \) to \( E_{n-1} \) if \( n \geq 1 \), but from state \( E_0 \) to state \( E_1 \) only.
2. If at time \( t \) the system is in state \( E_n \), the probability that between time \( t \) and time \( t+h \) a transition from state \( E_n \) to state \( E_{n+1} \) occurs equals \( \lambda_n h + o(h) \), and the probability of transition from \( E_n \) to \( E_{n-1} \) is \( \mu_n h + o(h) \) (if \( n \geq 1 \)).
3. The probability that in time interval from \( t \) to \( t+h \) more than one transition occurs is \( o(h) \).

Before we proceed, let us provide some examples of processes that can be classified as birth and death. As a first example, let us consider a stochastic process modeling the population in a closed system. The assumption that the system is closed is necessary to assure that the only mechanisms that the population can change are death and birth. Referring back to Definition 1 we can identify the following analogies:
1. State of the system $E_n$ represents the total population count at a given time $t$. One should note that although we can determine the population at an arbitrary time (that is the process is continuous in time), the actual values of $E_n$ can be only nonnegative integers, i.e. 0,1,2, … etc. Therefore, the process is continuous in time, but descrete in state space.

2. At any given state $E_n$, the population increases at the rate of $\lambda_n$ and decreases at the rate of $\mu_n$. Obviously, the population experiences growth if $\lambda_n > \mu_n$, and it is subject to decline if $\mu_n > \lambda_n$. In actual physical systems, no deaths or births can occur if the system is in state $E_0$. However, definition of the birth-and-death process allows for a nonzero birth rate even when the system is in state $E_0$.

3. Let’s assume that we count the population at times $t$ and $t+h$. If the time increment $h$ is kept small we expect the probability of birth to be given as the product of birth rate and the given time increment, i.e., $h\lambda_n$. Similarly, the probability of death is given as $h\mu_n$. Given that $h$ is infinitesimally small, the probability of both death and birth occurring within such a small time increment can be neglected, that is assumed as essentially zero.

4. As a final note, we point out that in general, birth and death rates are a function of the current population count. In other words, if the population grows, both the rate of birth and the rate of death can be expected to grow. Likewise if the population plummets, the rate of birth and the rate of death decrease as well. However, if the population is very large, the impact of the actual population count on the birth and death rates becomes smaller. In a boundary case for infinite population we would expect the rates of death and birth to remain constant.

As a second example, let us examine the modeling of traffic served in a cell of a cellular communication system.

1. The state of system $E_n$ represents the total number of users that are being served by a given cell. Unlike the previous example in which the set of possible states encompasses all positive integers, the possible states in this case are limited by the number of available resources at the cell site. In other words, $E_n \in \{0,1,2,\ldots,C\}$, where $C$ is the number of trunks (that is, voice channels) that are available at the site.

2. The process of birth is analogous to a new user trying to set up a call. Therefore, the birth rate $\lambda_n$ gives the rate at which the users request the service. In a similar way the death corresponds to a user that has completed the call and released the voice channel.

2.3.1 State Diagram Representation of Birth and Death Process

A useful visualization of the birth and death process is provided through the state transition rate diagram. An example of such a diagram is given in Fig. 2.8.
The number inside the circle indicates the state of the system. For example, in a cellular system this would be the number of users serviced by a given site. Values \( \lambda_i \) indicate the birth rates at each of the system states. Similarly, values \( \mu_i \) represent the death rates. The state diagram allows only “the nearest neighbor” transitions and only the birth transition is allowed from state zero.

State diagram representation of the birth and death process will be frequently used for analyses presented in subsequent sections. For that reason, we derive differential-difference equations for

\[
(P_n(t) = P_n \{ N(t) = n \}) ,
\]

that is, the probability that the system is in state \( E_n \) at time \( t \). Note that the derivation presented here is generalized, and as such, it is valid for any system that can be described using the birth and death processes.

If \( n \geq 1 \), the probability \( P_n(t+h) \) that at the time \( t+h \) system will be in the state \( E_n \) has four components listed as follows:

1. **The system was in state \( E_n \) at time \( t \) and no births or deaths have occurred.** Knowing that the probability of birth is \( \lambda_n h + o(h) \) and the probability of death is \( \mu_n h + o(h) \), this component can be expressed as:

\[
P_n^{(1)}(t+h) = P_n(t)[1 - \lambda_n h + o(h)][1 - \mu_n h + o(h)] = P_n(t)[1 - \lambda_n h - \mu_n h] + o(h) \tag{2.9}
\]

2. **The system was in state \( E_{n-1} \) at time \( t \) and a birth has occurred.** The probability of this event is given as:

\[
P_n^{(2)}(t+h) = P_{n-1}(t) \lambda_{n-1} h + o(h) \tag{2.10}
\]

3. **The system was in state \( E_{n+1} \) and a death has occurred.** The probability of this event is given as:

\[
P_n^{(3)}(t+h) = P_{n+1}(t) \mu_{n+1} h + o(h) \tag{2.11}
\]

4. **Two or more transitions have occurred.** By the properties of the birth and death process stated in Definition 2, this probability is:
\[ P_n^{(4)}(t + h) = o(h) \]  \hspace{1cm} (2.12)

From (2.9) through (2.12) we have:

\[ P_n(t + h) = \sum_{i=1}^{n} P_n^{(i)} + [1 - \lambda_n h - \mu_n h] P_n(t) + \lambda_{n-1} h P_{n-1}(t) + \mu_{n+1} h P_{n+1}(t) + o(h) \]  \hspace{1cm} (2.13)

or

\[ \frac{P_n(t + 1) - P_n(t)}{h} = -(\lambda_n + \mu_n) P_n(t) + \lambda_{n-1} P_{n-1}(t) + \mu_{n+1} P_{n+1}(t) + \frac{o(h)}{h} \]  \hspace{1cm} (2.14)

By letting \( h \to 0 \), (2.14) reduces to:

\[ \frac{dP_n(t)}{dt} = -(\lambda_n + \mu_n) P_n(t) + \lambda_{n-1} P_{n-1}(t) + \mu_{n+1} P_{n+1} \]  \hspace{1cm} (2.15)

Equation (2.15) is valid for \( n \geq 1 \). For \( n = 0 \), following the same procedure one obtains:

\[ \frac{dP_0(t)}{dt} = -\lambda_0 P_0(t) + \mu_1 P_1(t) \]  \hspace{1cm} (2.16)

If the initial state of the system is \( E_i \), then initial conditions are given as:

\[ P_i(0) = 1, \text{ and } P_j(0) = 0, \text{ for } j \neq i \]  \hspace{1cm} (2.17)

From (2.15) and (2.16), we see that the birth and death process can be described using an infinite set of differential equations, with initial conditions given in (2.17). Although it can be proven that the solution of these equations exists under very general circumstances [1], it can be rarely obtained in an analytical form.

The steady state solution of (2.16) and (2.17) are of a special practical interest. The steady state solution assumes that a sufficient time has elapsed and that the system has reached statistical equilibrium. In a steady state, all system state probabilities (\( P_n(t) \) values), become constant and hence the derivatives on the left-hand sides of (2.15) and (2.16) are equal to zero. Therefore, under the steady state assumptions

\[ 0 = \lambda_{n-1} P_{n-1} + \mu_{n+1} P_{n+1} - (\lambda_n + \mu_n) P_n, \text{ for } n \geq 1 \]  \hspace{1cm} (2.18)

and

\[ 0 = \mu_1 P_1 - \lambda_0 P_0, \text{ for } n = 0 \]  \hspace{1cm} (2.19)

Equation (2.19) can be rewritten as
\[ p_1 = \frac{\lambda_0}{\mu_1} p_0 \]  \hspace{1cm} (2.20)

Also, (2.18) can be rearranged in the form

\[ \mu_{n+1} p_{n+1} - \lambda_n p_n = \mu_n p_n - \lambda_{n-1} p_{n-1} \]  \hspace{1cm} (2.21)

Since (2.21) is valid for every \( n \), using (2.20) we can conclude that

\[ p_{n+1} = \frac{\lambda_n}{\mu_{n+1}} p_n \text{ for } n = 0,1,2,\cdots \]  \hspace{1cm} (2.22)

Using (2.22) we can compute

\[ p_1 = C_1 p_0 = \frac{\lambda_0}{\mu_1} p_0 \]  \hspace{1cm} (2.23)
\[ p_2 = C_2 p_0 = \frac{\lambda_1}{\mu_2} p_1 = \frac{\lambda_1 \lambda_0}{\mu_2 \mu_1} p_0 \]  \hspace{1cm} (2.24)
\[ p_3 = C_3 p_0 = \frac{\lambda_2}{\mu_3} p_2 = \frac{\lambda_2 \lambda_1 \lambda_0}{\mu_3 \mu_2 \mu_1} p_0 \]  \hspace{1cm} (2.25)

In general, we have

\[ p_n = C_n p_0 = \frac{\lambda_{n-1} \lambda_n \cdots \lambda_0}{\mu_n \mu_{n-1} \cdots \mu_1} p_0 \]  \hspace{1cm} (2.26)

Since the sum of all state probabilities has to be equal to 1,

\[ p_0 \left( 1 + \frac{\lambda_0}{\mu_1} + \frac{\lambda_1 \lambda_0}{\mu_2 \mu_1} + \cdots + \frac{\lambda_{n-1} \lambda_n \cdots \lambda_0}{\mu_n \mu_{n-1} \cdots \mu_1} + \cdots \right) = p_0 S = 1 \]  \hspace{1cm} (2.27)

Finally, as a summary, we have

\[ p_0 = P_r \{ N(t) = 0 \} = \frac{1}{S} \]  \hspace{1cm} (2.28)

and

\[ p_n = P_r \{ N(t) = n \} = \frac{C_n}{S} \]  \hspace{1cm} (2.29)

where

\[ C_n = \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n}, \text{ and } S = 1 + C_1 + C_2 + \cdots + C_n + \cdots \]  \hspace{1cm} (2.30)
From (2.28) through (2.30) we see that the birth and death process has a steady state solution if the sum $S$ converges. In such a case, there is a finite probability of a system occupying state zero. This would mean that from time to time the system “catches up” and manages to serve all users. On the other hand, if $S$ diverges, this indicates of an unstable system in which births are occurring at faster rates than deaths. For practical applications of the birth and death processes, we will assume that the system is not unstable, that a steady state exists, and that the state probabilities are constant and given by (2.29).

### 2.3.2 Little's Formula

Little's formula is a simple but very important equation that applies to any system in equilibrium in which customers arrive, spend some time and then depart. The formula is given by

$$L = \lambda W$$

(2.31)

where $L$ is the average number of customers in the system, $\lambda$ is the average rate of customer arrivals, and $W$ is the average time that customers spend in the system. The proof of (2.31) is relatively complex and is beyond the scope of this document. To get an intuitive understanding of Little's formula, consider a system with a single server and an infinite queue. If the average service time is $W$, the number of users that arrive while one user is being served is $\lambda W$. Since the resource is occupied, these users are placed in queue and the state of the system is described by (2.31). The most important aspect of (2.31) is its universal applicability, therefore it is used frequently throughout this document.

**Example 2.4.** As an illustration of a birth and death process, consider a queuing system having only one server. Assume that that the service request arrivals can be accurately modeled as a Poisson process with an average rate of $\lambda = 1 \text{min}^{-1}$, and that the average time required to service one request is given by $W_s = 0.5 \text{ min}$. Also assume an infinite queue capacity with a FIFO queuing discipline. This kind of queuing system can be used to model many practical "real life" scenarios. For example, it can be used to model the queue formed at the printer server, or the queue formed in a supermarket with only one cash register. Estimate the probability that exactly $n$ users are in the queue, an average number of users in the queuing system, and the average time that users spend in this queuing system.

First, we estimate the average death rate, that is, the average rate at which the users would be leaving the system providing that the server has no idle time. This rate is estimated as:

$$\mu = \frac{1}{W_s} = \frac{1}{0.5} = 2 \text{ min}^{-1}$$

(2.32)

Using (2.29) and (2.30) we have

$$C_n = \left(\frac{\lambda}{\mu}\right)^n = \left(\frac{1}{2}\right)^n = \frac{1}{2^n}$$

(2.33)

and
\[ S = 1 + C_1 + C_2 + \cdots = 1 + \frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^3} + \cdots = \frac{1}{1 - \frac{1}{2}} = 2 \quad (2.34) \]

Therefore, the probability of having exactly \( n \) users within the queue is given by:

\[ p_n = \frac{C_n}{S} = \frac{1/2^n}{2} = \frac{1}{2^{n+1}} \quad (2.35) \]

The average number of users in the system can be calculated as

\[ L = \sum_{n=0}^{\infty} n \cdot p_n = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{4} + 2 \cdot \frac{1}{8} + \cdots + n \cdot \frac{1}{2^{n+1}} + \cdots \quad (2.36) \]

Multiplying both sides in (2.36) with \( 1/2 \) we obtain

\[ \frac{1}{2} L = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{8} + 2 \cdot \frac{1}{16} + \cdots + (n-1) \cdot \frac{1}{2^{n+1}} + \cdots \quad (2.37) \]

Subtracting (2.37) from (2.38)

\[ \frac{1}{2} L = \frac{1}{4} + \frac{1}{8} + \frac{1}{2^{n+1}} + \cdots = \frac{1}{4} \left( 1 + \frac{1}{2} + \frac{1}{2^2} + \cdots \right) = \frac{1}{4} \cdot \frac{1}{1 - \frac{1}{2}} = \frac{1}{2} \quad (2.38) \]

Therefore the average number of users in the queuing system is given by

\[ L = 1 \]

The average time that users spend in the system can be calculated using Little's formula as

\[ W = \frac{L}{\lambda} = \frac{1}{1} = 1 \text{ min} \quad (2.39) \]

### 2.4 Kendall's Notation

Kendall's notation is frequently used for describing queuing systems of various properties. This is a shorthand notation in the following form

\[ A/B/C/K/m/Z \]

where the interpretation of individual terms is as follows:

- \( A \) - distribution of the interarrival times
- \( B \) - distribution of the service times
- \( C \) - number of servers within the service facility
- \( K \) - maximum number of users within the queuing system
Within Kendall’s notation for the description of the arrival process and service times, the following symbols are used:

- GI - general independent arrival/service times
- G - general (not necessarily independent) arrival/service times
- H_k - k-stage hyperexponential distribution
- E_k - Erlang-k distribution
- M - exponential distribution (Poisson process)
- D - constant interarrival/service times
- U - uniform distribution

As an illustration, consider the queuing system described in Example 2.4. In Kendall’s notation, this queue can be described as follows. Since the arrivals are modeled using the Poisson process \( A = M \). Due to exponentially distributed service times \( B = M \). Since there is only one server, \( C=1 \). Both the queue and the population are of an infinite size and therefore \( K=\infty \) and \( m = \infty \). As the queuing discipline is First-In-First-Out, \( Z = FIFO \). Therefore, Kendall’s notation for the queuing system in Example 2.4 is \( M/M/1/\infty/\infty/FIFO \). Very often, if the queue and population are infinite and the queuing service discipline is FIFO, the last three designators of the notation are omitted. In this example, the notation would reduce to \( M/M/1 \).

### 2.5 Examples

In this section we illustrate the application of the queuing theory in the analysis of some commonly encountered queuing systems. Two examples will be presented. The first example analyzes the problem of connecting two workstations to a central server. The second example shows the applicability of the queuing theory in the design of reliable microwave communication links.

**Example 2.5.** Consider a problem illustrated in Fig. 2.9.

Two work stations need to be connected to a single server and we examine two possible configurations that can be used to accomplish the task. In the first configuration, the connection is achieved by using two separate lines. The second configuration uses one line with a bandwidth that is two times larger. Let us assume that each workstation generates \( \lambda \) messages per second and that the average for the message delivery is given as \( 1/\mu \) for the individual lines and \( 1/(2\mu) \) for the line with the larger bandwidth. Both configurations in Fig. 2.9 can be modeled using the theory developed in previous sections. We will examine some performance matrix as they are observed from individual workstations.
Figure 2.9. Two different configurations examined in Example 2.5

**Configuration 1.** In configuration 1, we essentially have two separate $M/M/1$ queuing systems with the same performance. Using the results of the birth and death process analysis (c.f. Section 2.3), the probability of having exactly $n$ messages in a transmission line (or associated buffer), is given by

$$p_n = \frac{C_n}{S},$$

(2.40)

where

$$C_n = \left( \frac{\lambda}{\mu} \right)^n = \rho^n, \quad \rho = \frac{\lambda}{\mu}$$

(2.41)

and

$$S = 1 + \frac{\lambda}{\mu} + \left( \frac{\lambda}{\mu} \right)^2 + \left( \frac{\lambda}{\mu} \right)^3 + \cdots = \frac{1}{1 - \lambda/\mu} = \frac{1}{1 - \rho}$$

(2.42)

Therefore,

$$p_n = (1 - \rho)\rho^n$$

(2.43)

The average number of messages within each of the transmission lines is given by

$$\bar{n} = \sum_{n=0}^{\infty} np_n = 0 \cdot (1 - \rho)\rho^0 + 1 \cdot (1 - \rho)\rho + 2 \cdot (1 - \rho)\rho^2 + \cdots = \frac{\rho}{1 - \rho}$$

(2.44)

Using Little's formula, the average time required for the message delivery is given by
RF Engineering Continuing Education
Introduction to Traffic Planning

\[ W_1 = \frac{n_1}{\lambda} = \frac{\rho}{(1-\rho)\lambda} = \frac{1}{\mu - \lambda} \]

(2.45)

Therefore, in the first configuration each of the workstations experiences an average throughput

\[ R_1 = \frac{1}{W_1} = \mu - \lambda \]

(2.46)

**Configuration 2.** Configuration 2 can be seen as one \( M/M/1 \) queuing system with a birth rate of \( 2\lambda \) and a death rate of \( 2\mu \). Following the same approach as in the case of configuration 1, we obtain the following results

\[ p_n = (1-\rho)\rho^n = \left(1-\frac{2\lambda}{2\mu}\right)\left(\frac{2\lambda}{2\mu}\right)^n = \left(1-\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n \]

(2.47)

\[ n_2 = \frac{\rho}{1-\rho} = \frac{\lambda/\mu}{1-\lambda/\mu} \]

(2.48)

\[ W_2 = n_2 \frac{\lambda/\mu}{2\lambda(1-\lambda/\mu)} = \frac{1}{2(\mu - \lambda)} \]

(2.49)

and

\[ R_2 = \frac{1}{W_2} = 2(\mu - \lambda) = 2R_1 \]

(2.50)

Therefore, the second configuration is two times more efficient than the first one.

**Example 2.6.** In this example we illustrate the impact of the link diversity on the reliability of a microwave connection. Consider a microwave link with a hot standby [4]. Let us assume that a mean time between a single link failure is given as \( T_f \). When a link fails (either the main one or the hot standby), the mean time to repair is given by \( T_r \). If we assume the same reliability of the main link and the hot standby, let us estimate the reliability improvement over a system without the link diversity.

The microwave link in this example can be modeled as a birth and death process with just three states and the state diagram shown in Fig. 2.10.

![State diagram for the microwave system in Example 2.6](image)

Figure 2.10. State diagram for the microwave system in Example 2.6

30

Revision 2.0
The state of the system corresponds to the number of non-working links. In other words, state 0 corresponds to the case when both the main link and its hot standby are operational; state 1 corresponds to the case when one of the links fails; and state 2 corresponds to failure of both the main link and the hot standby. The birth and death rates are indicated in Fig. 2.10, where

\[ \lambda = \frac{1}{T_f}, \]  

and

\[ \mu = \frac{1}{T_r}. \]  

To calculate the mean time between the failure for the system with the link diversity we use the diagram in Fig. 2.10 to estimate the steady state rate at which the system reaches state 2. From Fig. 2.10, this rate can be calculated as

\[ \lambda_{2f} = \lambda \cdot p_1 \]  

where

\[ p_1 = \frac{C_1}{S} = \frac{2\lambda/\mu}{1 + 2\lambda/\mu} \]  

Therefore,

\[ \lambda_{2f} = \frac{2\lambda/\mu}{1 + \lambda/\mu} \]  

and the time between the failures becomes

\[ T_{2f} = \frac{1}{\lambda_{2f}} = \frac{1 + 2\lambda/\mu}{2\lambda^2/\mu} = \frac{1 + 2T_r/T_f}{2T_r/T_f^2} \]  

To illustrate the resulting improvement, let us consider the following numerical data. The average time between link failure is \( T_f = 4000 \) hours and the average repair time is \( T_r = 24 \) hours. When the link diversity is used, the average time between failures becomes

\[ T_{2f} = \frac{1 + 2 \cdot \frac{24}{4000}}{2 \cdot \frac{24}{4000^2}} = 337,333 \text{ [hours]} \]

which is a significant improvement.
3 Traffic Planning for Circuit Switched Voice Services

A dominant type of traffic in cellular communication systems of the first and the second generation is voice traffic. In all cellular standards, voice service is implemented as a circuit switched service. As explained in Section 1, the circuit switched nature of the voice service implies that a network allocates a cell site communication resource for the entire duration of the call. Strictly speaking, in cellular communication this assumption does not hold. Due to handoff, the occupancy of the cell site resource is generally shorter than the call duration. However, in practical traffic dimensioning, the effect of the handoff is usually neglected. This simplification can be qualitatively justified as follows:

1. In the state of traffic equilibrium there is approximately the same number of calls leaving and entering the cell site coverage area. Therefore, the average traffic load is not modified by the handoff.
2. The overhead traffic that results from the handoff processing is usually handled by a separate group of communication resources referred to as the control channels. Therefore, the actual handoff processing does not increase the traffic volume on the voice channels.
3. The service time for the voice call is usually assumed to be distributed exponentially. Due to so-called "memoryless" property of the exponential distribution, the incoming calls can be thought of as the continuation of the outgoing calls. Although this is truly not the case, we may assume that all calls terminate in the same cell where they originate.

3.1 Common Descriptors of the Circuit Switched Voice Service

There are several common terms used in everyday traffic engineering of cellular systems. In this section we define and illustrate their typical use.

3.1.1 Busy Hour

The volume of voice traffic carried by a cellular network is continuously changing with respect to time. Most systems will experience a low level of usage during the afternoon and early evening hours with peak levels of usage during the morning rush hour, lunch hour, and evening rush hour (see Fig. 3.1). The uninterrupted period of sixty minutes during which the traffic is at a maximum is known as the busy hour. The amount of traffic experienced during the busy hour is generally used as the basis for traffic calculations and resource dimensioning. Shorter peaks (less than sixty minutes) of traffic usage can be experienced due to special events such as traffic jams, automobile accidents or poor weather. However, system dimensioning which would cater to these special events would not be economically viable.

The busy hour can be defined as either fixed or bouncing. The fixed busy hour is a set period of 60 minutes that does not change from day to day. When the fixed busy hour is used, the data collection process is limited to one hour per day and the requirements for storage memory space are reduced. The bouncing busy hour is determined for each day independently. To determine
the bouncing busy hour, the switch has to collect traffic data over the entire 24 hours. Although twenty four hour data collection results in relatively large volumes of measured data, most contemporary switches are capable of providing reports based on the bouncing busy hour.

![Typical daily traffic usage in a cellular system](image)

**Figure 3.1.** Typical daily traffic usage in a cellular system

### 3.1.2 QoS parameters

In general, a definition of the quality of service depends largely on the type of communication service provided by the network. For the circuit switched voice, the most common indicator of the service quality is *probability of blocking*. The probability of blocking is defined as probability that a service request is denied due to all cell site trunks being occupied. In common engineering practice, the probability of blocking is commonly referred to as the *Grade Of Service* (GOS), and it can be formally defined as

\[
GOS = 1 - \frac{\text{Traffic Served}}{\text{Traffic Offered}} \tag{3.1}
\]

The GOS definition in (3.1) assumes no queuing. In other words, if the call attempt is made and all resources are occupied, the call is immediately rejected. This is analogous to the situation encountered in landline telephone networks when a user picks up the phone and gets a busy signal even before dialing the number. While in contemporary landline telephone networks this is a fairly rare event, the mobile communication networks are designed with a GOS target between 1 and 2%. In other words, we may expect that in a typical cellular network, 1 out of 50 calls may be blocked.
Some of the services provided in the circuit switched mode allow queuing of individual calls. For example, dispatch voice services as well as myriad of circuit switched data services may delay transmission until the channel becomes available. In such scenarios, GOS does not represent a valid measure of system quality and alternative indicators are used. The most common ones are probability of delay, probability of delay exceeding given limit, an average number of calls in the queue, maximum number of calls in the queue. Some of these parameters will be more thoroughly addressed in Section 3.2.2.

### 3.2 Trunking Models for Circuit Switched Services

In this section we will analyze two trunking models that are commonly used in traffic dimensioning of wireless cellular networks. The main difference between the two models is in the treatment of the service requests that occur while all servers are occupied. In the case of the Erlang B trunking model, requests that cannot find an available server are cleared from the system. This regime of operation is sometimes referred to as the "lost calls cleared" regime. The Erlang C model represents the other extreme. In the case of Erlang C, service requests made during time intervals when all servers are busy are placed in a queue of an infinite capacity. Therefore, no calls are lost, and this regime of operation is referred to as the "lost calls held" regime. There are numerous systems that can be closely modeled using either Erlang B or Erlang C formulas. However, in analyzing a particular traffic problem, an engineer needs to make sure that the assumptions of either Erlang B or Erlang C models are still valid.

#### 3.2.1 Erlang B Formula

The Erlang B formula is the most commonly used formula for the GOS calculation in a cellular system. The formula is applied on a per cell basis assuming that each cell site can be modeled as a queuing system of a type $M/M/C/C$. In other words, the assumptions used in derivation of the Erlang B formula are as follows:

1. Call arrival process is a Poisson process.
2. Service time (i.e. call holding time) is exponentially distributed.
3. There are $C$ identical servers within the service facility. For voice service this translates into $C$ voice channels available at the site.
4. There is no queuing in the system. Any call attempt that is made while all trunks are occupied is cleared from the system and is considered lost.

A commonly adopted QoS parameter in $M/M/C/C$ systems is the probability of call blocking or GOS. This is the probability that an incoming call will find all channels at the site occupied. The Erlang B loss formula is used for calculation of the GOS in $M/M/C/C$ systems.

A state transition diagram for a $M/M/C/C$ system is presented in Fig. 3.2.
Figure 3.2. State diagram for $M/M/C/C$ queuing system

From the state diagram we have

$$C_n = \frac{\lambda^n}{n! \mu^n} = \frac{a^n}{n!} \quad (3.2)$$

and

$$S = 1 + \frac{\lambda}{\mu} + \frac{\lambda^2}{2\mu^2} + \frac{\lambda^3}{3!\mu^3} + \cdots + \frac{\lambda^C}{C!\mu^C} = \sum_{k=0}^{C} \frac{(\lambda/\mu)^k}{k!} = \sum_{k=0}^{C} \frac{a^k}{k!} \quad (3.3)$$

where $a = \lambda/\mu$ is the offered traffic. Therefore, the probability of finding exactly $n$ occupied trunks is given by

$$p_n = \frac{C_n}{S} = \frac{a^n/n!}{\sum_{k=0}^{C} \frac{a^k}{k!}} \quad (3.4)$$

An incoming call will be blocked (i.e., a call request will be denied), if all of the trunks are occupied. The probability of such an event is given by

$$B[a, C] = \frac{a^C/C!}{\sum_{k=0}^{C} \frac{a^k}{k!}} \quad (3.5)$$

Equation (3.5) is the Erlang B loss formula.

To illustrate the use of (3.5), consider the following example.

**Example 3.1.** A cell site has 5 AMPS radios. The average rate of call origination within the cell site coverage area is 60 calls per hour. If the call holding times are distributed exponentially with an average of 90 seconds, calculate the blocking probability.

The average birth rate is given by: $\lambda = 60 \text{ calls/hour}$

The average death rate is given by: $\mu = \frac{1}{H} = \frac{1}{90/3600} = 40 \text{ calls/hour}$
The offered traffic is:
\[ a = \frac{\lambda}{\mu} = \frac{60}{40} = 1.5 \ E \]

Using (3.5), the blocking probability can be calculated as:
\[
B[a = 1.5, C = 5] = \frac{a^C / C!}{\sum_{k=0}^{C} \frac{a^C / C!}{k!} 1 + \frac{1.5^2}{2} + \frac{1.5^3}{3!} + \frac{1.5^4}{4!} + \frac{1.5^5}{5!}} \approx 0.0142
\]

To facilitate the use of the Erlang B formula in everyday engineering practice, (3.5) is frequently presented either in the form of parametric curves or in the form of a table. Erlang B curves are shown in Fig. 3.3, while Appendix A provides the Erlang B table.

**Figure 3.3.** Representation of the Erlang B blocking formula through a family of curves

To illustrate the use of the Erlang B table, consider the following examples.

**Example 3.2.** Determine the traffic capacity in erlangs for a 30-channel cell such that the GOS will not exceed a) 2% and b) 1% using the Erlang B table provided in Appendix A.

**Part a):** In Appendix A, find the row for \( N = 30 \). Find the intersection of this row with the column for 2% GOS. Read the corresponding traffic capacity to be 21.9 erlangs.
Part b): In Appendix A, find the row for N = 30. Find the intersection of this row with the column for 1% GOS. Read the corresponding traffic capacity to be 20.3 erlangs.

Example 3.3. Determine the number of voice channels required to support 20 erlangs (720 CCS) at a GOS of a) 2% and b) 1% using the Erlang B table provided.

Part a): In Appendix A, find the column for a GOS of 2% and follow this column down to a traffic capacity value of 20.2 erlangs. Follow this row to the left and read the number of voice channels to be 28.

Part b): In Appendix A, find the column for a GOS of 1% and follow this column down to a traffic capacity value of 20.3 erlangs. Follow this row to the left and read the number of voice channels to be 30.

Example 3.4. Determine the GOS that a 30 voice channel cell will provide with a) 28 erlangs of traffic and b) 20 erlangs of traffic using the Erlang B table provided.

Part a): In Appendix A, find the row for N = 30 and read across until a traffic capacity of 28.1 erlangs is found. Follow this column up and read the GOS to be 10%.

Part b): In Appendix A, find the row for N = 30 and read across until a traffic capacity of 20.3 erlangs is found. Follow this column up and read the GOS to be 1%.

Several performance indicators can be derived for the $M/M/C/C$ queuing systems. The values for some of them, including the ones derived above, are provided in Table 3.1.

Table 3.1. Summary of the performance parameters for the $M/M/C/C$ queuing system

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Symbol</th>
<th>Value (Formula)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Offered traffic</td>
<td>$a$</td>
<td>$a = \lambda / \mu$</td>
</tr>
<tr>
<td>Probability of having no calls at the cell site</td>
<td>$p_0$</td>
<td>$\frac{1}{1 + a + \frac{a^2}{2!} + \cdots + \frac{a^C}{C!}}$</td>
</tr>
<tr>
<td>Probability of having exactly $n$ calls at the cell site</td>
<td>$p_n$</td>
<td>$p_n = \frac{a^n}{n!} p_0$</td>
</tr>
<tr>
<td>Blocking probability</td>
<td>$B[a, C]$</td>
<td>$B[a, C] = \frac{a^C}{C!} p_0$</td>
</tr>
<tr>
<td>Average arrival rate of served calls</td>
<td>$\lambda_a$</td>
<td>$\lambda_a = \lambda(1 - B[a, C])$</td>
</tr>
<tr>
<td>Average number of active calls</td>
<td>$\bar{n}$</td>
<td>$\bar{n} = \lambda_a / \mu$</td>
</tr>
<tr>
<td>Average channel utilization</td>
<td>$\rho$</td>
<td>$\rho = \frac{\lambda_a}{\mu \mu}$</td>
</tr>
</tbody>
</table>
Example 3.5. Consider a GSM cell that implements GPRS service with 2 radios. Let us assume that on the voice side the average rate of the call origination is 300 calls/hour. Assuming that the average call holding time is 120s, calculate the average number of time slots available for the GPRS service.

The average birth rate: \( \lambda = 300 \text{ calls/hour} \)

The average death rate (per time slot): \( \mu = \frac{1}{120/3600} = 30 \text{ calls/hour} \)

The offered traffic is; \( a = \frac{\lambda}{\mu} = \frac{300}{30} = 10 \text{ erlangs} \)

Assuming that one of the time slots is used as a control channel the number of time slots (i.e. trunks) available for the voice service is given as

\[ C = 2 \cdot 8 - 1 = 15 \]

The probability of blocking at this cell site is given by

\[ B[10,15] = \frac{10^{15}/15!}{\sum_{k=0}^{15} 10^k/k!} = 0.0365 \]

Using Table 3.1, we find the average served traffic as

\[ a_s = \frac{\lambda}{\mu} = \frac{300(1 - B[\mu,C])}{30} = 9.635 \text{ erlangs} \]

The average utilization of time slots can be calculated as

\[ \rho = \frac{a_s}{C} = \frac{9.635}{15} = 0.64 \]

Therefore for about \( 1 - \rho = 36\% \) of time, each time slot is available for the GPRS data traffic. Alternatively, if we assume negligible setup overhead for the data traffic, the average number of time slots available to service GPRS can be calculated as

\[ C_{GPRS} = (1 - \rho)C = 0.36 \cdot 15 = 5.4 \]

As a final remark we note that the Erlang B formula is derived under the assumption of exponentially distributed service times. However, it has been proven [1,4] that this formula remains valid for an arbitrary distribution of the service times as well.
3.2.2 Erlang C Formula

The queuing system used for the derivation of the Erlang C formula is of type $M/M/C$. In other words, we assume the following:

1. Service request arrivals can be modeled as a Poisson process
2. Call holding times follow exponential distribution
3. The system has $C$ identical servers
4. The queue is of unlimited capacity

Unlike $M/M/C/C$ system which was used in the derivation of the Erlang B formula, $M/M/C$ does not reject any call request. Provided that the average birth rate is smaller than the average death rate, the system is stable and ultimately all requests will be served. For that reason, in $M/M/C$ systems, GOS provided by the Erlang B formula is inadequate and different performance indicators need to be examined. The type of the most relevant metric depends on the application. Some the most commonly used ones include

- Probability of service request delay
- Average delay for all requests
- Average delay for the requests that are placed in the queue
- 90% delay percentile
- The average number of requests in the queue, and so on.

![Figure 3.4. State transition model for a queuing system of type $M/M/C$](image)

A system satisfying the assumptions of the Erlang C formula can be modeled as a birth and death process with the state transition diagram shown in Fig. 3.4. Using the results of the queuing theory (c.f. Section 2.3), we can write:

$$p_n = \frac{C_n}{S}$$

(3.6)

where

$$C_n = \begin{cases} \frac{(\frac{\lambda}{\mu})^n}{n!} = \frac{a^n}{n!}, & n \leq C \\ \frac{(\frac{\lambda}{\mu})^n}{C^! C^{n-C}} = \frac{a^n}{C^! C^{n-C}}, & n > C \end{cases}$$

(3.7)

and

$$S = \sum_{n=0}^{\infty} p_n$$
RF Engineering Continuing Education
Introduction to Traffic Planning

\[ S = 1 + a + \frac{a^2}{2} + \cdots + \frac{a^c}{C!} + \frac{a^{C+1}}{C!} + \cdots \] (3.8)

If we define a new variable \( \rho = a/c \), (3.8) can be rewritten as

\[ S = 1 + a + \frac{a^2}{2} + \cdots + \frac{a^{C-1}}{(C-1)!} + \frac{a^c}{C!} \left(1 + \rho + \rho^2 + \cdots \right) = \sum_{k=0}^{C-1} k^a + \frac{a^c}{C! (1 - \rho)} \] (3.9)

Therefore,

\[ P_0 = \left( \sum_{k=0}^{C-1} \frac{a^k}{k!} + \frac{a^c}{C! (1 - \rho)} \right)^{-1} \] (3.10)

and

\[ P_n = \begin{cases} a^n \rho_0 & , n \leq C \\ \frac{a^c}{C!} \rho^{n-c} \rho_0 & , n > C \end{cases} \] (3.11)

Using (3.11) we can derive some important performance metrics for \( M/M/C \) systems. The first one is the probability that an arriving service request is placed in queue, that is, the probability of a request being delayed. The delay probability can be calculated as

\[ \Pr(>0) = \Pr\{N \geq C\} = \sum_{n=C}^{+\infty} P_n = 1 - \sum_{n=0}^{C-1} P_n \] (3.12)

\[ \Pr(>0) = 1 - \sum_{n=0}^{C-1} \frac{a^n}{n!} \rho_0 = 1 - \frac{\sum_{n=0}^{C-1} \frac{a^n}{n!}}{\sum_{n=0}^{C-1} \frac{a^n}{n!} + \frac{a^c}{C! (1 - \rho)}} \] (3.13)

or

\[ \Pr(>0) = \frac{\frac{a^c}{C! (1 - \rho)}}{\sum_{n=0}^{C-1} \frac{a^n}{n!} + \frac{a^c}{C! (1 - \rho)}} \] (3.14)

Formula (3.14) is commonly referred to as the Erlang C delay formula.

Now, we compute the average number of call requests that are in the system queue.

\[ L_q = \sum_{n=C}^{+\infty} (n - c) P_n = \sum_{k=0}^{+\infty} k P_{c+k} \] (3.15)

\[ L_q = \sum_{k=0}^{+\infty} k \frac{a^c}{C!} \rho^k \rho_0 = \rho_0 \frac{a^c}{C!} \sum_{k=0}^{+\infty} k \rho^k = \rho_0 \frac{a^c}{C!} \frac{\rho}{(1 - \rho)^2} \] (3.16)
or

\[ L_q = \Pr(>0) \frac{\rho}{1-\rho} = \Pr(>0) \frac{a/C}{1-a/C} = \Pr(>0) \frac{a}{C-a} \]  

(3.17)

The average time that a call request spends in the queue is given by Little's formula

\[ D_1 = W_q = \frac{L_q}{\lambda} = \Pr(>0) \frac{a}{C-a} \frac{1}{\lambda} = \Pr(>0) \frac{H}{C-a} \]  

(3.18)

where \( H = 1/\mu \) is the average call holding time.

The average delay experiences by the calls that are placed in the queue is given by

\[ D_2 = \frac{D_1}{\Pr(>0)} = \frac{H}{C-a} \]  

(3.19)

Finally, the probability of a delay exceeding some given time \( t \) can be calculated as

\[ \Pr(>t) = \Pr(>0) \exp\left(-\frac{C-a}{H} t\right) \]  

(3.20)

Proof for (3.20) can be found in [1].

Table 3.2 summarizes some of the relations that are valid for \( M/M/C \) queuing systems.

\textbf{Example 3.6.} Consider a dispatch system in which 100 users transmitting back to a central system using the \textit{push to talk} approach. The users can begin transmitting at an arbitrary instant in time and the lengths of their messages are distributed exponentially with a mean value of 6 sec. The system is operating using 5 channels. Assuming that each user generates 10 messages per hour, calculate the following:

- Probability of a message being delayed
- Average delay for all messages
- Average delay for delayed messages
- Probability that a message delay exceeds 4 seconds

The average offered traffic

\[ a = 100 \cdot \frac{10 \cdot 6}{3600} = 1.667 \text{ erlangs} \]

Average channel utilization is given as

\[ \rho = \frac{a}{C} = \frac{1.667}{5} = 0.333 \]

Probability of a delayed message is given by

\[
\begin{align*}
\Pr(>0) &= \sum_{n=0}^{C-1} \frac{a^n}{n!} + \frac{a^C}{C!(1-\rho)} = \frac{1.667^5}{5!(1-0.333)} + \frac{1.667^5}{5!(1-0.333)} = \frac{0.1608}{5.1503 + 0.1608} = 0.03
\end{align*}
\]
The average delay for all messages is given by

\[ D_1 = \Pr(>0) \frac{H}{C-a} = 0.03 \frac{6}{5-1.667} = 0.054 \text{ seconds} \]

The average delay for delayed messages

\[ D_2 = \frac{H}{C-a} = \frac{6}{5-1.667} = 1.8 \text{ seconds} \]

Probability of a delay exceeding 4 seconds can be calculated as

\[ \Pr(>4) = \Pr(>0) \exp \left( -\frac{C-a}{H} t \right) = 0.03 \exp \left( -\frac{5-1.667}{6} \cdot 4 \right) = 0.0033 \]

**Table 3.2.** Summary of the performance parameters for \( M/M/C \) queuing system

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Symbol</th>
<th>Value (Formula)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Offered traffic</td>
<td>( a )</td>
<td>( a = \lambda / \mu )</td>
</tr>
<tr>
<td>Channel utilization</td>
<td>( \rho )</td>
<td>( \rho = \frac{a}{C} )</td>
</tr>
<tr>
<td>Probability of having no calls at the cell site</td>
<td>( p_0 )</td>
<td>( \left{ \sum_{k=0}^{C-1} \frac{a^k}{k!} + \frac{a^C}{C!(1-\rho)} \right}^{-1} )</td>
</tr>
<tr>
<td>Probability of having exactly ( n ) calls at the cell site</td>
<td>( p_n )</td>
<td>( \left{ \begin{array}{ll} \frac{a^n}{n!} p_0 &amp; , n \leq C \ \frac{a^C}{C!} \rho^{n-C} p_0 &amp; , n &gt; C \end{array} \right} )</td>
</tr>
<tr>
<td>Probability of service request delay</td>
<td>( \Pr(&gt;0) ) or ( E_C[C,a] )</td>
<td>( \frac{a^C}{C!(1-\rho)} ) ( \sum_{n=0}^{C-1} \frac{a^n}{n!} + \frac{a^C}{C!(1-\rho)} )</td>
</tr>
<tr>
<td>Average number of requests in the queue</td>
<td>( L_q )</td>
<td>( \Pr(&gt;0) \frac{a}{C-a} )</td>
</tr>
<tr>
<td>Average time that all call requests spend in the queue</td>
<td>( D_1 ) or ( T_q )</td>
<td>( \Pr(&gt;0) \frac{H}{C-a} )</td>
</tr>
<tr>
<td>Average time that delay call request spend in the queue</td>
<td>( D_2 )</td>
<td>( \frac{H}{C-a} )</td>
</tr>
</tbody>
</table>
To facilitate the use of Erlang C in everyday engineering practice, (3.14) is frequently provided either in the form of curves or in the form of a table. Erlang C curves are given in Fig. 3.4 while in Appendix B we provide the Erlang C table. To illustrate the use of the Erlang C table, consider the following examples.

**Figure 3.4.** Representation of the Erlang C delay formula through a family of curves

**Example 3.7.** Determine the delay loss probabilities (Pr(> 0) and Pr(> t)) and delays (D₁ and D₂) for a cell with:

a) 24 voice channels if the busy hour traffic is 20 erlangs, the acceptable delay time is 4 seconds, and the average call holding time is 4 seconds
b) 20 voice channels if the busy hour traffic is 16 erlangs, the acceptable delay time is 3 seconds, and the average call holding time is 5 seconds.

**Part a):** In Appendix B, find the table for N = 24 and then find the row for \( a = 20 \) erlangs. Read across to find \( P(> 0) = 0.29807 \) or 29.8% and \( P(> t) = 0.00546 \) or 0.55% for \( t/H = 1.0 \). Calculate \( D₁ \) and \( D₂ \) using (3.18) and (3.19) as follows:

\[
D₁ = Pr(> 0) \frac{H}{C - a} = 0.298 \frac{4}{24 - 20} = 0.298 \text{ seconds}
\]
\[
D₂ = \frac{H}{C - a} = \frac{4}{24 - 20} = 1 \text{ second}
\]
Part b): In Attachment B, find the table for $N = 20$ and then find the row for $a = 16$ erlangs. Read across to find $Pr(> 0) = 0.256$ or 25.6% and $Pr(> t) = 0.0232$ or 2.32% for $t/H = 0.6$. Calculate $D_1$ and $D_2$ using (3.18) and (3.19) as follows:

$$D_1 = Pr(> 0) \frac{H}{C-a} = 0.256 \frac{5}{20-16} = 0.32 \text{ seconds}$$

$$D_2 = \frac{H}{C-a} = \frac{5}{20-16} = 1.25 \text{ seconds}$$

Example 3.8. Determine the offered traffic that

a) 24 voice channels can serve and still provide a delay probability of 2% or better assuming that the average call holding time is 4 seconds and that the acceptable delay is 4 seconds

b) 20 voice channels can serve and still provide a delay probability of 1% or better assuming that the average call holding time is 5 seconds and the acceptable delay is 3 seconds.

Part a): In Attachment B, find the table for $C = 24$ and read down the $P(>t)$ column for $t/H = 1.0$ to find a probability of delay greater than 4 seconds of 0.01851. Read across this row to find $a = 20.9$ erlangs.

Part b): In Attachment B, find the table for $C = 20$ and read down the $P(>t)$ column for $t/H = 0.6$ to find a probability of delay greater than 3 seconds of 0.00994. Read across this row to find $a = 15.2$ erlangs.

Example 3.9. Determine the number of voice channels

a) required to handle a busy hour traffic of 11 erlangs at a grade of service of 3% or better, given an acceptable delay time of 8 seconds and an average call holding time of 10 seconds

b) required to handle a busy hour traffic of 15 erlangs at a grade of service of 1% or better, given an average call holding time of 5 seconds and an acceptable delay time of 6 seconds.

Part a): Using a trial-and-error method, find that the table for $C = 14$ in Attachment B shows a probability of delay greater than 8 seconds to be 0.02748 for $t/H = 0.8$. Therefore, fourteen voice channels are required.

Part b): Using a trial-and-error method find that the table for $C = 18$ in Appendix B shows a probability of delay greater than 6 seconds to be 0.00987 for $t/H = 1.2$. Therefore, eighteen voice channels are required.
3.3 Optional: Traffic Planning for Circuit Switched Services in CDMA networks

Speech service is still the dominant application in the CDMA cellular systems of the second generation. Similar to other cellular technologies, speech service is implemented using circuit switching. However, due to unique properties of the CDMA access scheme, management of communication resources becomes equivalent to the management of the base station transmit power on the forward link, and management of the base station noise rise on the reverse link. In this section we provide analytical methodology that can be used in the process of the CDMA system's traffic planning. Due to the inherent asymmetry, two links are separately analyzed.

3.3.1 Reverse link

To analyze the reverse link of a CDMA cell, we start with by specifying the quality requirement for an individual link as

\[
\left( \frac{E_b}{N_T} \right)_i = \frac{W}{\alpha_i R_i} \cdot \frac{P_i}{I_{total} - P_i}
\]  

(3.21)

In (3.21), \( \left( \frac{E_b}{N_T} \right)_i \) represents the ratio of the signal energy and the total noise power for the \( i \)th connection, \( W \) is the hip rate, \( R_i \) is the data rate of the communication service, \( \alpha_i \) is the activity of the service, \( P_i \) is the received power on the \( i \)th connection, and \( I_{total} \) represents total noise which can be seen as an addition to the thermal noise and the co-channel noise produced by other users operating on the same channel. From (3.21), the required receive power for the \( i \)th user is given by

\[
P_i = \frac{1}{1 + \frac{W}{\alpha_i R_i \left( \frac{E_b}{N_T} \right)_i}} I_{total} = k_i I_{total}
\]  

(3.22)

where \( k_i \) specifies the factor of the overall receive power that can be contributed to the \( i \)th user.

The noise rise on the reverse link is a parameter used to characterize the level of the reverse link loading. It is defined as

\[
NR = 10 \log \left( \frac{I_{total}}{kTFW} \right)
\]  

(3.23)

where \( kTFW \) is the power attributed to the thermal noise.
With the aid of Fig. 3.5, we see that

\[ I_{\text{total}} - kTFW = \sum_{i=1}^{N} P_i = \sum_{i=1}^{N} k_i I_{\text{total}} \]  

(3.24)

Therefore, we have

\[ I_{\text{total}} = \frac{kTFW}{1 - \sum_{i=1}^{N} k_i} \]  

(3.25)

and

\[ NR = -10 \log \left( 1 - \sum_{i=1}^{N} k_i \right) = -10 \log (1 - \beta_{\text{ul}}) \]  

(3.26)

where \( \beta_{\text{ul}} = \sum_{i=1}^{N} k_i \) represents the uplink loading factor.

As the number of users increases, the loading factor increases as well and it becomes closer to 1. However, an increase of the loading factor is accompanied with an increase of the reverse link noise rise and as a result, it becomes progressively more difficult for each of the mobiles to maintain the reverse link connection. Ultimately, as \( \beta_{\text{ul}} \) approached unity, the noise rise increases without bound and it becomes infinite. This condition is commonly referred to as the reverse link pole point and it puts theoretical limit on the number of users that can be supported on the reverse link. Fig. 3.6 shows the relationship between the noise rise and the loading factor. Since the mobile received power needs to overcome the noise rise, it becomes clear that as the loading of the cell increases, the cell radius on the reverse link becomes smaller. This is commonly referred to as cell breathing. During the design phase the level of acceptable noise rise is specified in advance. Currently, most of the systems are designed to allow the noise rise from 2 to 4 dB. As can be seen from Fig. 3.6, this corresponds to the loading factors in the range from 0.4 to 0.6.
So far, we have considered a cell that operates in isolation. In reality, every cell is placed in a cluster of other CDMA cells. As a result, on the reverse link each cell experienced additional interference coming from users in adjacent CDMA cells. This increase is commonly modeled through a correction factor applied to the reverse link loading as

$$\beta_{UL} = \left(1 + I_{adj}\right) \sum_{i=1}^{N} \frac{1}{W + \frac{1}{\alpha R(E_b/N_T)}}$$ \hspace{1cm} (3.27)

Equation (3.27) is valid in a case of an arbitrary mixture of communication services. If the CDMA cell is supporting circuit switched voice service only, (3.27) simplifies as

$$\beta_{UL} = \left(1 + I_{adj}\right) \frac{N_C}{W + \frac{1}{\alpha R(E_b/N_T)}}$$ \hspace{1cm} (3.28)

or

$$N_C = \beta_{UL} \left[1 + \frac{W}{\alpha R(E_b/N_T)}\right] \frac{1}{1 + I_{adj}}$$ \hspace{1cm} (3.29)

where $N_C$ can be interpreted as the number of trunks that are available for the reverse link service. The probability of the outage is now given as $Pr\{N(t) > N_C\}$ and the CDMA cell can be modeled as a queuing system of $M/M/N_C/N_C$ type. As we have seen in Section 3.3.1, this type of system satisfies the assumptions of the Erlang B trunking model and we can perform the traffic analysis using formulas listed in Table 3.1.
As an illustration consider the following example.

**Example 3.10.** Consider a CDMA site in an all voice network having the following numerical parameters: \( W = 1.2288 \times 10^6 \) chips/second, \( \alpha = 0.5 \), \( R = 9.6 \) kb/sec, \( E_b/N_T = 7 \) dB = 5.16 and \( I_{adj} = 0.66 \). If the site is designed to operate at a loading level of \( \beta = 0.6 \), calculate the traffic volume that can be served by the site at the GOS of 2%.

Using (3.29) we have

\[
N = \beta UL \left( 1 + \frac{W}{\alpha R \cdot E_b/N_T} \right) \frac{1}{1 + I_{adj}} = 0.6 \cdot \left( 1 + \frac{1.2288 \times 10^6}{0.5 \cdot 9.6 \times 10^3 \cdot 5.16} \right) \frac{1}{1 + 0.66} = 18.29 \rightarrow 18 \tag{3.30}
\]

Using Erlang B blocking formula, we have

\[
0.02 = \frac{a^{18}/18!}{\sum_{k=0}^{15} \frac{a^k}{k!}} \rightarrow a = 11.49 \text{ erlangs} \tag{3.31}
\]

The value in (3.30) was obtained using the nominal value for the out of cell interference \( I_{adj} = 0.66 \). In a real system, this value is a function of the cell layout, traffic distribution, antenna patterns and other parameters which may be out of control of the traffic planner. Figure 3.7 shows the variation of the site's erlang capacity as a function of the out of cell interference. The curves are generated for two different vocoder rates.

![Figure 3.7](image-url)  
*Figure 3.7.* Traffic volume that can be carried on the reverse link at 2% GOS versus \( I_{adj} \)
3.3.2 Forward link

On the CDMA forward link, communication is conducted in the mode “one to many”. The quality of an individual forward link connection can be expressed as:

\[
\left( \frac{E_{b}}{N_{T}} \right)_{i} = \frac{W}{\alpha_{i}R_{bi} I_{total} - P_{i}/L_{i}} = \frac{P_{i}/L_{i}}{\alpha_{i}R_{bi} P_{BTS-TX}/L_{i} (h_{i} + I_{adj}) + kTFW}
\]

(3.32)

where \( P_{BTS-TX} \) is the total power transmitted by the serving base station, and \( L_{i} \) is the path loss between the serving base station and the \( i \)th mobile. In (3.32) we can identify three factors that contribute to the noise in the denominator. The first factor is \( kTFW \) and it represents omnipresent thermal noise. The second factor is \( P_{BTS-TX}/L_{i} \cdot h_{i} \) which represents a portion of the noise generated by the serving cell. In CDMA systems, each of the downlink signals is encoded with an orthogonal code. If there were no multipath propagation, the signals would remain orthogonal at the point of reception as well.

![Figure 3.8. Illustration of the out of cell interference on the downlink](image)

However, due to the multipath, some of the orthogonality is lost and that results in a level of self-jamming. The parameter \( h_{i} \) quantifies the degree of orthogonality retained on the downlink. The value \( h_{i} = 0 \) corresponds to a perfect orthogonality while \( h_{i} = 1 \) assumes no orthogonality.
at all. Finally, the third contribution to the downlink noise comes from the surrounding cells. Following the approach adopted for the uplink, the out of cell interference is expressed as a fraction of the in-cell interference, i.e. $I_{oc} = I_{adj} P_{TX_{-BS}} / L_i$. Unlike the uplink, where all mobiles at a given cell experience the same amount of the out of cell interference, on the downlink the interference depends on the position of the mobile. To get an idea of some typical values for $I_{adj}$, consider the situation shown in Fig. 3.8. All base stations are assumed to be equally configured and loaded to the same level. Two extreme cases can be identified. For the mobile located at point A, the signal coming from the serving base station is very strong and the amount of out of cell interference is negligible. On the other hand, for the mobile located at point B, the amount of out of cell interference is at a maximum and it can be estimated as:

$$I_{adj} = \frac{2R^{-n} + 3(2R)^{-n} + 6\left(R\sqrt{7}\right)^{-n}}{R^{-n}} = 2 + 3 \cdot 2^{-n} + 6 \cdot \left(\sqrt{7}\right)^{-n}$$

(3.33)

Assuming the path loss exponent of $n = 4$, we obtain

$$I_{adj} = 2 + 3 \cdot 2^{-4} + 6 \left(\sqrt{7}\right)^{-4} = 2.31$$

(3.34)

The average value of the $I_{adj}$ across the cell’s coverage area can be estimated as:

$$\bar{I}_{adj} = \frac{1}{\pi R^2} \int_{0}^{2\pi} \int_{0}^{R} I_{adj} r \, dr \, d\phi = \frac{2}{3} I_{adj} \approx 1.54$$

(3.35)

The majority of the interference on the downlink comes from the outside of the serving cell.

Now, by solving (3.32) for the average transmit power on the downlink channel we obtain:

$$\bar{P}_i = \frac{(E_b / N_T) \alpha_i R_i}{W} \left\{ P_{TX_{-BS}} \left(\bar{h} + \bar{I}_{adj}\right) + kTWF\bar{L} \right\}$$

(3.36)

where $\bar{L}$ is the average path loss. The average power transmitted from the base station can be estimated as

$$P_{TX_{-BS}} = \sum_{i=1}^{N_i} \bar{P}_i + P_{OH} = \sum_{i=1}^{N_i} \frac{(E_b / N_T) \alpha_i R_i}{W} \left\{ P_{TX_{-BS}} \left(\bar{h} + \bar{I}_{adj}\right) + N_0 \bar{L} \right\} + P_{OH}$$

(3.37)

where $P_{OH}$ is the power allocated to the overhead channels and $N_0 = kTWF$. Solving for $P_{TX_{-BS}}$, we obtain

$$P_{TX_{-BS}} = \frac{1}{1 - \beta_{DL}} \left\{ N_0 \bar{L} \sum_{i=1}^{N_i} \frac{(E_b / N_T) \alpha_i R_i}{W} + P_{OH} \right\}$$

(3.38)
where

$$\beta_{DL} = (\bar{h} + \bar{I}_{adj}) \sum_{i=1}^{N_C} \left( \frac{E_b}{N_T} \right) \alpha_i R_i \frac{W}{W}$$  \hspace{1cm} (3.39)$$

Equations (3.38) and (3.39) are valid in a most general situation of heterogeneous communication service. If the analysis is restricted to a voice only situation, (3.38) and (3.39) simplify as follows:

$$\beta_{DL} = (\bar{h} + \bar{I}_{adj}) N_C \left( \frac{E_b}{N_T} \right) \alpha R \frac{W}{W}$$  \hspace{1cm} (3.40)$$

and

$$P_{TX_{BS}} = \frac{1}{1 - N_C (\bar{h} + \bar{I}_{adj}) \left( \frac{E_b}{N_T} \right) \alpha R \frac{W}{W} \left( \frac{N_0 L N_C \left( \frac{E_b}{N_T} \right) \alpha R}{W} + P_{OH} \right) + \alpha R}$$  \hspace{1cm} (3.41)$$

Solving (3.41) for $N_C$, we obtain

$$N_C = \frac{P_{TX_{BS}} - P_{OH}}{N_0 L + P_{TX_{BS}} (\bar{h} + \bar{I}_{adj}) \left( \frac{E_b}{N_T} \right) \alpha R \frac{W}{W}}$$  \hspace{1cm} (3.42)$$

In a high capacity area, most of the interference on the downlink is induced by the CDMA system itself. Therefore the noise term in (3.42) can be neglected and the equation simplifies to

$$N_C = \frac{1 - P_{OH} / P_{TX_{BS}}}{(\bar{h} + \bar{I}_{adj}) \left( \frac{E_b}{N_T} \right) \alpha R \frac{W}{W}}$$  \hspace{1cm} (3.43)$$

Equation (3.42), does not take into account the loss of capacity that results from soft handoff. This can be taken into account through a multiplicative correction factor and (3.43) becomes

$$N_C = \frac{1 - P_{OH} / P_{TX_{BS}}}{(\bar{h} + \bar{I}_{adj}) \left( \frac{E_b}{N_T} \right) \alpha R \frac{W}{W} \cdot \frac{1}{S_{HO}}}$$  \hspace{1cm} (3.44)$$

Quantity $S_{HO}$ models the loss of capacity due to the soft handoff.

The quantity $N_C$ can be interpreted as the number of trunks that are available for the forward link voice service. The probability of the outage is now given as $Pr\{N(t) > N_C\}$ and the CDMA cell can be modeled as a queuing system of $M/M/N_C/N_C$ type. As we have seen in Section 3.3.1, this type of system satisfies the assumptions of the Erlang B trunking model and we can perform the traffic analysis using formulas listed in Table 3.1.
Example 3.11. Consider a cell site with the following forward link numerical parameters:  \( W = 1.2288 \times 10^6 \) chips/second, \( \alpha = 0.5 \), \( R = 9.6 \) kb/sec, \( E_b/N_T = 6 \) dB = 3.98, \( h = 0.1 \) and \( I_{adj} = 1.54 \). Assume that the total PA power is \( P_{0S_{_TX}} = 17 \) W, out of which 4W is allocated to the overhead channels. If the base station should not exceed 50% of its maximum power, calculate traffic in erlangs that can be supported at the GOS of 2%. Assume a soft handoff reduction factor of \( S_{HO} = 1.4 \).

Using (3.44) we obtain

\[
N_c = \frac{1 - P_{oh}/P_{TX_{_BS}}}{(h + I_{adj})(E_b/N_T)} \cdot \frac{W}{\alpha R} \cdot \frac{1}{S_{HO}} \cdot \frac{1 - 4}{0.75 \times 17} \cdot \frac{1.2288 \times 10^6}{0.5 \times 9.6 \times 10^3} \cdot \frac{1}{1.4} = 19.23 \rightarrow 19
\]

Using Erlang B blocking formula, we have

\[
0.02 = \sum_{k=0}^{19} \frac{a^k}{k!} \rightarrow a = 12.33 \text{ erlangs}
\]

3.4 Estimation of the Offered Traffic in Cellular Systems

Unlike fixed landline telephone systems where the location of every telephone is known, mobile users are distributed over the cellular system coverage area in relatively random fashion. In order to perform proper design and system resource dimensioning an accurate estimate of the mobile traffic needs to be obtained. To accomplish the task two estimates need to be provided. The first one is the estimate of the geographical distribution of the mobile users and the second one is the estimate of the traffic that an average user generates within the busy hour.

The geographic distribution of the mobile users is commonly estimated using Geographic Information Survey (GIS) data. Such data is available from many data providers and it includes various census and other statistical information collected over the past several years. Some GIS data in common use for mobile traffic estimation purposes is given by

- Average population density
- Average family income
- Land use type
- Road usage

Although many GIS data types provide useful information about the distribution of the mobile users, none of them gives a complete picture. For that reason, the geographic distribution of the users is commonly estimated as a combination of the GIS data. A fixed methodology for how to accomplish this does not exist. Most of the time it is left to the traffic planner to subjectively evaluate individual types of available GIS data and combine them in some logical sense. To illustrate the process, consider the following example.
Example 3.12. Distribution of mobile users within a given market is to be estimated. The following GIS and marketing data is available:

- Average population density
- Land use type with four morphological classifications – dense urban (DU), urban (U), suburban (SU) and open area (OA)
- Road use data with four road type classification and associated relative weights. The roads are classified as US interstates (weight of 10), US highways (weight of 5), major roads (weight of 4) and secondary roads (weight of 1).
- Total population within the market boundary and expected penetration rate

Using the above information, the distribution of mobile users can be obtained using the following procedure:

1. Convert all GIS data to relative demand grids. To convert the land use data to relative demand grid, relative weights need to be assigned to individual morphology types. Some typical assignments are: weight(DU) = 10, weight(U) = 50, weight(U) = 3, weight(SU) = 1.
2. Eliminate all GIS data outside of the market boundary
3. Determine the total number of users within the market boundary - \( N_{\text{users}} \)
4. Use the following formula

\[
DG = \left[ \frac{\rho_{\text{POP}}}{\max(POP_{DG})} \cdot POP_{DG} + \frac{\rho_{\text{LUSE}}}{\max(LUSE_{DG})} \cdot LUSE_{DG} + \frac{\rho_{\text{ROAD}}}{\max(ROAD_{DG})} \cdot ROAD_{DG} \right] \cdot N_{\text{users}} \cdot P_R \cdot (1 + R_F) \tag{3.45}
\]

where

- \( \rho_{\text{POP}} \) - fraction of users that are in place of their residence
- \( POP_{DG} \) - relative demand grid obtained from population density data
- \( \rho_{\text{LUSE}} \) - fraction of users distributed throughout the market area outside of roads
- \( LUSE_{DG} \) - relative demand grid obtained from land use data and associated relative weights
- \( \rho_{\text{ROAD}} \) - fraction of users that are located on the roads
- \( N_{\text{users}} \) - total population in the market
- \( P_R \) - market penetration rate
- \( R_F \) - roaming factor that models the increase in number of users due to roaming traffic

To implement the methodology described in Example 3.12, we need to be able to manipulate demand grids. The basic manipulations include conversion of demand grids from absolute to relative and relative to absolute as well as scaling and addition of demand grids. Most of the GIS data presentation software as well as some network planning tools [5] allow for these basic operations.
After the geographic distribution of the users throughout the market area is determined, we need to estimate the busy hour traffic generated by a single user. Estimation of the traffic per subscriber is not an easy task since it depends on many factors that may not be within control of the traffic planning engineer. Some of these factors are service price, service availability, customer cultural background, and many others. The traffic generated per subscriber is a function of two variables: the average call holding time and the number of phone calls within the busy hour. Table 3.3 shows a semiannual statistical report produced by the Cellular Telecommunication Industry Association (CTIA) [6]. The reports provide historical information about the average call holding time (CHT). Using the results shown in Table 3.3, we can have

\[
H_{\text{LOCAL}} = \frac{2.27 + 2.15 + \cdots + 2.56}{12} = 2.34 \, \text{[min]}
\]

and

\[
H_{\text{ROAM}} = \frac{2.74 + 2.79 + \cdots + 3.23}{12} = 2.97 \, \text{[min]}
\]

Therefore one roaming customer "counts" as approximately \(2.97/2.34=1.27\) local customers which should be taken into account in (3.45).

The average number of phone calls per user within the busy hour can be roughly estimated as follows. Table 3.3 reports an average monthly bill of 45 dollars. Analysis of the service plans for several large cellular and PCS providers reveals that 45 dollars corresponds to approximately 450 minutes of peak time usage\(^1\). The peak time is usually defined as the period from 8:00 AM to 8:00 PM during workdays (Monday to Friday). Therefore, an average user that spends 45 dollars for cellular service talks for about 450 minutes per month which corresponds to 450/24 = 18.75 minutes per day (peak time). If the average call holding time is 2.34 minutes, an average user places 18.75/2.34 = 8.01 phone calls during a workday. If we assume that 15% of the phone calls are placed within the busy hour, we obtain an estimate of 1.2 phone calls per user within the busy hour.

Knowing the average call holding time and the average number of phone calls within the busy hour, one can easily estimate the traffic generated by a single user. Using the above derived numerical values we have

\[
a_{\text{USER}} = \frac{n_{\text{calls}} \times \text{CHT}_{\text{av}}}{60} = \frac{1.2 \times 2.34}{60} = 46.8 \approx 50 \, \text{mE}
\]

(3.46)

In practice, the traffic generated per user in the busy hour is commonly assumed in the range from 20 to 100mE.

\(^1\) These numbers are determined by evaluating the service plans offered in the summer of 2001 within USA Florida markets.
### Table 3.3. Part of the CTIA semiannual report

<table>
<thead>
<tr>
<th>Date</th>
<th># Cellular subscribers</th>
<th># Cellular cell sites</th>
<th>Average monthly bill</th>
<th>Average CHT of local call</th>
<th>Average CHT for roaming call</th>
</tr>
</thead>
<tbody>
<tr>
<td>June 95</td>
<td>28,154,415</td>
<td>19,844</td>
<td>52.45</td>
<td>2.27</td>
<td>2.74</td>
</tr>
<tr>
<td>Dec 95</td>
<td>33,785,661</td>
<td>22,663</td>
<td>51.00</td>
<td>2.15</td>
<td>2.79</td>
</tr>
<tr>
<td>June 96</td>
<td>38,195,466</td>
<td>24,802</td>
<td>48.84</td>
<td>2.24</td>
<td>2.8</td>
</tr>
<tr>
<td>Dec 96</td>
<td>44,042,992</td>
<td>30,045</td>
<td>47.70</td>
<td>2.32</td>
<td>3.14</td>
</tr>
<tr>
<td>June 97</td>
<td>48,705,553</td>
<td>38,650</td>
<td>43.86</td>
<td>2.25</td>
<td>2.95</td>
</tr>
<tr>
<td>Dec 97</td>
<td>55,312,293</td>
<td>51,600</td>
<td>42.78</td>
<td>2.31</td>
<td>2.94</td>
</tr>
<tr>
<td>June 98</td>
<td>60,831,431</td>
<td>51,674</td>
<td>39.88</td>
<td>2.34</td>
<td>2.65</td>
</tr>
<tr>
<td>Dec 98</td>
<td>69,209,321</td>
<td>65,887</td>
<td>39.43</td>
<td>2.39</td>
<td>3.11</td>
</tr>
<tr>
<td>June 99</td>
<td>76,284,753</td>
<td>74,157</td>
<td>40.24</td>
<td>2.4</td>
<td>2.96</td>
</tr>
<tr>
<td>Dec 99</td>
<td>86,047,003</td>
<td>81,698</td>
<td>41.24</td>
<td>2.38</td>
<td>3.11</td>
</tr>
<tr>
<td>June 00</td>
<td>97,035,925</td>
<td>95,733</td>
<td>45.15</td>
<td>2.48</td>
<td>3.19</td>
</tr>
<tr>
<td>Dec 00</td>
<td>109,478,031</td>
<td>104,288</td>
<td>45.27</td>
<td>2.56</td>
<td>3.23</td>
</tr>
</tbody>
</table>

### 3.5 Case Study Example

For this example we will play the role of a traffic engineer for a PCS 1900 GSM service provider. The main purpose of this example is to illustrate the methodology provided in previous sections. We will assume that the preliminary coverage design has been completed and the intention is to implement a 7/21 frequency reuse plan. The GSM network provides a 200 kHz channel bandwidth and a time division multiple access scheme with eight time slots per carrier. However, not all of these time slots can be used as traffic channels. The restrictions shown in Table 3.4 apply.

#### Table 3.4. Carrier to voice channel relationships for a PCS 1900 GSM system

<table>
<thead>
<tr>
<th>Number of Radios</th>
<th>Number of Voice Channels</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>14</td>
</tr>
<tr>
<td>3</td>
<td>21</td>
</tr>
<tr>
<td>4</td>
<td>29</td>
</tr>
<tr>
<td>5</td>
<td>37</td>
</tr>
<tr>
<td>6</td>
<td>44</td>
</tr>
<tr>
<td>7</td>
<td>52</td>
</tr>
<tr>
<td>8</td>
<td>60</td>
</tr>
</tbody>
</table>

The system is planned to start out with tri-sectored sites throughout the system. The following information has been supplied:

- A GOS of 2% is required.
- The system is to provide for a 5% market penetration rate by the end of year one and is to focus only on the metropolitan areas (dense urban, urban, and suburban).
The coverage design is to provide for 90% coverage reliability outdoors throughout the dense urban, urban, and suburban areas. Sixty percent of the subscribers supported by a cell will be active during the busy hour. The number of call attempts that each active subscriber makes during the busy hour is one ($A_{sub} = 1$). The average call holding time during the busy hour is two minutes and has an exponential distribution ($H = 2$ min). Blocked calls are cleared from the system. No more than four radios can be used at a cell due to equipment limitations. Assume a uniform distribution of subscribers over the service area.

A demographic analysis has been performed for the market based on census statistics and the results are as follows in Table 3.5.

<table>
<thead>
<tr>
<th>Morphological Classification</th>
<th>Total Area (sq. miles)</th>
<th>Percent of Total Area</th>
<th>Total Population</th>
<th>Percent of Total Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dense Urban</td>
<td>1.06</td>
<td>0.00%</td>
<td>34,683</td>
<td>0.68%</td>
</tr>
<tr>
<td>Urban</td>
<td>29.26</td>
<td>0.06%</td>
<td>344,103</td>
<td>6.70%</td>
</tr>
<tr>
<td>Suburban</td>
<td>1,576.94</td>
<td>3.42%</td>
<td>2,937,042</td>
<td>57.22%</td>
</tr>
<tr>
<td>Rural</td>
<td>44,458.26</td>
<td>96.51%</td>
<td>1,817,070</td>
<td>35.40%</td>
</tr>
<tr>
<td>TOTAL</td>
<td>51,247.82</td>
<td>100%</td>
<td>5,132,898</td>
<td>100%</td>
</tr>
</tbody>
</table>

Nominal cell ranges have been calculated for each of the area types based on a propagation model, specified coverage reliability, and link budgets. The results are as follows in Table 3.6.

<table>
<thead>
<tr>
<th>Morphological Classification</th>
<th>Nominal Cell Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dense Urban</td>
<td>0.56 miles</td>
</tr>
<tr>
<td>Urban</td>
<td>0.96 miles</td>
</tr>
<tr>
<td>Suburban</td>
<td>2.89 miles</td>
</tr>
</tbody>
</table>

We are to determine the number of sites required to meet the coverage objectives based on the provided information and then determine the number of radios required on a per cell basis for each of the area types. We are then asked to make recommendations for additional sites to handle additional capacity.

Solution:

Our first step is to calculate the average busy hour traffic per subscriber ($T_{sub}$). We can do so by using the information provided as

$$T_{sub} = \frac{H}{OneHour} (A_{sub}) = \frac{2 \text{ min}}{60 \text{ min}} (1) = 0.0333 \text{ erlangs}$$
Since we are given the market penetration rate of 5% and we know the population of the dense urban, urban, and suburban areas, we can determine the number of subscribers on a per area basis. The results are as follows:

### Table 3.7. Determination of the number of subscribers

<table>
<thead>
<tr>
<th>Morphological Classification</th>
<th>Population</th>
<th>Market Penetration Rate</th>
<th>Number of Subscribers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dense Urban</td>
<td>34,683</td>
<td>5%</td>
<td>1,734</td>
</tr>
<tr>
<td>Urban</td>
<td>344,103</td>
<td>5%</td>
<td>17,205</td>
</tr>
<tr>
<td>Suburban</td>
<td>2,937,042</td>
<td>5%</td>
<td>146,852</td>
</tr>
<tr>
<td>TOTAL</td>
<td>3,315,828</td>
<td>N/A</td>
<td>165,791</td>
</tr>
</tbody>
</table>

Now that we know the number of subscribers, the percentage of subscribers that are active during the busy hour, and the average busy hour traffic of these subscribers we can determine what the offered traffic will be on a per area basis. The results are shown in Table 3.8.

### Table 3.8. Determination of the offered traffic

<table>
<thead>
<tr>
<th>Morphological Classification</th>
<th>Number of Subscribers (#subs)</th>
<th>% Active During Busy Hour</th>
<th>Average Busy Hour Traffic per Subscriber</th>
<th>Offered traffic [erlangs]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dense Urban</td>
<td>1734</td>
<td>60%</td>
<td>0.0333 erlangs/sub</td>
<td>34.65</td>
</tr>
<tr>
<td>Urban</td>
<td>17,205</td>
<td>60%</td>
<td>0.0333 erlangs/sub</td>
<td>343.76</td>
</tr>
<tr>
<td>Suburban</td>
<td>146,852</td>
<td>60%</td>
<td>0.0333 erlangs/sub</td>
<td>2934.10</td>
</tr>
<tr>
<td>TOTAL</td>
<td>165,791</td>
<td>N/A</td>
<td>N/A</td>
<td>3312.51</td>
</tr>
</tbody>
</table>

The next step in our analysis should be to determine the number of erlangs that each carrier can supply. The information provided to us allows us to determine the appropriate traffic model to use. By evaluating the provided information we determine that the Erlang B model is appropriate. By using the Erlang B table in Attachment A and information contained in Table 3.4, we determine the carried loads shown in Table 3.9.

### Table 3.9. Determination of maximum carried load in a GSM system for a 2% GOS

<table>
<thead>
<tr>
<th>Radios</th>
<th>Voice Channels</th>
<th>Carried Load (erlangs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7</td>
<td>2.94</td>
</tr>
<tr>
<td>2</td>
<td>14</td>
<td>8.20</td>
</tr>
<tr>
<td>3</td>
<td>21</td>
<td>14.0</td>
</tr>
<tr>
<td>4</td>
<td>29</td>
<td>21.0</td>
</tr>
<tr>
<td>5</td>
<td>37</td>
<td>28.3</td>
</tr>
<tr>
<td>6</td>
<td>44</td>
<td>34.7</td>
</tr>
<tr>
<td>7</td>
<td>52</td>
<td>42.1</td>
</tr>
<tr>
<td>8</td>
<td>60</td>
<td>49.6</td>
</tr>
</tbody>
</table>

The demographic analysis (refer to Table 3.5) has provided us with area percentages for each of the morphological areas. A nominal cell radius for each of these area types has also been provided, allowing us to determine the area covered by each of these sites (assuming a circular
coverage area). By dividing the total area of each land class type by the area covered by one cell in that land class type, we can obtain a rough estimate of how many cells are required to cover the area. The results are as follows:

**Table 3.10.** Determination of the number of required sites for coverage

<table>
<thead>
<tr>
<th>Morphological Classification</th>
<th>Nominal Cell Range (miles)</th>
<th>Area Covered by One Site (sq. miles)</th>
<th>Total Area to be Covered (sq. miles)</th>
<th>Number of Sites Required for Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dense Urban</td>
<td>0.56</td>
<td>0.99</td>
<td>1.06</td>
<td>2</td>
</tr>
<tr>
<td>Urban</td>
<td>0.96</td>
<td>2.90</td>
<td>29.26</td>
<td>11</td>
</tr>
<tr>
<td>Suburban</td>
<td>2.89</td>
<td>26.24</td>
<td>1576.94</td>
<td>61</td>
</tr>
<tr>
<td>TOTAL</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>74</td>
</tr>
</tbody>
</table>

Now that we know how many sites we should have in each area and the required offered load for each area, we can determine the number of traffic resources, or radios, required to meet the expected demand (refer to Table 3.11).

**Table 3.11.** Determination of the number of radios required per sector

<table>
<thead>
<tr>
<th>Morphological Classification</th>
<th>Number of Sites</th>
<th>Number of Sectors per Site</th>
<th>Required Offered Load per Area (erlangs)</th>
<th>Required Offered Load per Sector (erlangs)</th>
<th>Number of Radios per Sector</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dense Urban</td>
<td>2</td>
<td>3</td>
<td>34.65</td>
<td>5.78</td>
<td>2</td>
</tr>
<tr>
<td>Urban</td>
<td>11</td>
<td>3</td>
<td>343.76</td>
<td>10.41</td>
<td>3</td>
</tr>
<tr>
<td>Suburban</td>
<td>61</td>
<td>3</td>
<td>2934.10</td>
<td>16.03</td>
<td>4</td>
</tr>
</tbody>
</table>

Our results show that no more than four radios will be required at each site to provide a GOS of 2% to a subscriber base that is made up of 5% of the total market population, and of which only sixty percent are active during the busy hour. Therefore, no additional capacity sites will be required at the end of year one to meet the expected demand.
4 Traffic Planning for Packet Data Services

All second-generation cellular technologies have experienced a phenomenal growth. They have largely succeeded in delivering voice communication based on the paradigm “anytime and anywhere”. With the introduction of flexible billing plans, large coverage areas, roaming agreements and significant industry mergers, voice service has become a commodity with little pricing difference between various cellular providers. It is now evident that in order to remain competitive, cellular carriers must continue to enrich their offerings by providing additional communication services. The form of the services introduced to cellular networks is heavily influenced by the huge growth of the wireline data and the Internet. A noticeable trend is that most of the second-generation cellular providers are making considerable efforts to port some of the popular data applications such as e-mail and limited web browsing to their wireless platforms and offer them as a part of their service portfolio. However, in most cases, the introduction of wireless data comes at a fairly high cost. The second-generation technologies are not designed to handle data traffic. Since they are optimized for voice and are circuit switched, the data rates that they can support are fairly small (on the order of 10 Kb/sec). In addition, circuit switching makes the access scheme fairly expensive and that is a significant limiting factor standing in the way of widespread data service availability.

Third generation systems are being designed to address deficiencies of existing wireless cellular networks by providing access to packet data services. At present, the migration paths have been selected for all second-generation technologies and they are summarized in Fig. 4.1. As it can be seen, there are three major third generation technologies. CDMA2000 is successor to a very popular cdmaOne (IS-95A/B) and it will be deployed mostly in North and South America and some portions of Asia. EDGE is a TDMA based 3G technology that uses a 200 kHz channel and an advanced modulation scheme to provide high data rate throughputs necessary for the 3G communication services. This technology is a successor to the North American TDMA (IS-136) which is deployed mostly in North and South America. Finally, UMTS-FDD (W-CDMA) is a wide band CDMA technology that provides an evaluation path for both GSM and Japanese PDC. In addition, there are several 2.5G technologies that can be used as a transitional solution between GSM and W-CDMA. These technologies are High Speed Circuit Switched Data (HSCSD), General Packet Radio Service (GPRS), and Enhanced Data GSM Environment (EDGE). It is not necessary for GSM operators to deploy any of the 2.5G technologies. Therefore, many of them have announced a direct transition to either GPRS or W-CDMA.

The introduction of third generation technologies is aimed to deliver an entirely new set of communication services to the mobile market. Most of these services are packet data based and according to many market projections they are expected to become a dominant portion of traffic in future cellular networks. As an illustration, consider the chart in Fig. 4.2. The chart is created as a result of a study by the UMTS forum [5], and it predicts the expected revenue stream from all 3G services up to the year 2010. By the year 2004, the revenue generated by data services is expected to surpass that generated by the voice service. Accordingly, in the relatively near future, we may see wireless cellular networks as essentially data networks that dedicate only a portion of their resources to support circuit switched voice.
RF Engineering Continuing Education
Introduction to Traffic Planning

Figure 4.1. 3G evolution path for 2G cellular technologies

Figure 4.2. Global revenue projection for 3G telecommunication services
As evident from Fig. 4.2, the cellular networks of the future are expected to support a variety of communication services. The services differ in the requirements that they place on the communication network. Some of these differences are summarized in Table 4.1. Assuring that each of the services is delivered with sufficient quality will become a major part of the network design, optimization and maintenance. A first step in the network quality guarantee is the proper dimensioning of the network resources. Therefore, as the 3G services are rolled out, the importance of traffic planning and resource dimensioning is expected to increase.

<table>
<thead>
<tr>
<th>Type of service</th>
<th>Conversational</th>
<th>Streaming</th>
<th>Interactive</th>
<th>Background</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delay tolerance</td>
<td>Low</td>
<td>High</td>
<td>Medium</td>
<td>High</td>
</tr>
<tr>
<td>Jitter tolerance</td>
<td>Low</td>
<td>Low</td>
<td>Medium</td>
<td>High</td>
</tr>
<tr>
<td>Data rate requirement</td>
<td>Small to large</td>
<td>High</td>
<td>Low to medium</td>
<td>Low</td>
</tr>
<tr>
<td>Data symmetry</td>
<td>Symmetrical</td>
<td>Asymmetrical</td>
<td>Asymmetrical</td>
<td>Low</td>
</tr>
<tr>
<td>Reliability tolerance</td>
<td>High</td>
<td>High</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Typical applications</td>
<td>Circuit switched telephony, video conferencing, etc.</td>
<td>Audio and video broadcasting, etc.</td>
<td>E-commerce, location based services, web browsing, etc.</td>
<td>File transfer, e-mail, etc.</td>
</tr>
</tbody>
</table>

In this section we address some aspects of traffic planning in wireless cellular packed data networks. This area is relatively new and lacks general practical engineering experience. However, since data communication networks are so widespread on the wireline side much of the methodology used for their design can be used in the wireless surroundings as well.

### 4.1 Implementation of Data Services in Cellular Networks

There are two approaches to the implementation of packet data services. The principle difference is in the air interface utilization. According to the first approach, the packet data services are implemented through protocol changes on the air interface. The protocol changes allow sharing of the network's air interface resources between circuit switched services and packet data. This approach is implemented in GPRS, W-CDMA, EDGE and CDMA2000 EvDV². According to the second approach, the circuit switched and packet switched services are implemented through separate air interface. This approach is adopted by CDMA2000 EvDO³, which is an entire packet data network built around CDMA2000 air interface. In addition, there are several proprietary solutions that advocate a similar principle of operation. Both approaches have some advantages and disadvantages.

The first approach uses the capabilities of 3G technologies to the full extent. It allows dynamic allocation of the network resources between the data and circuit switched voice. On the outset, this seems like an optimal approach since it avoids segmentation of the network resources. However, in many instances this is not the case. The reason can be found in the different behavior of the data and voice services in the RF propagation environment. For example, due to

---

2 CDMA2000 1xEV DV stands for *CDMA2000 Evolution Voice and Data*
3 CDMA2000 1xEV DO stands for *CDMA2000 Evolution Data Only*
different processing gains, voice and data services placed on the same CDMA RF carrier have different coverage contours. This is generally undesirable and it causes some significant problems that need to be resolved during the system RF design phase.

The main advantage of the second approach is in its simplicity. Spectral separation of the voice and data traffic allows for their separate traffic management. By doing so, both sides can be dimensioned so that they have approximately the same RF performance. However, segmentation of the spectrum leads to sub-optimal use of network resources. To see that this is the case, consider the traffic volume traces for an imaginary site shown in Fig. 4.3.

![Traffic Volume Traces](image)

**Figure 4.3.** Traffic volume traces for voice and packet data

The peak hour for the packet data and voice traffic do not occur simultaneously. If the traffic dimensioning of the site is done separately for voice and data the number of allocated resources will be proportional to \( a + b \). On the other hand, if the two traffic types are handled at the same time, the number of required resources is proportionate to \( c \). From Fig. 4.3, it is obvious that \( c \leq a + b \). Hence, there is a loss in capacity due to the sub-optimal traffic balancing (See Example 2.5). Nevertheless, after taking RF factors into consideration, separate management of the voice and data traffic is appealing to some carriers, especially the ones following the IS-95 3G migration path.

While there are significant differences on the air interface side, the migration of core networks essentially follows the same path for all second generation technologies. An outline of the cellular network supporting packet data services is presented in Fig. 4.4. The network consists of essentially two independent domains: circuit switched domain and packet switched domain. The circuit switched domain supports the voice service, while the data is handled by the packet switched domain. Two domains share the same HLR and VLR databases.
The core network shown in Fig.4.4 is an evolution of the core networks currently used to support North American based systems (TDMA, CDMA, AMPS) – IS-41. The packet data domain of the network is designed around existing standards provided by the Internet Engineering Task Force (IETF). Therefore it consists of standard routers and switching equipment that already dominates the landline Internet data services. On the other hand, in GSM networks the packet data domain is implemented using a new set of logical entities called the service nodes: Service GPRS\(^4\) Service Node (SGSN) and Gateway GPRS Service Node (GGSN). Although slightly different in their implementation, the two network architectures are quite similar in the way they operate. In the near future we are likely to see their convergence.

### 4.2 Packet Data Traffic Modeling

In mobile cellular networks of third generation, circuit switching is used to support only voice traffic. All other traffic types are supported through packet switching. Therefore, packet switching is used to support a variety of potentially very different communication services. For example, both e-mail and WWW browsing are extremely popular packet data services. However, they have substantially different characteristics. While e-mail usually involves text messaging and exchange of relatively small documents, WWW browsing implies transfer of text, images and very often multimedia data. As a result, the two applications require substantially

\(^4\) GPRS stands for – General Packet Radio Service
different data rates. Furthermore, e-mail is usually delivered on a best effort basis and since users are well aware of that, there are no serious delay requirements. On the other hand, while Web browsing, users expect relatively short response and load times, and if they are forced to wait for a long time they are inclined to render the service unusable.

In this section we provide some insight into the traffic modeling of packet data networks. Due to a variety of data service types and flexibility imbedded in the packet switching we will see that there is no unique approach that can be applied to all situations. Despite extensive research in the field, an "Erlang B" equivalent for data traffic still does not exist. However, there are some general methods that allow us to perform effective network dimensioning in many situations of great practical interest. Some of these methods will be explained in the following sections.

### 4.2.1 Description of Packet Data Traffic

As discussed in Section 1.2, packet data source does not occupy the network resources in situations that don not require data transfer. This allows the network to share its resources with many users counting on the fact that all of them will not be active at the same time. In most networks, the packet generation process can be seen as a random process characterized with several parameters. Some of those parameters along with their definition are given as follows.

**Average Rate.** The average rate is defined as the average number of data packets that a source sends per unit time. The average rate is usually determined by observing the behavior of the packed data source over some large time interval. As an example, consider the transmission pattern between two computers shown in Fig. 4.5. The two computers are not active all the time and the average rate becomes a measure of their overall activity.

![Diagram of data communication pattern between two computers](image)

**Figure 4.5.** Example of data communication pattern between two computers

**Peak Rate.** While the average rate describes the long term activity of the packet source, the peak rate is used to describe its short term behavior. Unlike voice transmissions, which usually can be seen as a continuous stream at a relatively small data rate, most data applications are intermittent. Data transmission is followed by intervals of inactivity and vice versa. However, during periods of activity, data applications usually make an attempt to capture the largest
possible bandwidth so that the information can be delivered as quickly as possible. This kind of behavior is usually referred to as "bursty" and in general it is considered to be one of the prime characteristics of data traffic. The peak rate represents a convenient measure of the application burstiness.

**Burstiness.** The formal definition of the burstiness can be given as

\[
\text{Burstiness} = \frac{\text{Peak Rate}}{\text{Average Rate}}
\]

An alternative measure that can be used to characterize the burstiness of a data source is the probability of source activity. By definition it is the activity of the source being active at any given time and it can be calculated as

\[
\text{Source Activity} = \frac{1}{\text{Burstiness}} = \frac{\text{Average Rate}}{\text{Peak Rate}}
\]

For example, if a packet data source has a peak rate of 144 Kb/sec and an average data rate of 8 Kb/sec, its burstiness factor is given as 144/8=18 and the probability of its activity is about 6%. One should note that the above definitions of the burstiness and source activity capture only a portion of the source behavior. Two sources can have the same burstiness but completely different distributions of activity times. As we will see in the following sections, time domain effects (interarrival and service times) are of major importance in data traffic modeling.

**Priority.** Cellular networks that support packet data usually have built in mechanisms to provide different priority to different services. These mechanisms need to be taken into account when performing the traffic dimensioning. The most common approach is to give the highest priority to the circuit switched voice and utilize the unused resources to provide service on a best effort bases. However, with the rollout of the 3G services we may expect an entire toolbox of features that allow discrimination between different traffic types on the basis of delay tolerance, error tolerance, data rate requirement, and many other parameters.

### 4.2.2 Self Similarity in Data Traffic

The fundamental parts of traffic modeling are the probability distribution of the interarrival and service times. Although Poisson modeling produces tractable and elegant analytical results, it has been discovered that there are many realistic traffic processes for which this modeling approach is not applicable. In general, the Poisson model is usually adequate for traffic processes that involve immediate human activity. Examples of such processes include circuit switched telephony, SMS services, e-mail with no attachments, voice mail and dispatch radio. Due to exponential distribution of the service times these processes are mostly characterized with a large number of short messages and a relatively small number of long messages. For example, no one types a 3000 page e-mail message. On the other hand, extensive research in the properties of the data traffic has revealed that many data services do not have completely Poisson like behavior. For most data services, the service request arrival process is random and it can
still be modeled accurately as Poisson. However, the distribution of the service times may not be exponential at all. To understand the difference, consider our everyday experience with the Ethernet traffic in the local area network. Although the majority of traffic still consists of short messages (e-mails, short file transfers, etc.), it is not unusual to use network resources to transfer large documents, pictures and other large files. In extreme situations, even gigabytes of data may be transported during the backup processes. In short, data traffic shows much larger variability and the service time may range from several milliseconds to several hours.

The non-exponential distribution of the service time in data networks was noticed by many researchers. An extensive measurement campaign conducted by Willinger et. al. [8], discovered that data traffic exhibits long term dependence uncharacteristic for Poisson processes. They have used the concept of self-similarity to mathematically describe their observations. This concept is somewhat abstract and it extends beyond the scope of this paper. However, we will try to illustrate the essence of their observations through the following example.

Consider two processes shown in Figs. 4.6 and 4.7. The process shown in Fig. 4.6 is a Poisson process model for the traffic in a circuit switched network.

![Figure 4.6. Aggregation of the Poisson processes](image)

Figure 4.6. Aggregation of the Poisson processes
At the average traffic volume of 10E, the traffic is quite bursty. Points where the instantaneous traffic volume exceeds the average by as much as 60% are not rare. However, as the average of traffic is increased we see that the burstiness becomes smaller. Although the traffic variations become larger in the absolute sense, they become smaller relative to the mean. The actual numerical data characterizing the traces given in Fig. 4.6 are shown in Table 4.2. We see that as the volume of the traffic becomes larger, the traffic tends to become smoother. As a numerical indication of the traffic burstiness, the table provides the volume threshold exceeded in 2% of the time. This threshold corresponds to the number of traffic resources that would have to be allocated so that the traffic could be served at the GOS of 2%. As we expect, as the volume of the traffic increases, the number of resources needed to maintain the same GOS becomes smaller relative to the average traffic volume. For example, to serve 10E at GOS of 2%, we need approximately 16 trunks, which is 60% higher than the average traffic value. However, to provide the same GOS to the traffic volume of 100E, the necessary number of trunks is only 112, which exceeds the average traffic volume by only 12%.

**Table 4.2.** Numerical data associated with the process given in Fig. 4.6

<table>
<thead>
<tr>
<th>Mean µ</th>
<th>St. Dev. σ</th>
<th>σ/µ</th>
<th>2% threshold T₂%</th>
<th>T₂%/µ</th>
<th>(T₂%/µ - 1) × 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.55</td>
<td>2.27</td>
<td>0.22</td>
<td>16.90</td>
<td>1.60</td>
<td>60</td>
</tr>
<tr>
<td>20.29</td>
<td>3.01</td>
<td>0.15</td>
<td>28.17</td>
<td>1.39</td>
<td>39</td>
</tr>
<tr>
<td>49.85</td>
<td>4.79</td>
<td>0.10</td>
<td>60.23</td>
<td>1.21</td>
<td>21</td>
</tr>
<tr>
<td>99.65</td>
<td>6.70</td>
<td>0.07</td>
<td>112.03</td>
<td>1.12</td>
<td>12</td>
</tr>
</tbody>
</table>
Now, let us consider the process shown in Fig. 4.7. The four traces have the same time average as the corresponding traces in Fig. 4.6. However, the statistical behavior of the data is quite different. Unlike the Poisson process which with increased volume becomes smoother, the process in Fig. 4.7 retains its burstiness, even for large traffic volumes. In other words the statistical behavior of the process remains the same for several aggregation levels. This property is commonly referred to as the self-similarity. Over past decade or so, many researchers have presented statistical analyses of data traffic, which demonstrate its self-similar behavior. For comparison purposes the numerical data characterizing the traces in Fig. 4.7 is shown in Table 4.3. We see that the ratio of the 2% threshold and the average traffic volume remains approximately the same for all of the traces. Therefore, to guarantee GOS of 2% the number of allocated resources should be at least 60% higher than the average traffic volume for all cases.

### Table 4.3. Numerical data associated with the process given in Fig. 4.7

<table>
<thead>
<tr>
<th>Mean $\mu$</th>
<th>St. Dev. $\sigma$</th>
<th>$\mu/\sigma$</th>
<th>2% threshold $t_{2%}$</th>
<th>$t_{2%}/\mu$</th>
<th>$(t_{2%}/\mu - 1) \times 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.52</td>
<td>2.94</td>
<td>0.31</td>
<td>16</td>
<td>1.68</td>
<td>68</td>
</tr>
<tr>
<td>20.36</td>
<td>5.00</td>
<td>0.25</td>
<td>31</td>
<td>1.52</td>
<td>52</td>
</tr>
<tr>
<td>51.85</td>
<td>14.32</td>
<td>0.29</td>
<td>83</td>
<td>1.60</td>
<td>62</td>
</tr>
<tr>
<td>106.05</td>
<td>29.61</td>
<td>0.30</td>
<td>170</td>
<td>1.60</td>
<td>60</td>
</tr>
</tbody>
</table>

#### 4.2.3 Pareto Distribution of the Service Times

As mentioned in the previous section, the statistics of the data traffic exhibit self-similar behavior. In its most elementary interpretation, this means that the data traffic remains bursty even when aggregated from many different sources. One is naturally interested in why this is occurring. Careful analysis of the measured traffic data [9] shows that distribution of service times for data services frequently does not follow exponential distribution. Unlike voice or low-end data services (SMS, paging, e-mails with no attachments, etc.) where the main sources of information are humans, data services involve communication between humans and machines or even between machines. In such circumstances, exchanging large volumes of data is not a rare event. As an example, consider the Web browsing process. Just a couple of mouse clicks on the user's side may start an online radio service which can last for hours, or trigger download of a high resolution image file. In both cases, the communication may require network resources for a very long time and result in a transfer of huge amounts of data. Therefore, when specifying distribution of the service times for various data services it is customary to utilize some kind of a "heavy tail" probability density function. In recent literature, the most common probability density function type used to describe the distribution of the service time in data networks is the Pareto distribution. The reason resides in its relatively simple analytical form and flexible parameterization that allows fitting of a wide range experimentally obtained probability density functions.

The analytical form of the Pareto density function is given as:
where $x_0$ is the smallest message size, $x_m$ is the largest message size and $n$ is the distribution shape parameter that determines the burstiness of the data service. Figure 4.8 shows the plots of Pareto distributions for various values of the shape parameter $n$. We see that as $n$ gets smaller the tail of the distribution becomes "heavier". This results in a more pronounced self-similar behavior, and the data remains bursty even when aggregated from many different and independent sources. The level of burstiness depends on the nature of data traffic. Some typical values found through analysis of the measurements obtained for different traffic types are provided in Table 4.4. The table includes references to the published literature documenting the measurement process and obtained values.

![Pareto and exponential probability density functions](image)

**Figure 4.8.** Pareto and exponential probability density functions

<table>
<thead>
<tr>
<th>Data traffic type</th>
<th>$n$</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>WWW</td>
<td>1.1-1.5</td>
<td>[9,10]</td>
</tr>
<tr>
<td>Ethernet</td>
<td>1.1-1.4</td>
<td>[8]</td>
</tr>
<tr>
<td>FTP</td>
<td>0.9-1.4</td>
<td>[10,11,12]</td>
</tr>
<tr>
<td>E-mail (SMTP)</td>
<td>1.18-1.48</td>
<td>[10,11,12]</td>
</tr>
<tr>
<td>Rlogin and Telnet</td>
<td>1.1-1.4</td>
<td>[10,11,12]</td>
</tr>
</tbody>
</table>
As a final remark, we note that the values in Table 4.4 were obtained using the measurements on fixed landline networks. Since the rollout of the 3G services is yet to begin, similar measurements for the wireless cellular networks are still unavailable. At this time, we assume that the mobile data services and their landline counterparts exhibit similar behavior.

We demonstrate the effect of the non-exponentially distributed service times using the following example.

**Example 4.1. (M/G/1 queuing system).** The M/G/1 queuing system is a single server system with general distribution of service times. The equations that describe its performance are given in Table 4.5 [13]. We should note that the key parameter present in all equations is the ratio of the standard deviation of the service time to its mean. Intuitively speaking, this can be expected.

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Symbol</th>
<th>Value (Formula)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Offered traffic</td>
<td>$a$</td>
<td>$a = \lambda / \mu$</td>
</tr>
<tr>
<td>Server utilization</td>
<td>$\rho$</td>
<td>$\rho = a / C = a$</td>
</tr>
<tr>
<td>Scaling factor</td>
<td>$K$</td>
<td>$\frac{1}{2} \left[ \frac{1}{1 + \left( \frac{\sigma_{T_s}}{T_s} \right)^2} \right]$</td>
</tr>
<tr>
<td>Average number of users in the system</td>
<td>$L$</td>
<td>$\rho + \frac{\rho^2 K}{1 - \rho}$</td>
</tr>
<tr>
<td>Average number of users in the queue</td>
<td>$L_q$</td>
<td>$\frac{\rho^2 K}{1 - \rho}$</td>
</tr>
<tr>
<td>Average time that users spend in the system</td>
<td>$T$</td>
<td>$T_s + \frac{\rho T_s K}{1 - \rho}$</td>
</tr>
<tr>
<td>Average time that users spend in the queue</td>
<td>$T_q$</td>
<td>$\frac{\rho T_s K}{1 - \rho}$</td>
</tr>
</tbody>
</table>

A large standard deviation to mean ratio indicates a large variability of the data relative to its mean. Considering the data given in Table 4.3, we see that this is one of the main properties of the self-similar data. Of a special interest is to determine $\left( \frac{\sigma_{T_s}}{T_s} \right)^2$ for the two commonly used service time probability density functions: exponential and Pareto.

For the exponential distribution we have

\[
T_s = \int_0^{\infty} x \cdot \mu \exp(-\mu x) dx = \frac{1}{\mu} \int_0^{\infty} x \mu \exp(-\mu x) d\mu = \frac{1}{\mu} \Gamma(2) = \frac{1}{\mu} \tag{6.2}
\]
RF Engineering Continuing Education
Introduction to Traffic Planning

\[ E\{x^2\} = \int_0^{+\infty} x^2 \mu \exp(-\mu x) dx = \frac{1}{\mu^2} \int_0^{+\infty} (x\mu)^2 \exp(-\mu x) d(\mu x) = \frac{1}{\mu^2} \Gamma(3) = \frac{2}{\mu^2} \quad (6.3) \]

Therefore,
\[ \sigma_{Ts}^2 = E\{x^2\} - E\{x\}^2 = \frac{2}{\mu^2} - \frac{1}{\mu^2} = \frac{1}{\mu^2} \quad (6.4) \]

and
\[ \left( \frac{\sigma_{Ts}}{T_s} \right)^2 = \frac{1/\mu^2}{1/\mu^2} = 1 \quad (6.5) \]

For the Pareto probability density function, we have
\[ E\{x\} = \int_{x_0}^{x} x \cdot \frac{n}{1-\left(\frac{x_0}{x_m}\right)^2} \cdot \frac{x_0^n}{x^{n+1}} dx = \frac{n}{n-1} \cdot \frac{x_m}{x} \left[ 1 - \left(\frac{x_0}{x_m}\right)^n \right] - 1 \quad (6.6) \]

and
\[ E\{x^2\} = \int_{x_0}^{x} x^2 \cdot \frac{n}{1-\left(\frac{x_0}{x_m}\right)^2} \cdot \frac{x_0^n}{x^{n+1}} dx = \frac{n}{n-2} \cdot \frac{x_m}{x} \left[ 1 - \left(\frac{x_0}{x_m}\right)^n \right] - \frac{n}{n-2} \cdot \frac{x_m}{x} \left[ 1 - \left(\frac{x_0}{x_m}\right)^{n-1} \right] \quad (6.7) \]

Therefore,
\[ \left( \frac{\sigma_{Ts}}{T_s} \right)^2 = \frac{E\{x^2\}}{E\{x\}^2} - 1 = \frac{(n-1)^2}{n(n-2)} \cdot \left[ 1 - \left(\frac{x_0}{x_m}\right)^n \right] \cdot \frac{1 - \left(\frac{x_0}{x_m}\right)^{n-2}}{1 - \left(\frac{x_0}{x_m}\right)^{n-1}} - 1 \quad (6.8) \]

Equations (6.5) and (6.8) can be used to evaluate the performance of the system for different levels of self similarity. Normalizing the average service time to unity, we obtain the values given in Table 4.6.

**Table 4.6. Parameters for the service times PDSs used in Example 4.1**

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Average service time</th>
<th>Parameters</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential</td>
<td>1</td>
<td>( \mu = 1 )</td>
<td>1</td>
</tr>
<tr>
<td>Pareto (n = 1.9)</td>
<td>1</td>
<td>( x_0 = 0.5005 )</td>
<td>4.26</td>
</tr>
<tr>
<td>Pareto (n = 1.4)</td>
<td>1</td>
<td>( x_0 = 0.3049 )</td>
<td>13.47</td>
</tr>
<tr>
<td>Pareto (n = 1.1)</td>
<td>1</td>
<td>( x_0 = 0.1822 )</td>
<td>20.30</td>
</tr>
</tbody>
</table>
Using the values for the scaling factor $K$ for Table 4.6, we can examine the performance of the $M/G/1$ queuing system when presented with data of different self similarity levels. For example, Fig. 4.9 shows normalized average queuing time as a function of the resource utilization.

![Figure 4.9. Normalized average queuing time for different distribution of the service times](image)

As can be seen, as the burstiness of the traffic increases (indicated by smaller value for $n$ in the Pareto distribution), the average time that the request spends in the queue increases as well. Therefore, to guarantee the same level of performance for the busty data, the resource utilization needs to be limited.

### 4.3 Data Services in Cellular Networks

Unlike voice communication, data communication encompasses a large variety of services. Each service is characterized by a set of statistical parameters that determine its performance within a network with limited communication resources. From the standpoint of traffic planning, the most salient property of the data service is the distribution of required service times. Based on the service time distribution we can identify two groups of data services. The first group consists of data services with exponentially distributed service times. We refer to that group as the low-end data services. The second group consists of data services with service times following Pareto distribution. We refer to the second group as the high-end data services. In this section we provide some engineering guidelines for modeling and traffic planning of the two data service groups.
4.3.1 Low-End Data Service in Data Only Networks

The low-end data services are usually implemented as low rate data communication involving humans at both ends. Typical representatives of such service are short message service, e-mail without attachments, dispatch voice and voice mail. All of the above services are characterized with a large number of short length messages. Large messages are rare and probability of having a large messages decreases rapidly as a function of the message size. This kind of service time distribution can be approximated as exponential and that makes the theory developed in Section 3.2.2 applicable. In other words, the low-end data services can be successfully modeled using the Poisson process assumptions and Erlang C based traffic planning formulas. As an illustration, consider the distribution of the message lengths for an e-mail service with no attachments shown in Fig.4.10. The exponential nature of the distribution is evident.

![Figure 4.10. Histogram of e-mail messages with no attachments](image)

In contemporary cellular systems data services may be either allocated a separate set of communication resources or they can share the same resources with the circuit switched voice. Analysis and dimensioning of data only networks is somewhat easier; for low-end data services, it can be accomplished using the Erlang C based formulas. Since this topic was thoroughly covered in Section 3.2.2, in this section we just illustrate the process through the following example.

**Example 4.2.** Consider a cellular system that has two downlink control channels for delivery of short messages. Let us assume that the average rate of message arrivals is $\lambda = 20$ messages per second. The message lengths are exponentially distributed with average length of 1 Kb and the average data rate on the channel is 15 Kb/sec. Calculate the following:

1. Probability of message delay
2. Average delay for all messages
3. Average delay for delayed messages
4. Probability of a message delay exceeding 1 second

Since the message lengths are distributed exponentially, the traffic model that can be used is Erlang C. The formulas for this model are summarized in Table 3.2.

The average death rate in the system is given by

\[ \mu = \frac{R}{L} = \frac{15 \text{ kb/sec}}{1 \text{ kb}} = 15 \text{ sec}^{-1} \]

Therefore, the offered traffic and average server utilization can be calculated as

\[ a = \frac{\lambda}{\mu} = \frac{20}{15} = 1.333 \text{ erlangs} \]
\[ \rho = \frac{a}{C} = \frac{1.333}{2} = 0.667 \]

Probability of having zero messages in the system queue

\[ p_0 = \left\{ \sum_{k=0}^{c-1} \frac{a^k}{k!} + \frac{a^c}{C!(1 - \rho)} \right\}^{-1} = \left\{ 1 + 1.333 + \frac{1.333^2}{2!(1 - 0.667)} \right\}^{-1} = .2 \]

The probability of the message delay can be calculated as

\[ \Pr(> 0) = \frac{a^c}{C!(1 - \rho)} \cdot p_0 = \frac{1.333^2}{2!(1 - 0.667)} \cdot 0.2 = 0.5336 \]

Therefore, approximately 53.36% of the messages are delayed. The average delay for all messages is

\[ D_1 = \Pr(> 0) \cdot \frac{H}{C - a} = \Pr(> 0) \cdot \frac{1/\mu}{C - a} = 0.5336 \cdot \frac{1/15}{2 - 1.333} = 0.0534 \text{ second} \]

The average time that delay messages spend in the queue

\[ D_2 = \frac{H}{C - a} = \frac{1/15}{2 - 1.333} = 0.1 \text{ second} \]

Probability of a message delay exceeding 1 second is

\[ \Pr(t > 1) = \Pr(t > 1) \cdot \exp[-\mu(C - a)] = 0.5336 \cdot \exp[-15 \cdot 1 \cdot (2 - 1.333)] = 2.41 \times 10^{-5} \]
4.3.2 Low –End Data Services in Mixed Traffic Networks

Some cellular systems implement the packed data service over the same communication resources that are used for the circuit switched voice service. Such systems can be seen as priority systems. The resources are allocated to the data service only when they are not used by the circuit switched voice. This concept is illustrated in Fig. 4.11.

**Figure 4.11.** Sharing the communication resources between voice and data

A typical cellular network is dimensioned to satisfy a GOS of 2% for the circuit switched voice within the busy hour. This results in a relatively large percentage of idle time. The resources that are idling on the voice side can be used for the support of data traffic. Figure 4.12 shows the capacity in erlangs that is available for the data service at a given value of the voice traffic and the GOS requirement.

**Figure 4.12.** Distribution of erlang capacity between voice and data
Use of the curves shown in Fig 4.12 can be illustrated with the following example.

**Example 4.3.** Consider an implementation of a low-end data service in a cellular system with 24 digital trunks that is dimensioned to handle circuit switched voice traffic with GOS of 2% within the busy hour. Estimate the average erlang capacity available for data channels. If the data rate of a channel is 8 Kb/sec, estimate the average busy hour data capacity in Kb/sec. Repeat the calculation assuming a GOS of 1%.

At 2% GOS, 24 trunks can handle 16.63 E of circuit switched voice traffic. The average erlang capacity of the data service is given as:

\[ a_{\text{data}} = 24 - 16.63 = 7.37 \text{ E} \]

The data capacity in Kb/sec can be calculated as

\[ C_{\text{data}} = 7.37 \cdot 8 = 58.96 \text{ Kb/sec} \]

If the system is dimensioned with the GOS of 1% we obtain

\[ a_{\text{voice}} = 15.29 \text{ E} \]
\[ a_{\text{data}} = 24 - 15.29 = 8.71 \text{ E} \]
\[ C_{\text{data}} = 69.64 \text{ Kb/sec} \]

### 4.3.3 High-End Data Services

The high-end data services are characterized with non-exponential distributions of service times. Typical examples of the high-end data services are file transfer, streaming audio and video, Web browsing and FTP. All of the high-end data service types are characterized with high level of data burstiness. Since the service times are not exponentially distributed, Poisson modeling cannot be applied.

Accurate traffic modeling of the high end data services is a subject of intensive research. An excellent survey of the existing methods can be found in [14]. The approach that is presented in this report is based on the general formula for queuing time estimation of the GI/G/C\(^5\) type queuing systems. The formula is developed by Arnold Allen and John Cunneen [1], and is given by

\[
T_q = \frac{E_c[C, a]}{2C(1 - \rho)} \cdot \left\{ \left( \frac{\sigma_{T_a}}{T_a} \right)^2 + \left( \frac{\sigma_{T_s}}{T_s} \right)^2 \right\}
\]

(6.9)

where

\[ T_q \] - average message delay time

---

5 Following Kandell's notation rules, GI/G/C signifies general and independent distribution of the interarrival times, general distribution of service times, C trunks, infinite queuing capacity, infinite population and FIFO queuing discipline.
The formula in (6.9) has been developed as a generalization of experimental results and not by any formal proof. However, it provides reasonably good results for many queuing systems of practical interest while still being relatively easy to compute. To examine the queuing delay prediction that is obtained by using (6.9), consider the following example.

**Example 4.3.** Use the formula given in (6.9) to predict the average queuing delay in the following queuing systems $M/M/C$ and $M/G/1$. Using the estimate of the delay, determine the expressions for the average throughput in the number of requests per unit time.

For $M/M/C$ we have

$$\frac{\sigma_{T_a}}{T_a} = 1, \quad \text{and} \quad \frac{\sigma_{T_s}}{T_s} = 1 \quad \text{(see Example 4.1, Equation (6.5))}$$

Thus

$$T_q = \frac{E_c[C,a]T_s}{2C(1-\rho)}(1+1) = \frac{E_c[C,a]}{C(1-\rho)}T_s = \frac{E_c[C,a]}{C-a}$$

which is identical to the expression provided in Table 3.2. Therefore, in the case of $M/M/C$, (6.9) gives the exact estimate for the queuing time.

For $M/G/1$ we have

$$\frac{\sigma_{T_a}}{T_a} = 1$$

and (6.9) becomes

$$T_q = \frac{E_c[C,1]T_s}{2C(1-\rho)}\left[ 1 + \frac{\sigma_{T_s}^2}{T_s^2} \right] = \frac{aT_s}{2(1-\rho)} \cdot \left[ 1 + \frac{\sigma_{T_s}^2}{T_s^2} \right] = \frac{\rho}{1-\rho}T_sK$$

which is the same expression provided in Table 4.5. Therefore, for the case of $M/G/1$, (6.9) gives the exact estimate as well.

In both cases, the throughput for number of messages per unit time can be determined as

$$R = \frac{1}{T_s + T_q}$$
The formula given in (6.9) can be used for traffic planning of high-end data services in cellular networks. The procedure is essentially identical to that used in the case of low-end data services discussed in previous sections the only difference being the probability distribution function of the service time. As discussed in Sections 4.2.2 and 4.2.3, for most data services types, we find that PDF of service times can be approximated as Pareto with properly chosen distribution parameters. Standard deviation to mean ratio for the Pareto distribution can be calculated in accordance to (6.8) while some typical values for the distribution shaping parameter are provided in Table 4.4. Detailed explanations of high-end data service dimensioning process for GPRS systems is provided in Sections 4.4. Here, we will demonstrate the procedure using the following example.

**Example 4.4. (E-mail service over CDPD).** Cellular Digital Packed Data (CDPD) is a packed data service developed to work with existing AMPS cellular infrastructure. It uses 30Khz radio channels and it provides native data rate of 19.2 Kb/sec in a packed data mode. Due to various overhead bits as well as the error correction coding, the rate of the CDPD channels as seen by the application is on the order of 9.6Kb/sec. Most commonly, the CDPD service is implemented through a set of separate radios. In this configuration, the system may be regarded as a data only system.

Let us consider an implementation of e-mail service over CDPD channels. Let us assume that the service follows Pareto distribution with parameters listed in Table 4.7. If the cell site has 2 CDPD radios and if the average number of service requests is 10 per minute, calculate the average throughput experienced by an individual user.

<table>
<thead>
<tr>
<th>Table 4.7. Parameters for e-mail service used in example 4.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
</tr>
<tr>
<td>Shape parameter</td>
</tr>
<tr>
<td>Minimum message size</td>
</tr>
<tr>
<td>Maximum message size</td>
</tr>
<tr>
<td>Average message size</td>
</tr>
</tbody>
</table>

The average throughput experienced by an individual user can be calculated as

$$R_{av} = \frac{\bar{x}}{T_s + T_q}$$

where $\bar{x}$ is the average message length, $T_s$ is the average service time and $T_q$ is the average time that a user service request spends in the queue. The average service time is a function of the effective CDPD channel and it can be calculated as:

$$T_s = \frac{\bar{x}}{R_0} = \frac{3.32}{9.6} = 0.35 \text{ sec}$$

---

6 For Pareto distribution $\bar{x} = \frac{n}{n-1} \cdot \frac{x_0^n}{1-\left(\frac{x_0}{x_m}\right)^n} \cdot \left[ \frac{1}{x_0^{n-1}} - \frac{1}{x_m^{n-1}} \right]$
The average queuing time can be estimated using Allen-Cunneen's formula, that is

\[ T_q = \frac{E_c[C,a,T_s]}{2C(1-\rho)} \left[ 1 + \left( \frac{\sigma_{T_s}}{T_s} \right)^2 \right] \]

In this particular case

\[
C = 2
\]
\[
a = \frac{\lambda \cdot \bar{x}}{R_0} = 10 \frac{1}{60} = 0.35 = 0.0583
\]
\[
\rho = \frac{a}{C} = \frac{0.0583}{2} = 0.0292
\]
\[
\left( \frac{\sigma_{T_s}}{T_s} \right)^2 = \frac{(1.4-1)^2}{1.4 \cdot (1.4-2)} \left[ 1 - \left( \frac{1}{1600} \right)^{1.4} \right] \cdot \left[ \frac{1}{1 - \left( \frac{1}{1600} \right)^{1.4-2}} - 1 \right] = 300.057
\]

Therefore, the queuing time is calculated as

\[ T_q = \frac{0.0017 \cdot 0.35}{2 \cdot 2 \cdot (1-0.0292)} \cdot [1 + 300.05] = 0.45 \]

Finally, the throughput per user becomes

\[ R_{av} = \frac{\bar{x}}{T_s + T_q} = \frac{3.32}{0.35 + 0.45} = 4.15 \text{ Kb/sec} \]

We see, that due to congestion, each of the users receives throughput which is less than half the data rate that can be delivered through CDPD channel.

### 4.4 Optional: Traffic Planning for Packet Data in GSM/GPRS Networks

The GPRS is the packet data services implemented in GSM mobile networks. It is designed to allow for smooth and cost effective transition of GSM towards the mobile network of the third generation. The implementation of GPRS over GSM creates a network with mixed traffic types. Also, a GSM/GPRS network implements service priority since the circuit switched voice has the service precedence over the packet data traffic. Within the GPRS, service prioritization is supported between various data traffic types as well. However, at least in the initial deployment stages, the data service is provided on a best effort basis.

---

7 Using standard deviation to mean ratio for the Pareto distribution given in (6.8)
Within a cellular network that supports voice and data over the same communication channels, the highest traffic loading does not have to coincide with the busy hour for either data or voice service. A case in point is illustrated in Fig. 4.13.

As shown in Fig. 4.13, to perform a proper dimensioning of a GPRS network we need to estimate the following values:

1. Peak circuit switched traffic in erlangs
2. Peak packet data traffic in Kb/sec
3. Circuit switched traffic load during the total traffic peak
4. Packet data traffic load during the traffic peak

For the analysis presented in this section, we assume that the above listed estimates are available. An example of the methodology that can be used to estimate the level of the circuit switched load is provided in Section 3.4. On the other hand, in the case of data services estimation of the traffic load is considerably dependent on the operator's business plan and marketing strategy.

The GPRS supports four different error control coding schemes: CS-1 to CS-4 (see Table 4.8). The coding schemes offer different levels of data protection at the expense of the available date rate. Which coding scheme is used depends on the quality of the radio channel. As the quality of the channel degrades (most often as a result of cochannel interference), the coding is changed to a coding scheme offering a higher level of error correction capability. When performing the traffic dimensioning, we usually assume that the system operates at a high loading point at which it needs to cope with a high level of self-induced interference. For that reason the dimensioning is usually performed assuming the coding scheme CS-1. Correction of the obtained results for other coding schemes is relatively straightforward. As an illustration, consider the following example.
**Example 4.4.** Consider a GSM/GPRS base station designed to support the aggregate data throughput of 40Kb/sec using coding scheme CS-1. Assume that 50% of users with the cell site's coverage area operate under favorable RF conditions that allow them to use coding scheme CS-2. Estimate the aggregate throughput that can be supported by the base station.

The aggregate throughput can be calculated as

\[ R_{agg} = 0.5 \cdot 40 + 0.5 \cdot \frac{13.4}{9.05} \cdot 40 = 49.62 \text{ [Kb/sec]} \]

**Table 4.8.** Summary of the GPRS coding schemes

<table>
<thead>
<tr>
<th>Coding Scheme</th>
<th>Code type [15]</th>
<th>Data rate [Kb/sec]</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS-1</td>
<td>1/2 convolutional</td>
<td>9.05</td>
</tr>
<tr>
<td>CS-2</td>
<td>2/3 convolutional</td>
<td>13.4</td>
</tr>
<tr>
<td>CS-3</td>
<td>3/4 convolutional</td>
<td>15.6</td>
</tr>
<tr>
<td>CS-4</td>
<td>no coding</td>
<td>21.4</td>
</tr>
</tbody>
</table>

**4.4.1 Estimation of the GPRS Data Capacity**

Due to GSM's TDMA air interface architecture, the number of trunks assigned to a base station is a multiple of eight. In GSM, the trunk is essentially a time slot on one of the GSM transceivers. Some of the time slots are permanently allocated to control and signaling information, therefore, they are not available for carrying either voice or data traffic. The number of the channels dedicated to the overhead is configurable and Table 4.9 provides a summary of some possible configurations. The table also provides the level of circuit switched traffic that can be supported at GOS of 1 and 2%.

**Table 4.9.** Some common configurations of a GSM base station

<table>
<thead>
<tr>
<th>Number of transceivers</th>
<th>Number of time slots available for traffic</th>
<th>Voice capacity at GOS of 1% [E]</th>
<th>Voice capacity at GOS of 2% [E]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>1.91</td>
<td>2.28</td>
</tr>
<tr>
<td>2</td>
<td>14</td>
<td>7.35</td>
<td>8.20</td>
</tr>
<tr>
<td>3</td>
<td>22</td>
<td>13.65</td>
<td>14.90</td>
</tr>
<tr>
<td>4</td>
<td>30</td>
<td>20.34</td>
<td>21.93</td>
</tr>
</tbody>
</table>

The objective of traffic dimensioning in GPRS networks may be formulated in many different ways. The approach presented here is trying to estimate the maximum aggregate data traffic that can be supported per GSM/GPRS sector under the following assumptions:

1. Coding scheme is CS-1
2. Threshold throughput per time slot is 5 Kb/sec
3. Dominant type of data service is WWW browsing, which means that the distribution of service times is Pareto with an appropriately chosen set of parameters.
The average throughput per GPRS time slot can be estimated as

\[ R_{av} = R_0 \frac{T_s}{T_s + T_q} = \frac{R_0}{1 + T_q/T_s} \]  \hspace{1cm} (6.10)

where \( R_0 \) is the maximum data rate available at a given GPRS coding scheme (for CS-1, this rate is 9.05 Kb/sec), \( T_s \) is the average service time per one user service request, and \( T_q \) is the average time that a request spends waiting within the system's queue. We assume that the size of the queue is large so that there is no loss of data due to insufficient buffer sizes. In addition, the service requests are arriving randomly and independently of each other. In essence, we are modeling the GPRS as a queuing system of the \( M/G/C \) type where \( C \) is the average number of time slots available for the packet data. In section 4.3.3 we showed that the average queuing time in \( GI/G/C \) systems can be estimated using Allen-Cunnen's formula, which is given as [1]:

\[ T_q = \frac{E_C [C_D \cdot a_D] T_s}{2C_D (1 - \rho)} \left\{ 1 + \left( \frac{\sigma_{Ts}}{T_s} \right)^2 \right\} \]  \hspace{1cm} (6.11)

where \( a_D \) is the average data traffic load in erlangs and \( C_D \) is the average number of time slots that are available for GPRS service.

Using (6.10) and (6.11) we obtain

\[ \frac{R_0}{R_{av}} = 1 + \frac{E_C [C_D \cdot a_D]}{2C_D (1 - \rho)} \left\{ 1 + \left( \frac{\sigma_{Ts}}{T_s} \right)^2 \right\} \]  \hspace{1cm} (6.12)

or

\[ \frac{R_0}{R_{av}} = 1 + \frac{E_C [C_D \cdot a_D]}{2(C_D - a_D)} \left\{ 1 + \left( \frac{\sigma_{Ts}}{T_s} \right)^2 \right\} \]  \hspace{1cm} (6.13)

Since the dominant part of the data traffic is the WWW browsing, we assume that the distribution of the service times is given as Pareto. Standard deviation to mean ration for the Pareto distribution is given in (6.8). Substituting (6.8) into (6.13), we obtain

\[ \frac{R_0}{R_{av}} = 1 + \frac{E_C [C_D \cdot a_D]}{2(C_D - a_D)} \left\{ (n-1)^2 \frac{(n-1)}{n(n-2)} \left[ 1 - \left( \frac{x_0}{x_m} \right)^n \right] \right\} \left[ 1 - \left( \frac{x_0}{x_m} \right)^{a-2} \right] \]  \hspace{1cm} (6.14)

Equation (6.14) allows us to calculate the erlang data capacity of the GPRS site. To illustrate the calculation procedure, consider the numerical data given in Table 4.10.
Table 4.9. Numerical data characterizing WWW browsing data service

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Symbol</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Threshold throughput</td>
<td>$R_{av}$</td>
<td>5 Kb/sec</td>
</tr>
<tr>
<td>Pareto distribution shape</td>
<td>$n$</td>
<td>1.1</td>
</tr>
<tr>
<td>parameter</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minimum message size</td>
<td>$x_0$</td>
<td>1 Kb</td>
</tr>
<tr>
<td>Maximum message size</td>
<td>$x_m$</td>
<td>1.6 Mb</td>
</tr>
</tbody>
</table>

Using (6.14), and numerical data from Tables 4.8 and 4.9, we obtain the results shown in Table 4.11.

Table 4.11. Calculation of the GPRS capacity for GOS of 2% and 1%

<table>
<thead>
<tr>
<th># TX</th>
<th># TS</th>
<th>GOS of 2%</th>
<th>GOS of 2%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Voice</td>
<td>Data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>traffic</td>
<td>traffic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[E]</td>
<td>[E]</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>2.28</td>
<td>3.72</td>
</tr>
<tr>
<td>2</td>
<td>14</td>
<td>8.20</td>
<td>5.80</td>
</tr>
<tr>
<td>3</td>
<td>22</td>
<td>14.90</td>
<td>7.10</td>
</tr>
<tr>
<td>4</td>
<td>30</td>
<td>21.93</td>
<td>8.07</td>
</tr>
</tbody>
</table>

Figure 4.14. Aggregate CS-1 capacity for different base station configurations

$R_{agg}$ is the aggregate throughput that can be achieved using the coding scheme CS-1.
Table 4.11 provides GPRS capacity estimates for systems that are designed to handle voice service at a given GOS requirement. In other words, no additional radios are introduced in support of the data service only. For example if a sector has 3 radios, according to Table 4.11, it can support 14.90 E of voice traffic at 2% GOS and about 36.47 Kb/sec of additional aggregate CS-1 throughput on the packet data side.

Equation (6.14) can be used to solve for GPRS capacity for any allowed combination of number of radios and circuit switched load. Using numerical data provided in Table 4.9, we generate a family of curves shown in Fig. 4.14. The curves can be used for the traffic dimensioning of GPRS sites. We will illustrate the procedure through the following examples.

**Example 4.5.** Determine the number of GSM radios at the GSM/GPRS site required to support at least 7.5 erlangs of voice traffic at 2% GOS and aggregate CS-1 packet data throughput of 80 Kb/sec. If 50% of the users are in area of C/I that allows for CS-2 coding, what is the available aggregate throughput?

The point on the curves that satisfies both requirements is (9.7E, 80 Kb/sec) – See Fig. 4.15. Therefore, the site requires at least three GSM radios.

If 50% of the mobiles are using the CS-2 coding scheme, the aggregate throughput can be estimated as

\[
R_{agg} = 0.5 \cdot 80 + 0.5 \cdot \frac{13.4}{9.05} \cdot 80 = 99.23 \text{ Kb/sec}
\]

This corresponds to an increase of 24%. Since the coding scheme depends primarily on the distribution of C/I, it should be obvious that the optimization of the frequency plan and the interference control have major impact on the GPRS capacity.

**Figure 4.15.** Dimensioning of the GSM/GPRS site using the capacity curves in Fig. 4.14
**Example 4.6.** Consider a GSM provider trying to role out the GPRS service within an urban core area. We will assume that the traffic is following fairly uniform geographical distribution and that the most dominant traffic type is WWW browsing (curves in Fig. 4.14 can be used). Other relevant data is given in Table 4.12.

**Table 4.12.** Numerical data for Example 4.6

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amount of spectrum</td>
<td>5 MHz</td>
</tr>
<tr>
<td>Reuse strategy</td>
<td>(N = 4/12)</td>
</tr>
<tr>
<td>Peak circuit switched load</td>
<td>420 E</td>
</tr>
<tr>
<td>Site configuration</td>
<td>3-sectored</td>
</tr>
</tbody>
</table>

Determine:

1. Number of sites necessary for handling the circuit switched voice at 2% GOS.
2. Aggregate GPRS capacity per site (for CS-1 coding scheme).
3. If the system is to provide aggregate capacity of 2.5 Mb/sec while serving the peak voice traffic load, how many sites need to be installed?

1. The provider operates in 5 MHz of spectrum, the total number of GSM/GPRS ARFCN’s can be determined as

\[
N_{ARFCN} = \frac{5000 - 200}{200} = 24
\]

Since the frequency reuse scheme is \(N = 4/12\), there are two radios per sector. Assuming fourteen time slots for voice and data traffic (see Table 4.8), the total voice capacity per sector at 2% GOS is given as 8.20 E. Therefore, a single GSM site in this system can accommodate \(3 \times 8.20 = 24.60\) erlangs of voice traffic. The number of sites necessary to provide service for 420 erlangs is given as

\[
N_{CELL} = \frac{420}{24.6} = 17.07 \rightarrow 18
\]

2. Using the curves in Fig. 4.14 for a site with 2 radios and circuit switched voice traffic of 8.2E we obtain the GPRS capacity of 28.33 Kb/sec. Therefore, the aggregate GPRS capacity for the system is given as

\[
R_{agg} = 28.33 \times 18 \times 3 = 1.529 \text{ Mb/sec}
\]

3. If the system is to serve 2.5Mb/sec of the aggregate traffic the following equations need to be satisfied.
\[ N_{CELL} \cdot a_v \geq 420 \text{ E} \quad (6.15) \]
\[ N_{CELL} \cdot R_D \geq 2500 \text{ Kb/sec} \quad (6.16) \]

Substituting (6.16) into (6.15) we obtain
\[
\frac{R_D}{a_v} \leq 5.95 \quad (6.17)
\]

Equation (6.17) is plotted in Fig. 4.16. The intersection of (6.17) and the throughput curve for the two GPRS radios gives the desired operating point for the site. At the desired operating point the voice traffic capacity is 6.8E and the GPRS throughput available per site is 40 Kb/sec. Therefore the number of sites is given as
\[
N_{CELL} = \frac{1}{3} \cdot \frac{420}{6.8} = 20.58 \to 21
\]

As a test, we determine the aggregate GPRS throughput as
\[
R_{agg} = 21 \cdot 40 \cdot 3 = 2520 \text{ Kb/sec}
\]

Figure 4.11. Determining the operating point for a GSM site in Example 4.6
5 Traffic Trending and Growth Planning

The purpose of traffic trending is to predict the growth of the traffic in cellular networks. Accurate predictions of the traffic growth help us anticipate a shortage of traffic resources before it actually happens. Knowing when the resource shortage will occur allows for appropriate resource provisioning. In other words, an RF engineer can assure the timely purchase of additional radio resources, devise a new frequency plan, and develop an optimal strategy for system growth. This way, the system can grow without compromising the service quality. Most of the switch data processing software tools provide routines that can be readily used for traffic trending. However, the effectiveness of these routines usually depends on the engineer's ability to interpret the data and assure its consistency. In this section, we briefly review the process of traffic trending and provide some basic theory and practical examples of how it can be done.

5.1 Linear Regression

Trending of the traffic data is usually done using linear regression. In this section, we provide description of this simple but powerful technique.

Consider the measurement of traffic volume within the busy hour for an imaginary cell presented in Fig. 5.1. We can see that the points follow a trend and the task is to project the traffic volume at some time in advance. By doing so, we can assure that before the capacity of the cell is exceeded, we introduce some new communication resources so that the GOS is kept below the desired target. The amount of data in this example is far from sufficient (it is available for only fifteen consecutive days), but it is used for illustrative purposes. A more realistic example will be provided in the next section. To perform the projection of the traffic we assume that there exist a linear relationship between the time and the volume of the served traffic. In other words, we assume that the traffic at any given day's busy hour can be calculated as

\[ y = mx + b \] (5.1)

where \( m \) and \( b \) are parameters that need to be estimated from the measurement data. The data used to generate the plot in Fig. 5.1 are given in Table 5.1.

<table>
<thead>
<tr>
<th>Point</th>
<th>Value</th>
<th>Point</th>
<th>Value</th>
<th>Point</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.94</td>
<td>2</td>
<td>5.96</td>
<td>3</td>
<td>6.08</td>
</tr>
<tr>
<td>4</td>
<td>5.97</td>
<td>5</td>
<td>6.16</td>
<td>6</td>
<td>6.16</td>
</tr>
<tr>
<td>7</td>
<td>6.04</td>
<td>8</td>
<td>6.11</td>
<td>9</td>
<td>6.04</td>
</tr>
<tr>
<td>10</td>
<td>5.96</td>
<td>11</td>
<td>6.30</td>
<td>12</td>
<td>6.17</td>
</tr>
<tr>
<td>13</td>
<td>6.27</td>
<td>14</td>
<td>6.28</td>
<td>15</td>
<td>6.28</td>
</tr>
<tr>
<td>16</td>
<td>6.20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Suppose, that we have found the slope and intercept of the ‘best fit’ line. We can use (5.1) to predict what is the value of \( y \) at point \( x_i \). That is:

\[ \hat{y}_i = mx_i + b \]  

(5.2)

On the other hand, from our measurements we know that the value of \( y \) at data point \( x_i \) is \( y_i \). The error we make if we use the straight line approximation is given by:

\[ e_i = \hat{y}_i - y_i = mx_i + b - y_i \]  

(5.3)

![Figure 5.1. Measurements of busy hour traffic volume at the imaginary cell site](image)

Since we are not concerned with the sign of the error we will consider the square of (5.3):

\[ e_i^2 = (mx_i + b - y_i)^2 \]  

(5.4)

The total error that we make for all of the points is given as:

\[ E = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (mx_i + b - y_i)^2 \]  

(5.5)

The error function in (5.5) depends entirely on slope \( m \) and intercept \( b \). We want to choose \( m \) and \( b \) in such a manner so that (5.5) is minimized. Taking the partial derivatives of (5.5) and solving for \( m \) and \( b \) we obtain:

\[ m = \frac{AN - BC}{ND - B^2} \]  

(5.6)
and

\[ b = \frac{DC - BA}{ND - B^2} \]  

(5.7)

where

\[ A = \sum_{i=1}^{N} x_i y_i, \quad B = \sum_{i=1}^{N} x_i, \quad C = \sum_{i=1}^{N} y_i, \quad D = \sum_{i=1}^{N} x_i^2, \]  

(5.8)

and \( N \) is the total number of points.

For our specific example, using the data provided in Table 5.1, we obtain

\[
\begin{align*}
A &= 0 \cdot 5.94 + 1 \cdot 5.96 + \cdots + 15 \cdot 6.20 = 741.17 \\
B &= 0 + 1 + 2 + \cdots + 15 = 120 \\
C &= 5.94 + 5.96 + \cdots + 6.20 = 97.92 \\
D &= 0 + 1^2 + 2^2 + \cdots + 15^2 = 1240 \\
N &= 16
\end{align*}
\]

Therefore,

\[ m = \frac{741.17 \cdot 16 - 120 \cdot 97.92}{16 \cdot 1240 - 120^2} = 0.036 \]  

(5.10)

and

\[ b = \frac{1240 \cdot 97.92 - 120 \cdot 741.17}{16 \cdot 1240 - 120^2} = 5.971 \]  

(5.11)

The coefficients in (5.10) and (5.11) can be used to predict the traffic demand. For example, at the end of the next fifteen days, the predicted traffic volume is given as

\[ a(15 + 15) = 0.036 \cdot 30 + 5.971 = 7.051 [\text{erlangs}] \]  

(5.12)

Knowing the projected traffic load and assuming that the system is dimensioned to operate below a GOS of 2%, we can determine the number of necessary channels. In this case, using the Erlang B table we determine that the imaginary cell needs 13 channels.

### 5.2 Special Events and Seasonal Trends

The data presented in Fig. 5.1 and Table 5.1 do not represent real measurements of the cell site traffic. In reality the variations in the traffic volume over the given period of time are considerably larger. Some of the variations may be explained by the randomness of the traffic process but we can identify several other reasons as well. Most of these reasons can be classified as either special events or seasonal trends.

The special events can be seen as events that lead to extraordinary change in the cellular traffic volume over a relatively short period of time. Typical examples of what can be considered as special events are traffic jams and delays on the highway, bad weather, accidents, important sports competitions, political or other social gathering. As an example consider a part of call
statistics data for a site in the Seattle area recorded during the earthquake on February 28, 2001. The earthquake occurred at 9:54 AM local time and, as can be seen from Table 5.2, it resulted in an almost 20 fold increase in call origination (267 to 5555 call attempts per hour). Since the site has a limited number of traffic resources, the increase in traffic is accompanied by a sharp degradation in system performance. The dropped call rate increased from 0.39 to 2.98% and the blocked call rate reached almost 6%.

Table 5.2. Part of the cell site traffic report for a site in the Seattle area during 2001 earthquake

<table>
<thead>
<tr>
<th>Hour</th>
<th>Orig. failure</th>
<th>%Drops</th>
<th>%Block</th>
<th>Attempts</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>0.00%</td>
<td>0.39%</td>
<td>0.00%</td>
<td>267</td>
</tr>
<tr>
<td>9</td>
<td>0.84%</td>
<td>0.90%</td>
<td>0.00%</td>
<td>346</td>
</tr>
<tr>
<td>10</td>
<td>53.03%</td>
<td>1.68%</td>
<td>3.77%</td>
<td>1327</td>
</tr>
<tr>
<td>11</td>
<td>53.89%</td>
<td>2.98%</td>
<td>5.92%</td>
<td>5555</td>
</tr>
<tr>
<td>12</td>
<td>14.58%</td>
<td>1.45%</td>
<td>0.00%</td>
<td>2873</td>
</tr>
</tbody>
</table>

The seasonal trends are seen as a longer term variations of the traffic volume. They are usually results of seasonal changes in the user density within the cell site's coverage area. As an example, consider a beach site in the winter period. The number of calls and accordingly the carried traffic for this site will be very low. However, the situation will change drastically during the summer months.

Special events and seasonal trends are treated differently from a planning standpoint. Since they are largely unpredictable, special events are usually ignored. The data recorded during special events is usually eliminated from the trending process. On the other hand, the seasonal trends are highly predictable and a well-dimensioned cellular system should maintain a high degree of performance consistency regardless of the season.

Example 5.1. Let us consider the data in Table 5.1. Assume that due to some special event the traffic recorded on the 4\textsuperscript{th} day is changed from 5.97 E to 8.73 E. Using all data points, perform the traffic trending and predict the traffic volume at the end of the next fifteen days.

Repeating the trending process we obtain

\[ A = \sum_{i=1}^{16} x_i y_i = 749.39 \]
\[ B = \sum_{i=1}^{16} x_i = 120 \]
\[ C = \sum_{i=1}^{16} y_i = 100.68 \]
\[ D = \sum_{i=1}^{16} x_i^2 = 1240 \]
\[ N = 16 \]
\[ m = \frac{AN - BC}{ND - B^2} = \frac{749.39 \cdot 16 - 120 \cdot 100.68}{16 \cdot 1240 - 120^2} = -0.017 \]
RF Engineering Continuing Education
Introduction to Traffic Planning

\[ b = \frac{DC - BA}{ND - B^2} = \frac{1240 \cdot 100.68 - 120 \cdot 749.39}{16 \cdot 1240 - 120^2} = 6.418 \]  \hspace{1cm} (5.13)

and

\[ d(15 + 15) = -0.017 \cdot 30 + 6.418 = 5.915 \text{ [erlangs]} \]  \hspace{1cm} (5.14)

Comparing (5.12) and (5.14) we notice that they are quite different (approximately 17%). The difference is a result of the special event point. In general, linear regression is quite sensitive to the "outliers". For that reason, to obtain accurate trending results, special points are usually eliminated from the measured data.

5.3 Case Study: Growth planning for a site in a cellular network

In this section, we present the example of traffic trending using real data traffic measurements obtained at a commercially operated site. The data used for trending in this case study is presented in Fig. 5.2. The measurements are performed on a site in a relatively busy area of a fast growing PCS system.

![Figure 5.2. Measurements of busy hour cell site traffic](image)

Even at the first glance we notice that the real data show entirely different behavior from our imaginary measurements discussed in the previous section. The most dominant feature is its seven day periodicity which reflects the dependence of the traffic volume on the working week cycle. Also, the data has several peaks that can be attributed to special events. Finally, although the increase trend is quite noticeable, we may observe the difference in the traffic increase rate before and after January 1st. We perform the trending of this data with the goal to estimate the number of trunks required so that the GOS is kept below 2% until June 30th.
From Fig. 5.2, we see substantial difference between traffic volumes recorded on weekdays and weekends. Since the traffic planning is performed to guarantee satisfactory GOS within the busy hour of a working day, the first step in the data trending is elimination of the weekend data. Figure 5.3 shows the same data that is presented in Fig 5.2 with the weekend data points eliminated. Two things are readily noticed. First, overall variability of the data is reduced and we can easily identify two linear segments. Second, even after removal of the weekend data, there are several data points that seem to deviate substantially from the general trends.

The analysis of these points reveals that they were recorded during the holidays. During the Thanksgiving and Christmas holidays, the cellular traffic is decreased. On the other hand, Halloween is characterized with traffic volume that is almost two times greater when compared to the levels recorded during the preceding and following weeks. For traffic planning purposes, the holidays are usually treated as special events and are eliminated from the measured data. After this is done, we obtain the graph shown in Fig. 5.4. On this graph, we can clearly identify two segments. The break point between the segments is January 1st. The data before the break point show a slower growth. In addition, integration of a microcell that occurred on January 1st has decreased the traffic served by the cell. In short, due to various reasons, the trend before and after the breakpoint is quite different. Since the configuration of the network is changed, the data before the breakpoint is eliminated from the trending and we obtain the graph shown in Fig. 5.5.

Using the linear trending methodology described in the previous section we determine the parameters of the best-fit line. The slope of the line is calculated as 0.1014 E/day and the intercept is 7.909 E. Extrapolating the line to June 30th we obtain a traffic projection of 26.16 E. Assuming a GOS of 2% and using the Erlang B table we obtain the number of required resources as 35.
Figure 5.4. Traffic measurement from Fig. 5.3 after the elimination special events

Figure 5.6. Projection of the traffic volume for the cellular site
6 Methods for Traffic Congestion Control in Cellular Systems

In the initial stage of a cellular system commercial deployment, the volume of served traffic is relatively small. During this stage a majority of operational problems are related to cell settings on both the RF and system parameters side. For example, initial settings for neighbor list, handoff thresholds or transmit power are a result of system planning and are usually derived from CAD network planning tools. Due to the limited accuracy of such tools, the system settings cannot be perfect. As a result, the initial stage of system deployment is used to determine more appropriate values for various RF and parameter settings. New values are commonly determined through the analyses of field collected and switch reports.

As a cellular system matures, the volume of served traffic is increased. As a result, the system starts experiencing congestion problems. Since cellular traffic distribution throughout the system is non-uniform, the congestion does not occur everywhere in the system simultaneously. The level of site congestion at a particular site can be accurately determined through careful examination of switch reports which is routinely done by traffic planning engineering teams.

In this section, we examine some possible methods that can be followed to alleviate traffic congestion. In general the problem can be approached in two ways. The first approach is to add new resources. This can be done either through the upgrade of existing cell sites or through the addition of new cells. The second approach involves traffic balancing between the cells and it is accomplished either through RF or system parameter optimization. Both approaches play significant parts in everyday practice and here we present the methodology and some of the results that can be achieved. A brief survey of possible traffic congestion control methods is shown in Fig. 6.1

![Figure 6.1. Some methods for traffic-congestion control](image-url)
6.1 Congestion Control Through Addition of Resources

The addition of the traffic resources is a relatively straightforward method for addressing traffic congestion. To add new resources, we can either purchase new radios and upgrade the existing site, sectorize the existing sites, or add additional sites.

6.1.1 Addition of new radios/trunks

Integration of new radios provides additional traffic capacity to a given cell site. Ideally, a decision on the number of radios to be added should be made on the basis of a traffic trending process. That way, the system capacity is increased in timely manner and no degradation of the GOS is experienced. However, in practice, traffic engineers are sometimes faced with a site that is already congested and an estimate needs to be made on the number of required resources that will bring the GOS within the desired target value (usually 2%). The estimation process can be illustrated with the following example.

Example 6.1. Consider an IS-136 site serving 4.65 E of traffic using 3 radios (8 trunks for voice and one for control channel). The site is experiencing a GOS of 7%. Using the Erlang B formula, estimate the number of radios required to bring the GOS at the site to 2%.

The site is serving 4.65 E at a GOS of 7%. Using Table 3.1, we can easily determine the offered traffic at the site

\[
\frac{a_{offered}}{1 - \frac{GOS}{100}} = \frac{4.65}{1 - \frac{7}{100}} = 5 \text{ E}
\]

Using the Erlang B formula (or table in Appendix A), we determine that in order to support 5 E of offered traffic at a GOS of 2%, the cell site needs 9.9 → 10 trunks. The addition of an extra radio to the site would provide a total of 11 trunks and brings the GOS to 0.83 % which is well within the required target.

The addition of new radios can cause some practical implementation problems. The two most important are

- Ability of existing base station hardware to support additional radios
- Frequency plan for the sites in the area may require some changes

6.1.2 Sectorization

When growth has reached a point where an engineer can no longer simply add channels to a site due to the limitations of the frequency plan or to hardware issues, the next step would be to change from an omni configuration to a sectored configuration. If we change to a sectored configuration without changing the amount of traffic resources available at the site, we will actually decrease the amount of capacity the site can support (refer to Example 6.2). Therefore,
we must increase the total amount of traffic channels at a site to support the same amount of traffic we had with an omni configuration.

**Example 6.2.** Using the Erlang B table in Attachment A, determine

a) the amount of traffic supported by an omni directional site with 36 voice channels at a GOS of 2% and
b) the amount of traffic supported by a three sector site with 12 voice channels per sector (36 channels total) and a GOS of 2%.

**Part a):** Using the Erlang B table in Attachment A with N = 36 and a GOS of 2%, find that A = 27.3 erlangs.

**Part b):** Using the Erlang B table in Attachment A with N = 12 and a GOS of 2%, find that A = 6.61 erlangs. The three sectors combined will support a traffic capacity of 3 x 6.61 erlangs = 19.83 erlangs.

By changing to a sectored configuration and keeping the same amount of voice channels at this site we have reduced the capacity that this site can support by 7.47 erlangs. This relates to a reduction in trunking efficiency of 27.4%.

Once we have gone from the omni configuration to a sectored configuration we can continue to add channels to each of the sectors until once again the site reaches limitations associated with equipment and the frequency reuse plan. At this point other options need to be considered.

### 6.1.3 Cell Splitting and Microcell Deployment

The next step in relieving traffic congestion is to perform a cell split. This process provides additional traffic resources to the same geographical area while still maintaining a sound frequency reuse plan. Example 6.2 demonstrates a simple approach to determine the amount of off-loading that side splitting could provide.

**Example 6.2.** The traffic usage of faces B and C in the diagram in Fig. 6.2 have started to exhibit blocking that is greater than the design objective of the system. System parameters have already been modified to relieve the congestion and each of the sectors already has the maximum amount of traffic channels possible. The traffic engineers of this system have proposed a side split (4:1). If the growth rate of sectors B and C is known to be $G_B$ and $G_C$, respectively, and the usage of sectors B and C is known to be $x_B$ and $x_C$, respectively, determine the growth rates and projected usage of sectors S1, S2, and S3 of the new cell. Assume a uniform traffic distribution over faces B and C.
The area covered by faces B and C can be divided into eight triangles each with equal area. The area covered by sector S1 of the new cell can be represented by 1/8 times the area of sector B plus 1/8 times the area of sector C. The area covered by sector S2 of the new cell can be represented as 2/8 times the area of sector B. The area covered by sector S3 of the new cell can be represented as 2/8 times the area of sector C. Based on these analogies, we can determine the growth rates and usage of sectors S1, S2, and S3 of the new cell as well as the new usage rates of the original faces B and C (refer to Table 6.1).

### Table 6.1. Projected usage and growth rates as a result of side splitting

<table>
<thead>
<tr>
<th>Sector</th>
<th>Area Covered</th>
<th>Projected Usage</th>
<th>Growth Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>$A_B$</td>
<td>$(5/8)(x_B)$</td>
<td>$G_B$</td>
</tr>
<tr>
<td>C</td>
<td>$A_C$</td>
<td>$(5/8)(x_C)$</td>
<td>$G_C$</td>
</tr>
<tr>
<td>S1</td>
<td>$A_1 = (1/8)(A_B) + (1/8)(A_C)$</td>
<td>$x_1 = (1/8)(x_B) + (1/8)(x_C)$</td>
<td>$G_1 = (G_B + G_C)/2$</td>
</tr>
<tr>
<td>S2</td>
<td>$A_2 = (2/8)(A_B)$</td>
<td>$x_2 = (2/8)(x_B)$</td>
<td>$G_2 = G_B$</td>
</tr>
<tr>
<td>S3</td>
<td>$A_3 = (2/8)(A_C)$</td>
<td>$x_3 = (2/8)(x_C)$</td>
<td>$G_3 = G_C$</td>
</tr>
</tbody>
</table>

In practice, cell site arrangement rarely follows the hexagonal grid pattern. In addition, distribution of the traffic is non-uniform. For that reason, network planning tools are frequently used to estimate the effect of the new site integration. As a minimum, an evaluation of a cell split should take the estimate of the traffic offloading as well as the changes to the RF and system parameters that need to be implemented on the surrounding sites.

The microcell implementation can be seen as a special form of cell splitting. A new cell site with a small coverage area is placed in a traffic "hot-spot" area to capture a good portion of traffic causing congestion on the macrocell level. Microcell antennas are usually mounted below the rooftops and their target coverage area rarely exceeds one mile in radius. Deployment of microcells provides a substantial increase in capacity due to its ability to super-reuse of frequencies. This concept can be illustrated with the aid of Fig. 6.3.
The coverage overlap between the macrocell sites prevents tight reuse of frequencies. Therefore, depending on the technology, in regular FDMA mode, macrocells are planned using a frequency reuse strategy of $N = 9$, $N = 7$ or $N = 4^9$. Since microcell coverage areas are disjointed, their signals do not interfere. As a result, they can be planned with a much tighter reuse pattern. In some instances, if a high capacity needs to be provided and the microcells are sufficiently separated, the same set of frequencies can be reused for all microcell sites in the area. This approach is sometimes referred to as the super-reuse.

### 6.2 Congestion Control Through Cell Site Coverage Balancing

From the traffic planning standpoint, the most optimal situation arises when the traffic load is uniformly distributed between the cell sites. However, in reality, spatial density of mobile users exhibits a non-uniform character. On a global scale, some of that non-uniformity is taken into account during the system design phase. As a result, the sizes of the cell site coverage areas depend on the estimate of traffic density. For example, cell sites in urban and dense urban areas have at least a three to four times smaller coverage footprint than all sites located in suburban and rural area.

Despite planning on the global level some localized areas of the system may experience highly uneven traffic demand between adjacent cells. In such circumstances balancing the coverage and capacity between the cells may be the most efficient way to control traffic congestion.

---

9 In GSM networks with frequency hopping the reuse can be increased to $N=3$ or even $N=1$. 
6.2.1 Optimization of System Parameters

Changing the system parameters can be the simplest way to provide traffic balancing. A simple change to the neighbor list (a list of allowable handoff candidates) could provide temporary off-loading of some of the traffic congestion. In addition to the neighbor list, one could fine-tune other system parameters such as the handoff threshold or the handoff hysteresis values between the site that is experiencing the traffic congestion and its neighboring sites. These parameters can be adjusted so that handoffs will occur earlier as a mobile travels away from the cell that is congested, or later as the mobile moves towards the congested cell. Caution should be taken when making these types of changes as even though they could provide a solution to the traffic congestion, they could decrease the call quality as well. For example, changing the handoff hysteresis parameters may result in premature handoffs and degradation in the voice quality as well as the drop call statistics. Therefore, changing system parameters should be accompanied with appropriate RF optimization.

6.2.2 Overlay and Underlay Configurations

An overlaid/underlaid subcells feature is used in cellular networks to increase system capacity without building the new sites or adding more frequencies. This method utilizes the second frequency reuse pattern with the shorter reuse distance. The cells using the new reuse pattern should be restricted in size to make a shorter reuse distance possible without causing excessive co- or adjacent channel interference. These additional cells are called overlaid subcells and the original cells are called underlaid subcells (Fig. 6.4.).

![Figure 6.4. Overlaid and underlaid subcells](image-url)

The available frequencies are divided between the overlaid and underlaid subcells. Each overlaid subcell covers a smaller area than its corresponding underlaid subcell, and therefore, the layer of
overlaid subcells can reuse frequencies more often. Consequently, the number of the channels available per cell can be increased providing increased system capacity.

To illustrate the improvements resulting from underlay/overlay configuration, consider the following example.

**Example 6.3.** Assume that the total number of channels in an imaginary cellular network equals 126. Using just one layer of cells with a 7/21-reuse pattern, each sector could be assigned a maximum of six channels. With underlaid/overlaid cell structure, available channels can be divided in the following way: 42 channels remain in an underlaid frequency plan and 84 channels are assigned to the overlaid cells. Each underlaid cell now supports up to two channels using 7/21 reuse, while overlaid cells with a tighter reuse of 4/12 can support up to seven channels. Therefore, the total number of channels allocated in each sector can be increased from 6 to 9. At a 2% GOS, this results in a capacity improvement from 2.28 to 4.34, which is a 90% capacity improvement.

### 6.2.3 Hierarchical cell structure

The HCS feature divides the cellular system into several layers (typically two or three as shown in Fig. 6.5) and implements dedicated logic that allocates the channels to the mobiles according to the built-in priorities between layers. Changing the priority between the cells can be used to steer the traffic away from the congested sites and towards the sites having a surplus of traffic resources.

![Layered cell structure](image)

**Figure 6.5.** Layered cell structure

The gains that can be achieved using the HCS are similar to the overlay/underlay. Additional flexibility derived from the fact that various HCS levels do not need to be co-located allows much more flexible design and it can provide substantial capacity improvements. The hierarchical cell structure is not supported in all cellular standards. The technologies that provide HCS support are: IS-136, GSM and UMTS-FDD.
7 References


