1 Introduction

Telecommunication traffic, or simply traffic, can be defined as the flow of information messages though 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>
<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 re-allocation 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 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 attempt 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](image)

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.

![Graphs showing number of call attempts](image)

**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.

![Histogram of call holding time measurements](image)

**Figure 2.3.** Histogram of the call holding time measurements

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

$$A = \frac{t}{T} \quad \text{[E]}$$  \hspace{1cm} (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.

![Diagram of resource occupancy](image)

**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

\[
A = \frac{t_1}{T} + 2 \frac{t_2}{T} + \cdots + C \frac{t_C}{T} = \sum_{n=1}^{C} n \frac{t_n}{T}
\]  

(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

\[
A_{offered} = \frac{\lambda T_s}{T}
\]  

(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

\[
A_{offered} = \frac{\lambda T_s}{T} = \frac{100 \times 90}{3600} = 2.5E
\]  

(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

\[
A_{served} = \sum_{n=1}^{C} n \frac{t_n}{T}
\]  

(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...