# Introduction to

# Basic Teletraffic Theory

Mr. H. Leijon, ITU

**UNION INTERNATIONALE DES TELECOMMUNICATIONS**
**INTERNATIONAL TELECOMMUNICATION UNION**
**UNION INTERNACIONAL DE TELECOMUNICACIONES**

# Basic Teletraffic Theory (T)

## INTRODUCTION

**Contents**

1. Background
2. Scope and nature of Teletraffic Theory
   - Input process
   - Service mechanism
   - Queue discipline
   - Conservation of flow
3. Mathematical modeling

## 1. Background

The development of telephone traffic theory started at the beginning of this century. The pioneering achievement in this field was that of the Dane A.K. Erlang, whose works were published between 1909 and 1928. Among those who built upon Erlang's ideas should also be mentioned the Swede, Conny Palm, whose papers during the period 1936-1946 (1957) contributed to give traffic theory its present stringency. Many other persons of various nationalities also have contributed to the development of the present theory.

The traffic theory that can be applied to practical cases is based on the assumption of statistical equilibrium, which implies that it can only deal with cases subject to stationary conditions.

For non-stationary conditions, no practical methods of calculations have yet been devised. The theoretical background for dealing with such cases was, however, presented in Palm's doctoral thesis of 1943, in which he made a study of variations in the call intensity. But it is already now quite possible to deal with non-stationary traffic cases with the aid of computer simulations. The theories considered here, however, will be confined to stationary conditions.

Existing theories use different combinations of assumptions, and the derivations from 1909 to date build upon different levels of knowledge and use partly different terminology. A direct review of the derivations for different cases originally presented would therefore not provide a clear survey of the ability of the theory to describe different cases occurring in practice. It has therefore been preferred to present the traffic theory in a more general form, from which various special cases can then be derived.

In the first section, therefore, the characteristics which are common to different methods of grouping, both for loss and delay systems, will be derived. It will then be preferable to deal separately with the full availability group in a loss system and with the full availability group in a delay system. The theories for gradings and link systems are dealt with in later chapters, in which it will easily be understood how the general theory is applied to these cases.

## 2. Scope and nature of Teletraffic Theory

Teletraffic Theory can be considered as queuing theory applied for telecommunication systems. The general concept queuing theory is concerned with the mathematical analysis of systems subject to demands whose occurrences and lengths can, in general, be specified only probabilistically. For example, consider a telephone system whose function is to provide communication paths between pairs of telephone sets (customers) on demand. The provision of a permanent communication path between each pair of telephone sets would be astronomically expensive and perhaps impossible. In response to this problem, the facilities needed to establish and maintain a talking path between a pair of telephone sets are provided in a common pool, to be used by a call when required and returned to the pool when no longer needed. This introduces the possibility that the system will be unable to set up a call on demand because of a lack of available equipment at that time. Thus, the question immediately arises: how much equipment must be provided, so that the proportion of calls experiencing delays will be below a specified acceptable level? Questions similar to that just posed arise in the design of many systems quite different in detail from a telephone system: how many beds should a hospital provide? How many data terminals can a time-shared computer service? These questions share a common characteristic: in each case, the times at which requests for service will occur and the lengths of times that these requests will occupy facilities cannot be predicted except in a statistical sense. Although these systems are usually very complex, it is often possible to abstract from the system description a mathematical model whose analysis yields useful information.

Consider the following model. Customers request the use of a particular type of equipment (server). If a server is available, the arriving customer will seize and hold it for some length of time, after which the server will be made immediately available to other incoming or waiting customers. If the incoming customer finds no available server, he then takes some specified action such as waiting or going away. Hence the model is defined in terms of three characteristics: the input process, the service mechanism and the queue discipline.

The input process describes the sequence of requests for service. Often, for example, the input process is specified in terms of the distribution of the lengths of time between consecutive customer arrival instants. The service mechanism is the category that includes such characteristics as the number of servers and the length of time that the customers hold the servers. For example, customers might be processed by a single server, each customer holding the server for the same length of time. The queue discipline specifies the disposition of blocked customers (customers who find all servers busy). For example, it might be assumed that blocked customers leave the system immediately or that blocked customers wait for service in a queue and are served from the queue in their arrival order.

Now, consider the following model. Two cities are interconnected by a group of n telephone trunks (servers). Suppose that arrivals finding all trunks busy do not wait but immediately depart from the system. (Technically, no "queuing" occurs.) What proportion of incoming calls (customers) will be unable to find and idle trunk (and thus are lost)?

We wish to derive a formula that will predict the proportion of calls lost as a function of the demand; that is, we wish to derive a formula that allows estimation of the number of trunks required to meet a prespecified service criterion from an estimate of the telephone traffic load generated between the two cities. The great practical value of any model that leads to such a formula is obvious.

We shall now give a heuristic derivation of the required formula, using a concept of great importance in science and engineering, that of conservation of flow. The following derivation is heuristic, so no one is expected to understand it completely; the "derivation" is a plausibility argument and is correct in certain circumstances.

When the number of customers in the system is *j*, the system is said to be in state $E_j$ (j = 0, 1, .... n). Let $P_j$ be the proportion of time that j trunks are busy; $P_j$ is the proportion of time the system spends in state $E_j$. Denote by $\lambda$ the call arrival rate; $\lambda$ is the average number of requests for service per unit time. Consider first the case j < n. Since calls arrive with overall rate $\lambda$, and since the proportion of time the system spends in state $E_j$ is $P_j$, the rate at which the transition $E_j \rightarrow E_{j+1}$ occurs (the average number of such transitions per unit time) is therefore $\lambda P_j$. Now, consider the case when j = n. Since the state $E_{n+1}$ represents a physically impossible state (there are only n trunks), the transition $E_n \rightarrow E_{n+1}$ is zero. Thus the rate at which the upward transition $E_j \rightarrow E_{j+1}$ occurs is $\lambda P_j$ when j = 0, 1, ..... n - 1 and is Zero when j = n.

Let us now consider the downward transitions.

$E_{j+1} \rightarrow E_j$ (j = 0, 1, ....., n - 1)

Suppose that the mean holding time (the average length of time a call holds a trunk) it $\tau$, then if a single trunk is busy, the average number of calls terminating during an elapsed time $\tau$ is l; the termination rate for a single call is therefore $l/\tau$. Similarly, if two calls are in progress simultaneously and the average duration of a call is $\tau$, the average number of calls terminating during an elapsed time $\tau$ is 2; the termination rate for two simultaneous calls is therefore $2/\tau$. By this reasoning, then, the termination rate for j + 1 simultaneous calls is $(j + 1)/\tau$. Since the system is in state $E_{j+1}$ a proportion of time $P_{j+1}$, we conclude that the downward transition $E_{j+1} \rightarrow E_j$ occurs at rate $\frac{j+1}{\tau} \cdot P_{j+1}$ transitions per unit time (j = 0, 1, .... n - 1).

We now apply the principle of conservation of flow. If the system is to be in statistical equilibrium, that is if the relative proportion of time that the system spends in each state is to be a stable quantity, then the upward transition $Ej \rightarrow Ej_{+1}$ must occur with the same rate as the downward transition $E_{j+1} \rightarrow E_j$. Thus, we have the so called statistical equilibrium balance equations.

$$\lambda \cdot P_j = (j+1) \ . \ \tau^{-1} \ . \ P_{j+1} \qquad \text{(j = 0, 1, j....., n - 1)} \qquad \text{(TIN 2.1)}$$

These equations can be solved recurrently; the result, which expresses each $P_j$ in terms of the value $P_o$ is:

$$P_j = \frac{(\lambda \cdot \tau)^j}{j!} \cdot P_o \qquad \text{(j = 0, 1, ....., n)} \qquad \text{(TIN 2.2)}$$

Since the numbers $\{P_j\}$ are proportions, they must sum to unity:

$$P_o + P_1 + \ \ldots\ldots \ P_n = 1 \qquad \text{(TIN2.3)}$$

Using the normalization equation (TIN 2.3) together with equation (TIN 2.2), we can determine $P_o$:

$$P_o = (\sum_{k=0}^{n} \frac{(\lambda \cdot \tau)^k}{k!})^{-1} \qquad \text{(TIN 2.4)}$$

Thus we obtain for the proportion $P_j$ of time that j trunks are busy, the formula:

$$P_j = \frac{(\lambda.\tau)^j / j!}{\sum_{k=0}^{n} (\lambda.\tau)^k / k!} \qquad \text{(TIN 2.5)}$$

An important observation to be made from formula (TIN 2.5) is that the proportions $\{P_j\}$ depend on the arrival rate $\lambda$ and the mean holding time $\tau$ only through the product $\lambda.\tau$. This product is a measure of the demand made on the system; it is often called the offered load and given the symbol $A = \lambda \cdot \tau$. The numerical values of A are expressed in units called erlangs (erl), after the Danish mathematician A.K. Erlang, who first published the formula (TIN 2.5) in 1917. When j = n in formula (TIN 2.5), the right-hand side becomes the well-known Erlang loss formula, denoted in Europe by $E_{1n}(A)$:

$$E_{1n}(A) = \frac{A^n / n!}{\sum_{k=0}^{n} A^k / k!} \qquad \text{(TIN 2.6)}$$

The point to be made here is that some potentially useful mathematical results have been derived using only heuristic reasoning. The question we must now answer is: What are the conditions under which these results are valid?

More precisely, what assumptions about the input process and service mechanism are required for the validity of formulas (TIN 2.5) and (TIN 2.6)? Can the assertion that the downward transition rate is proportional to the reciprocal of the mean holding time be justified? What is the relationship between the proportion $P_j$ of time that j calls are in progress and the proportion $\Pi_j$, say, of arriving calls that find j other calls in progress? How widely applicable is the conservation-of-flow analysis? How does one handle processes for which this type of analysis is inapplicable?

Questions of this nature sometimes require highly sophisticated mathematical arguments. We shall however take a middle ground with regard to the use of advanced mathematics. The material should be accessible to a student who understands applied probability theory and related areas of mathematics.

## 3. Mathematical modeling

From the foregoing, we understand that traffic theory consists of the mathematical modeling of a telecommunications system (or some part of it) and its behaviour when demands are made on it or by it. All such theory is a modeling exercise: we cannot set up and examine a truly identical system, and so instead we construct a (hypothetical) simplified version, with well-defined inputs, and analyse that. The validity and usefulness of the theory we develop therefore rests entirely upon the answer to the question: how satisfactory is the model? If we have little confidence in that, then no matter how sophisticated our mathematics we can have little confidence in the final results of the theory.

In order to set up our model, we must take careful consideration of a number of points:

1) What part, exactly, of a system are we interested in? Can we separate out the relevant section, and look at this in isolation; or must we treat it all at once?

2) What is the precise technical behaviour of the section we have decided on, in terms of operating times, limitations on access, dead times, detailed response to a demand, etc.?

3) How does the input stream of demands on the system behave?

4) What information do we want from our model, and how accurate must it be?

The points are all mutually dependent. It is perhaps easier to start with the incoming demand stream, and enquire how this is structured: which will itself involve some assumptions that the sources of demand - e.g. subscribers -

are reacting to the “usual” system behaviour in their “usual” way. Typically, this stream of demands is governed very largely by chance, so that the methods of Stochastic processes will be appropriate. We then need to know:

a) what (in probabilistic language) is the *arrival process* of demands? And

b) what is the distribution of the work that they bring?

The description of the arrival process may require more or less detail. In a low-congestion system offered fresh traffic by a multitude of independent subscribers, we may assume with high accuracy that (at any rate over not-too-long periods of time) the arrivals are *pure-chance* - i.e. that they form a Poisson process. If we are examining a system offered overflow traffic, a more complex description will be needed; and more complex still if we expect a significant proportion of repeat-attempts.

Now consider the second point - the distribution of the work brought by a single demand. Indeed, this may have considerable qualitative variation: if we are concerned with circuit occupancy, the “work” consists of a single continuous holding-time; whereas if we are modeling the common control of a complex processor system, it may be a sequence of disjoint tasks, of widely different types and duration. Two particularly common and important cases are however where the holding-time has a negative-exponential distribution, and where it is effectively deterministic - i.e. a constant.

We can now turn our attention to point (4) above - what information do we want from the model. This naturally requires at any rate some knowledge of the system behaviour (are blocked calls lost, or do they queue, for instance), and an understanding of the nature of the input process. Since that input is stochastic, the output from our model will be probabilistic, and may consist of probabilities of loss, mean delays, percentiles of processor occupancy or similar quantities. It might even be that this is only the first stage in the construction of a more complex model, in which case we may need to know complete details of the distributions of delays, of overflowing calls, or of some other quantity which affects the rest of the system. At this stage, point (1) is considered.

We are now ready to attend to the details of the system engineering, point (2) above; and it is at this stage that our mathematical model takes shape, and it becomes clear whether we have any hope of an analytic treatment. Finally, the whole process is repeated, until we have confidence that it is consistent and that the behaviour of the system is indeed compatible with the assumptions on the incoming demand stream and vice versa.

We assume then that a model has been, or can be, set up. It will necessarily be approximate in some way or other, and we must estimate how much effect this approximation will have on the results of model. If the answer is too much effect, we must reshape the model. The ultimate requirement is always for a number or set of numbers, that is, for a numerical calculation: so useful tradeoffs can often be made between approximate models and approximate calculations.

It may in fact not be possible to specify the input data of subscriber behaviour in as much detail as it needed, because the relevant quantities are unknown or even unmeasurable. In such circumstances, the approaches are reasonable: to decide what The Reasonable Man should do, and make a decision that that is what the system shall be dimensioned to (with, of course, appropriate safeguards for the system itself!); or to analyse several models, differing only in the subscriber behaviour, and present a range of results for final decision-taking on other grounds.

Fortunately, most quantities of interest as output from the mathematical models are remarkably robust with respect to variations in the input processes (provided, of course, that certain critical parameters like the overall traffic offered are kept constant), and so even quite simple models give remarkably accurate and useful results. It should never be forgotten however, that no matter how elaborate the mathematical solution,

A BAD MODEL MEANS UNRELIABLE RESULTS