15. TELERAFFIC THEORY

15.1. Introduction

The telecommunication traffic theory is the application of the mass-servicing theory for telecommunication systems. The theory of the telecommunication traffic was founded by the Danish A. K. Erlang and published between 1909-1928. The traffic theory applied for practical cases is based upon the condition of the statistical equilibrium.

Telecommunication systems are built upon sets of different resources (switching units, transmission channels, etc.). Due to economical reasons, the number of such resources is limited in a system thus the customers have to share them somehow. This may happen so that a servicing unit is engaged during a call and it is disengaged, i.e. 'given back' to the common resource at the end of the call. This method implies the chance that it is impossible to set-up a call when all servicing devices of the system are engaged simultaneously. Such unpleasant event is called *congestion*, and from this point of view, services are qualified by the so-called GOS (grade of service). The aim of the traffic design is to provide for a sufficient number of servicing devices and to maximize their utilization.

15.2. Terms and Definitions

To make the following discussion easier, some terms and definitions are mentioned in advance:

- *call intensity* (λ)**:** the sum of the demands directed to a switching unit, group of circuits, or customers per unit of time.
- *holding time* (*h*): the time for which the servicing device is engaged by an acknowledged demand.
- *traffic intensity*: the sum of the holding times of calls simultaneously in progress during a particular period of time. Let us take first the sum of the holding times:

$$\sum_{i=1}^{z} h_i = z \cdot \bar{h} \qquad (15.1)$$

where h_i is time of the i-th call, z is the number of the calls and $\bar{h}$ is the average holding time. If the period of time during which the holding times were counted is T then the average of the occupations is:

$$\frac{1}{T}\sum_{i=1}^{z} h_i = \frac{1}{T} \cdot z \cdot \bar{h} = l \cdot \bar{h} \qquad (15.2)$$

where λ is the call intensity and

$$\lambda \cdot \bar{h} = A \qquad (15.3)$$

gives the average of the simultaneous occupations for a given period of time, generally called traffic intensity or simply traffic. Traffic is a quantity with no dimension, the word Erlang is, however, used with the value to indicate that the number is a term used in *telecommunication*.

- *offered traffic (A)*: amount of traffic offered for a group of devices corresponding to the theoretical description of the given traffic situation (a presumed quantity).
- *carried traffic (Y)*: traffic carried (transmitted) by a certain group. It can be used both for theoretical description and -since it can be measured- for actual situation as well.
- *busy hour*: a daily period of one hour in which the traffic is the greatest. The time of the busy hour generally depends on the calendar days.
- *time-consistent busy hour*: period of one hour starting at the same time each day in which the traffic intensity is the maximum for the group of devices and the days examined. In all likelihood, the customers would like to have a satisfying GOS even in the busiest hours of the year. Value of the traffic during this period is, however, practically impossible to measure, further system with satisfying GOS even fur such a high traffic would be too expensive and would not be well utilized in the other hours. Therefore it is more reasonable to choose in the design a smaller traffic value which is exceeded only in a few days of the year.

15.2.1. Mathematical Model of Telephone Traffic

The *input process*, the *service procedure* and the *servicing rules* are the characteristic features of a model. The input process is defined by the distribution of time passing between the arrivals of two consecutive call demands. The service procedure is determined by the number of the service units, by the distribution of the service (holding) times and by the access mode to the service units. The service rules dispose of the congested demands. In *loss* systems the congestion is resolved by clearing the congested demands while in *delay* systems the calls form a queue and are serviced e.g. in the order of arrivals.

Two terms of the congestion are used: *Time congestion* is the proportion of the time during which all accessible service units are simultaneously engaged. *Call congestion* is the proportion of those calls which were rejected in a loss system or had to wait in a delay system.

The different cases are based upon the following conditions:

1.) The principle of the statistical equilibrium may be used.
2.) The operation of individual traffic sources is independent of the state of the other sources.
3.) Time between two consecutive calls has a negative exponential distribution.
4.) Time of the individual occupations is independent of other occupations.
5.) Duration of the individual occupations has a negative exponential distribution.
6.) The fate of the unsuccessful calls is regulated by deterministic rules.

The traffic is considered as events generated by individual sources each capable to initiate simultaneously only one call. The number of the sources can be finite or infinite, the traffic offered by them, however, is always finite!

The distribution of arrival times has a negative exponent and their average is $1/\lambda$:

$$F(t) = 1 - e^{-lt} = P(\leq t) \qquad (15.4)$$

The distribution of times between an arbitrary point of time and a call follows the same exponential function as the distribution of the times between the calls and it is independent of t (i.e. memoryless). The number of arriving calls is described by Poisson distribution.

The holding time distribution is given by the holding time average h in the negative exponent:

$$F(t) = 1 - e^{-\frac{i}{\bar{h}}} = P(\leq t) \qquad (15.5)$$

If the distribution of the holding times is exponential and provided the number of independent occupations is *i*, then the number of ending calls within time *t* is

$$m_t(t) = i \cdot t / \bar{h} \qquad (15.6)$$

The number of the call occurrences can be calculated from the call intensity of the free traffic sources (λ'):

$$\lambda_i = (S\text{-}i)\cdot\lambda' \quad \text{(Bernoulli, Engset)} \qquad (15.7)$$
$$\lambda = \lambda_i \quad \text{(Poisson,Erlang)} \qquad (15.8)$$

where *i* is the number of engaged traffic sources, *S* is the number of traffic sources and λ_i is the frequency of call intensity (occurrence).

An important element of the service mechanism is the mode of accessing (grouping) the service units which can be as follows:

1.) *Full-availability group* in which any input has access to any output; a free output can thus always be accessed by a given input regardless of the occupations between other inputs and outputs. As shown in Fig. 15.1. a.), inputs and outputs are interconnected through the switching elements located at the cross-points of vertical and horizontal lines. Because of this arrangement, the array of switches is called the switch matrix.

To analyze the traffic behaviour, the model shown in Fig 15.1.c) is used. Here the small circles arranged in a straight line represent the corresponding inputs and outputs. The presentation most frequently used for the full-availability switch matrix is shown in Fig. 15.1. b).

2.) *Limited-availability* (or grading) group was used in space division exchanges to increase the throughput of the circuits.

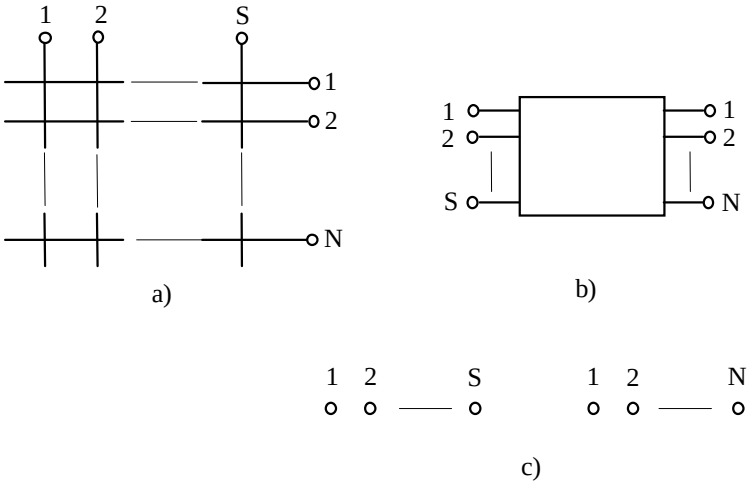

Fig 15.1. Different Presentation of Full-Availability Groups

3.) *Link system* in which the input/output interconnections are realized by two or more serially connected full-availability switch matrices with a small number of cross-points (see Fig. 15.2.). The link system minimizes the number of cross-points, its application began in cross-bar systems using precious metal contacts.

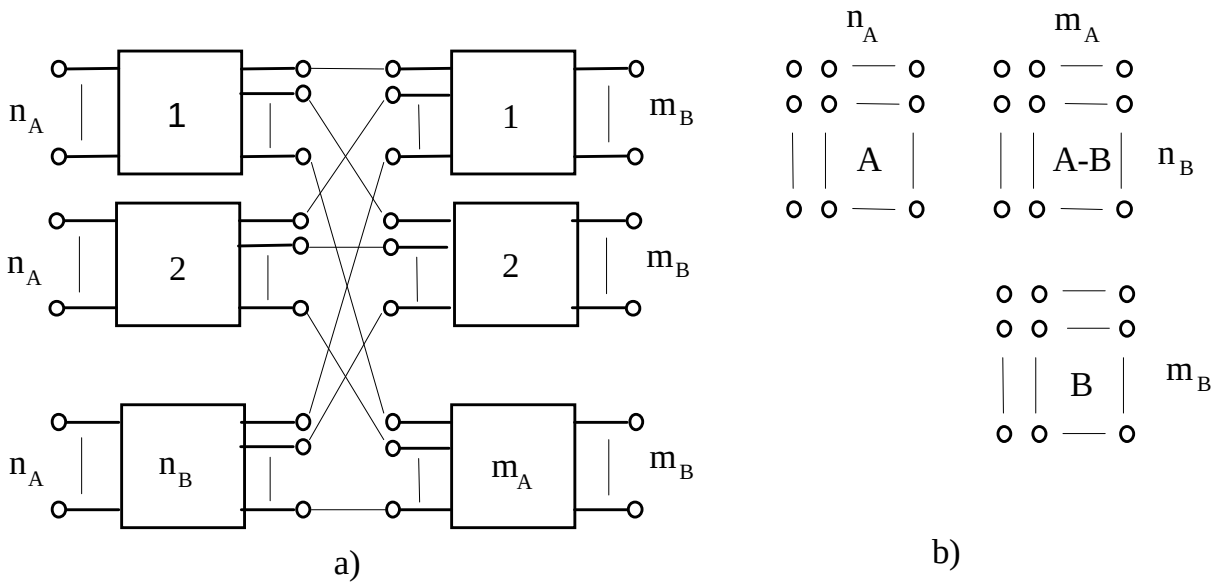

Fig 15.2. Different Presentation of Two-Stage Link Systems

15.2. Loss Systems

The principle of statistical equilibrium can be applied for the busy hour since during this time the traffic is neither increasing nor decreasing but it is fluctuating near the average value. This is only possible if the frequency of transitions from the (*i*-1) occupations to *i* occupations is the same as inversely.

15.3.1. Full-Availability Group

Let *P*(*i*) be the probability of *i* simultaneous occupations out of *N* service units (*i* = 0, 1, ... *N*). P(*i*) represents also the proportion of the time during which *i* simultaneous occupations exist. Let λ_i denote the call intensity, m_i the call terminations per unit of time in a system with *i* occupations.

Using the principle of the statistical equilibrium:

$$P(i+1)\cdot m_{i+1} = P(i)\cdot l_i \quad \text{from which} \quad P(i+1) = P(i)\cdot\frac{\lambda_i}{\mu_{i+1}} \qquad (15.9)$$

(Note that $l_N = 0$ in loss systems!)

a.) *Erlang-type system*

If the call intensity is constant (i.e. *S*>>*N*) then $l_i = l$. Substituting $m_{i+1} = h/(i+1)$ and $l \cdot h = A$ into equation (15.9):

$$P(i+1) = P(i)\,\frac{A}{i+1} \qquad (15.10)$$

With recursion from *i* = 0 and the full series of events:

$$P(i) = \frac{\frac{A^i}{i!}}{\sum_{i=0}^{i=N} \frac{A^i}{i!}} \tag{15.11}$$

Time congestion is defined as:

$$E = \sum_{i \geq N} P(i) \tag{15.12}$$

so that the time congestion of Erlang-type systems is

$$E_N(A) = \frac{\frac{A^N}{N!}}{1 + A + \frac{A^2}{2!} + \ldots + \frac{A^N}{N!}} \tag{15.13}$$

The call congestion (*B*) is defined as

$$B = \frac{\sum_{i \geq N} l_i P(i)}{\sum_{i \geq 0} l_i P(i)} \tag{15.14}$$

Substituting $\lambda_i = \lambda$ we obtain

$$B_N(A) = P_N(A) = E_N(A) \tag{15.15}$$

which is called the first Erlang formula. Tables or diagrams are used for the quick evaluation of the Erlang formula B, as is shown in Fig. 15.3. Usually, loss is given as the parameter but any other variable can be used instead.

The carried traffic:

$$Y = \sum_{i=1}^{N} i \cdot P(i) \tag{15.16}$$

can be expressed also as

$$Y = A \cdot [1 - E_N(A)] \tag{15.17}$$

The average usage in the case of random hunting of the service units is

$$a = \frac{A}{N} \cdot [1 - E_N(A)] \tag{15.18}$$

The usage of the *i*-th service unit in the case of sequential hunting:

$$a_i = A \cdot [E_{i-1}(A) - E_i(A)] \tag{15.19}$$

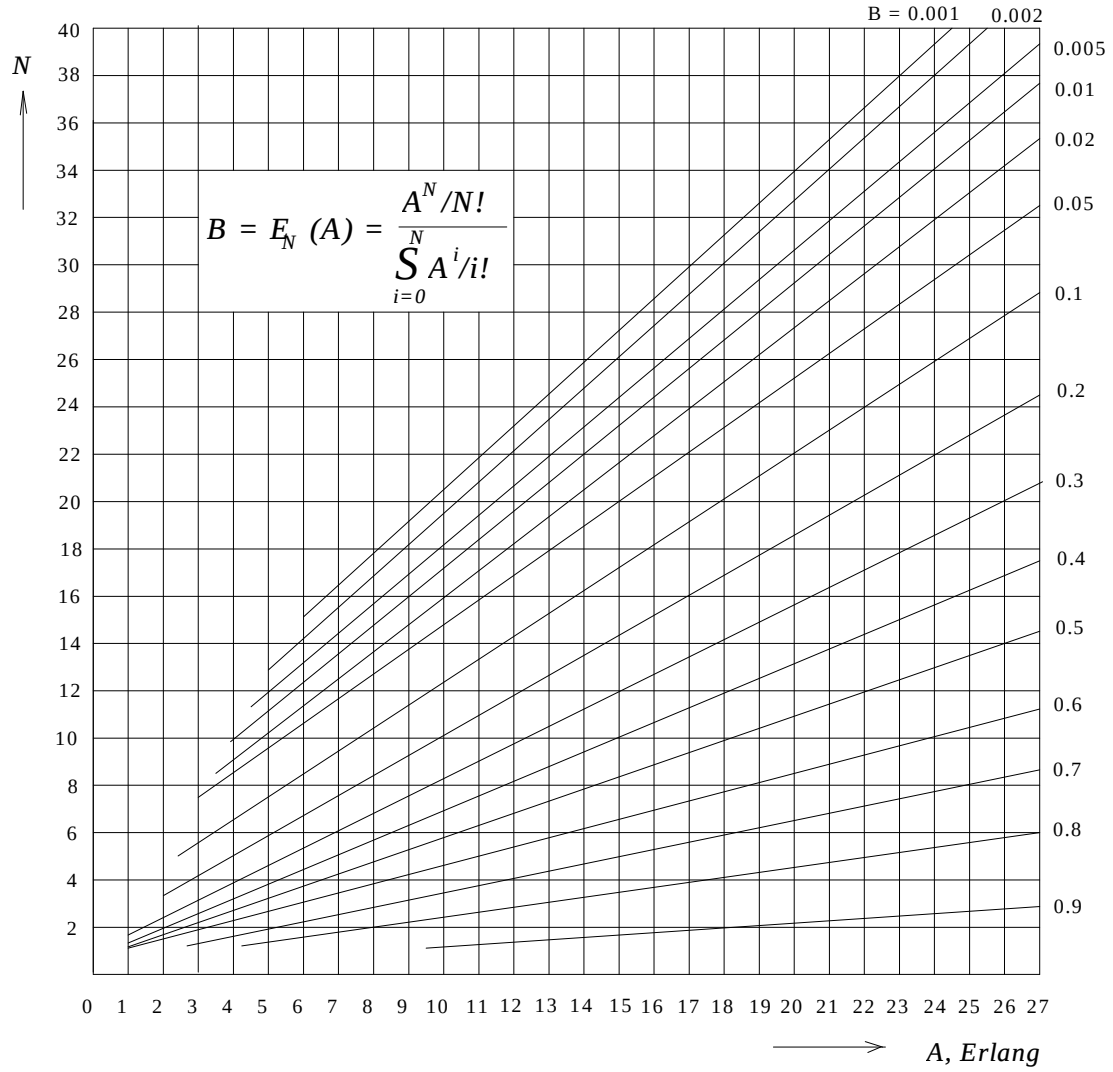

Figure 15.3 Graphical Representation of the Erlang Formula

b.) *Varying call intensity.*

If S is not much greater than N, the call intensity depends on the number of the engaged service units. Substituting λ_i from (15.7) into eq. (15.9) representing the principle of the statistical equilibrium:

$$P(i+1) = P(i)\cdot(S-i)\cdot\frac{\lambda'\bar{h}}{i+1} \qquad (15.20)$$

Using notation $\lambda'\cdot h = \alpha$ (offered traffic of the free traffic-source in a unit of time!) and starting recursion with $i = 0$ we have

$$P(i) = P(0)\cdot\binom{S}{i}\cdot\alpha^i \qquad (15.21)$$

P(0) can be determined from the total event and the well-known binomial distribution is obtained for $S = N$:

$$P(i) = \binom{S}{i}\cdot\frac{a^i}{(1+a)^S} = \binom{S}{i}\cdot a^i(1-a)^{S-i} \qquad (15.22)$$

where a = $\frac{\alpha}{1+\alpha}$ is the traffic offered by the traffic source in the unit of time.

The call congestion (*B*) is zero here and the time congestion is:

$$E = P(N) = a^S \qquad (15.23)$$

15.3.2. Link Systems

The exact calculation of the traffic situations of a link system is very complex. The approximation which can be practically used are based almost exclusively on the theory worked out by Jacobaeus.

Suppose that the state of occupation of the links and that of the outputs is independent, and that every free pair of link and output can be engaged with the same probability (random search) and that the entire congestion is small. The two stage link system is the simplest way to demonstrate the Jacobeous' theory but it can be extended to three or more stages as well.

According to the actual parameters of the input switch matrix (**A** stage), the link system can be basic ($n = m$), expanded ($n < m$) or concentrated ($n > m$). Let *a*, *b* and *c* denote the occupations of an input, a link and an output, respectively. Let $G(p)$ be the probability that p outputs out of m are engaged and $H(m-p)$ the probability that $m-p$ links leading to free outputs are engaged (see Fig. 15.4.). (The positions of $G(p)$ and $H(m - p)$ can, of course, be swapped).

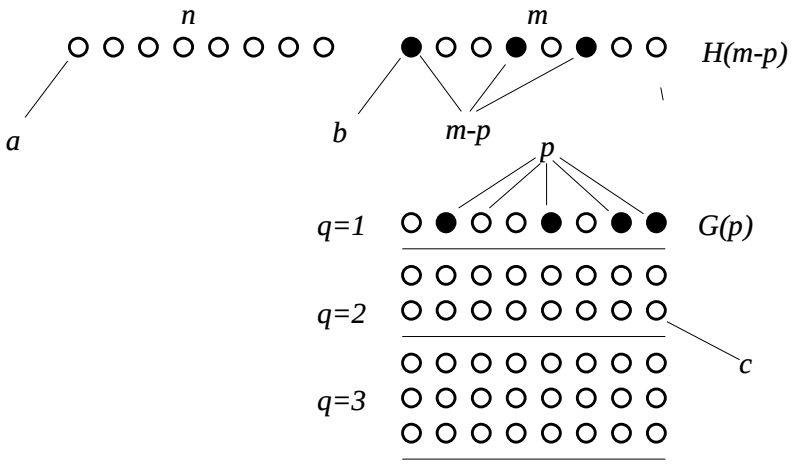

Figure 15.4. Traffic Model of the Two-stage Link System

Obviously, time congestion occurs if the two events are simultaneous. The probability of congestion for $n = m$ is:

$$E = \sum_{p=0}^{p=m} G(p)\cdot H(m-p) \qquad (15.24)$$

The probability of the output occupation can be given either by the Erlang or by the Bernoulli distribution while for the link occupation, the Bernoulli distribution is used. If the Erlang distribution is used for $G(p)$ and the Bernoulli distribution for $H(m-p)$, then

$$E = \sum_{p=0}^{p=m} \frac{\frac{(mc)^p}{p!}}{\sum_{i=0}^{i=m} \frac{(mc)^i}{i!}} b^{m-p} \qquad (15.25)$$

where $m \cdot c$ is the outgoing traffic (m). The expression can be rewritten as

$$E = \frac{E_m(mc)}{E_m\left(m\frac{c}{b}\right)} \qquad (15.26)$$

where the Erlang loss of the m lines in the case of $m \cdot c$ amount of offered traffic is in the numerator while the loss of the same number of lines for the fictive traffic $m \cdot c/b$ is in the denominator.

Generally, the individual directions can be accessed from the matrices B also from more than one output ($q \neq 1$). The congestion is then:

$$E = \frac{E_{mq}(mqc)}{E_{mq}\left(mq\frac{c}{b}\right)} \qquad (15.27)$$

Describing the occupation of the outputs also by the Bernoulli distribution:

$$E = (b + c^q - bc^q)^m \qquad (15.28)$$

15.4. Delay Systems

Besides the probability of waiting, delay systems are characterized by the average waiting time, by the probability of waiting more than a given time interval, and by the expected length of the waiting queue. The exact solution can be given for negative-exponent service time distribution or for constant holding time system with one service unit.

The analysis is given for the so-called Erlang-type system. The previous conditions are thus extended by the further ones:

- each customer who has to wait keeps on waiting for the service,
- the amount of the offered traffic is less than the number of the service units ($A < N$),
- service is done in the order of the arrivals,
- the queue is not limited (infinite number of waiting positions).

The analysis of the delay system is based also on the principle of the statistical equilibrium but the possible states of the system do not end at N but in the case of the simultaneous occupation of N service units, there may be j customers in the queue ($j = 0, 1 \ldots \infty$).

In eq. (15.9) representing the principle of the statistical equilibrium, the rate of increase exists also for $i \geq N$ (λ), the rate of decrease exist always for $i \geq N$ (N/h) because it is assumed that only serviced calls leave the system. Hence

$$P(i+1) = P(i)\cdot\frac{A}{i+1} \qquad \text{if } i \le N\text{-}1 \tag{15.29}$$

$$P(i+1) = P(i)\cdot\frac{A}{N} \qquad \text{if } i > N\text{-}1 \tag{15.30}$$

With the help of the recursion and using the condition of the full event:

$$P(i) = \frac{\dfrac{A^i}{i!}}{\displaystyle\sum_{i=0}^{N-1}\frac{A^i}{i!} + \frac{A^N}{N!}\frac{N}{N-A}} \qquad \text{if } i \le N\text{-}1$$

and

$$P(i) = \frac{\dfrac{A^N}{N!}\left(\dfrac{A}{N}\right)^{i-N}}{\displaystyle\sum_{j=0}^{N-1}\frac{A^j}{j!} + \frac{A^N}{N!}\frac{N}{N-A}} \qquad \text{if } i > N\text{-}1$$

Using equations (15.12) and (15.14), it turns out that the time congestion and the call congestion are of the same value. This is stated by the second (or *D*) Erlang formula, which gives the probability of waiting $P(t > 0)$ as:

$$D_N(A) = \frac{\dfrac{A}{N!}\dfrac{N}{N-A}}{\displaystyle\sum_{j=0}^{N-1}\frac{A^j}{j!} + \frac{A^N}{N!}\frac{N}{N-A}} \tag{15.31}$$

The relation between the Erlang B and D formulae is:

$$D_N(A) = \frac{N\cdot E_N(A)}{N - A\cdot\left[1 - E_N(A)\right]} \tag{15.32}$$

As it follows from the above conditions, $Y = A$, i.e. the carried traffic is equal to the offered traffic. Without derivation, the distribution function of the waiting times is as follows:

$$P(>t) = D_N(A).e^{-\frac{N-A}{\bar{h}}t} \tag{15.33}$$

If a call has to wait then the probability of waiting longer than a given period of time is:

$$P_w(>t) = \frac{P(>t)}{P(>0)} = e^{-\frac{N-A}{\bar{h}}t} \tag{15.34}$$

The expected value of waiting times is given by eq. (15.35) for all calls and by equation (15.36) for the waiting calls:

$$\tau_w = \frac{\bar{h}}{N-A}D(A) \tag{15.35}$$

$$\tau_{wait} = \frac{\bar{h}}{N - A} \qquad (15.36)$$

The expected length of the waiting queue is:

$$(q) = \frac{A}{N - A} D_N(A) \qquad (15.37)$$

For systems which can be characterized by constant holding times (h), an exact solution can be given if N = 1. The values are halves of those obtained for the exponential holding time:

$$\tau_w = \frac{h}{2} \frac{a}{1 - a} \qquad (15.38)$$

$$\tau_{wait} = \frac{h}{2} \frac{a}{1 - a} \qquad (15.39)$$

where a = offered traffic = carried traffic <1.

Control questions

1. What is the aim of the traffic design in telecommunication?
2. What is the definition of the time-congestion and that of the call-congestion?
3. What is the principle of the statistical equilibrium and how is it used?
4. What are the parameters of link systems?
5. What are the characteristic parameters of delay systems?

Examples

1. How many per cent of calls has to pay at least two tariff units if one unit is paid for each commenced 3-minute interval and the average holding time is 2 minutes?
Solution: Using the inverse of the equation (15.5): $P(t > 3) = e^{-3/2} = 22\%$.

2. Suppose a full-availability loss system containing 5 circuits. What will be the values of the carried traffic for the sequential and for the random hunting, provided the offered traffic is 2 Erlangs?
Solution: The traffic of the individual circuits in the case of sequential hunting is given by (15.19) and congestion can be computed from the following recursive relation:

$$E_i(A) = \frac{A \cdot E_{i-1}(A)}{i + A \cdot E_{i-1}(A)}$$

i	$A[E_{i-1}(A) - E_i(A)]$	a_i
1	2·(1-0.66667)	0.66666
2	2· (0.66667-0.4)	0.53334

3	2· (0.4-0.21053)	0.37895
4	2· (0.21053-0.09524)	0.23058
5	2· (0.09524-0.0367)	0.11708
	carried traffic:	1.92662
	loss = 2·0.0367	0.0734
		2.00002

In case of random hunting, the traffic of one servicing unit is (see equation 15.18.)

$$a = \frac{1.92662}{5} = 0.38532$$

3. What should be the minimum number of lines if 20 Erlangs of offered traffic has to be serviced at a congestion not greater than 0.002?
Solution: Reading out from Fig. 15.3.: N ≥ 33.

4. How can the average carried traffic of the circuits of full-availability lines be evaluated, if the number of lines is increasing and the congestion is 0.005?
Solution: Considering eq. (15.18) for the average usage of the circuits, 0.005 can be neglected with respect to 1 so that A/N approximation may be used. From the diagrams of the Fig. 15.3.:

5. How can the average usage of circuits for a full-availability group of 10 circuits be evaluated as a function of the offered traffic?
Solution: Using diagrams of Fig. 15.3. and the equation (15.18):

A, Erlang	$E_{10}(A)$	a, Erlang
3.1	0.001	0.31
4.5	0.01	0.45
7.5	0.1	0.68
12.0	0.3	0.84
18.3	0.5	0.92

It can be shown that for every finite N a $\rightarrow 1$ if A $\rightarrow \infty$.

6. Suppose an Erlang-type full-availability delay system with N = 30 service units, offered traffic of which is 700 calls/hour and the average holding time is 108 s. What is the value of the offered traffic?

Solution: From (15.3): $A = l \cdot \bar{h} = \frac{700}{3600} \cdot 108 = 21\,\text{Erlang}$

What is the probability of waiting?

From the relation (15.32) and from the Fig. 15.3.:

$$D_{30}(21) = \frac{30 \cdot E_{30}(21)}{30 - 21[1 - E_{30}(21)]} = \frac{30 \cdot 0.015}{30 - 21[1 - 0.015]} = 0.048$$

What is the average waiting time of the waiting calls?

From the eq. (15.36): $t_w = \frac{108}{30-21} = 12\,s$

What is the probability for a waiting call that it has to wait longer than 24 s?

From the equation (15.35): $P_w(t > 24s) = e^{-9 \cdot \frac{24}{108}} = 0.1353$

References

[1] R. Syski, Introduction to Congestion Theory in Telephone Systems, Oliver and Boyd, Edinburg and London, 1960

[2] L. Kleinrock: Sorbanállás, kiszolgálás, Mûszaki Könyvkiadó, Budapest, 1979

[3] A távközlési forgalom tervezése. (CCITT 1984) Közlekedési Dokumentációs Vállalat, Budapest, 1986