Fundamental Concepts

From TETRAPRO, edited by Mr. H. Leijon, ITU

UNION INTERNATIONALE DES TELECOMMUNICATIONS
INTERNATIONAL TELECOMMUNICATION UNION
UNION INTERNACIONAL DE TELECOMUNICACIONES

Basic Teletraffic Theory (T)

FUNDAMENTAL CONCEPTS (TFC)

Contents

1. Traffic unit

Call intensity

The telephone traffic is generated by a number of subscribers. When a subscriber marks his intention to make a telephone call by lifting his receiver, the local telephone exchange receives an impulse which starts a number of actions to make it possible to receive numerical information from the caller, so that the telephone exchange later on can connect the calling subscriber to the desired subscriber. When the exchange is ready to receive the desired number, a dialling tone is sent to the caller. The caller can then dial the desired number and the telephone exchange connects the caller to the desired number.

The setting-up procedure involves generally that the call is switched over a number of selectors. These selectors may be all situated in the calling subscriber's own local exchange. However, when the telephone network is large enough, the switching may pass over several switching stages in different exchanges. Consequently, one switching unit handling the call will request another switching unit which may be geographically near or far, to continue the switching procedure. When doing so, information concerning the desired destination must also be transferred. Finally, when the last switching stage is passed, the ringing of the called subscriber indicates that someone wants to speak with him.

The setting-up procedure for a telephone call involves consequently a series of requests for the call to be processed. The total number of such requests to a switching unit - or subscriber - per time unit is the call intensity.

Consequently, the first switching unit in the selection chain receives calls from subscribers, the following switching units receive calls from the previous one, and finally, the called subscribers receive calls from the last link in the switching and other common equipment taking part in the setting-up procedures will also receive calls. One can therefore define the call intensity for each distinct part of the switching chain from calling to called subscriber.

Holding time

Each call that is accepted means that some devices are occupied a certain time. The length of this holding time depends on how fast the switching or processing work can be done and on whether the device is kept during the following switching moments and during the possibly following conversation between the two subscribers or not. The holding times can consequently vary from call to call, depending on the type of switching device used and its function in the setting-up procedure and on the outcome of the call.

Number of simultaneous occupations

If we consider a single switching device, it is easily understood that the occupation of this device depends on how many calls it receives and how long each call will occupy this device.

If we now consider a group of devices where every device can only handle one call at a time, we understand that the number of occupied devices may vary as shown in Figure TFC 1/1.

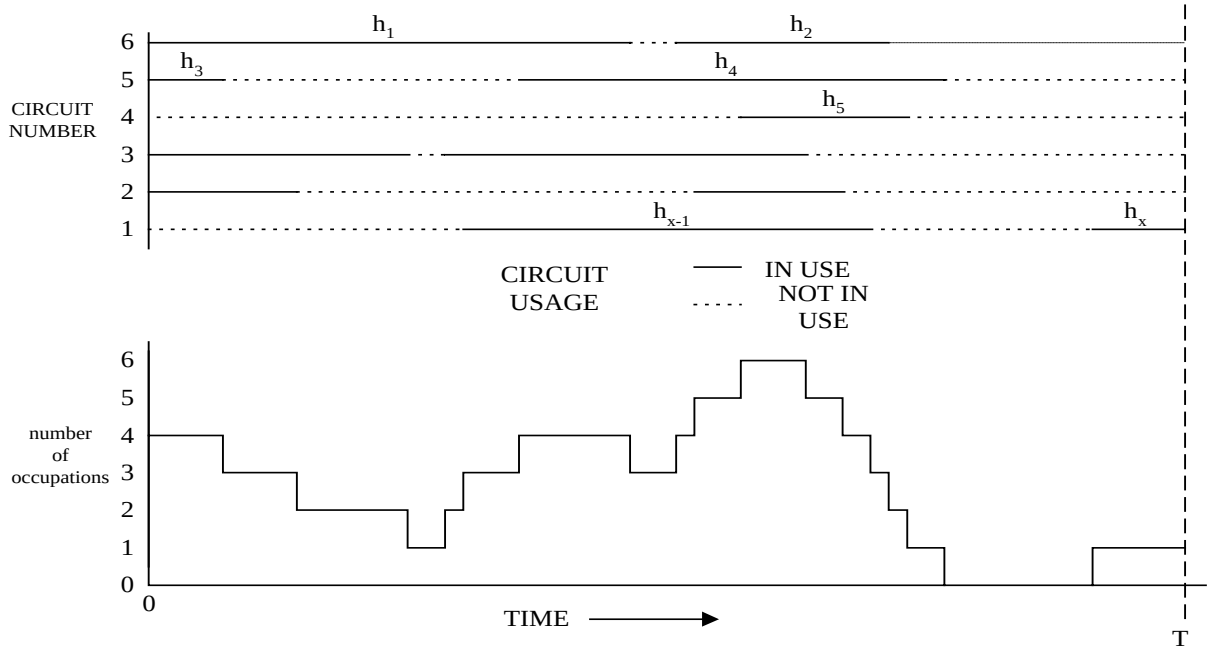

Figure TFC 1/1 : Occupations in a group of 6 devices

The total work performed by the group can, for the considered interval, be expressed as

$$x\bar{h} = \sum_{i=1}^{x} h_i$$

if it handled x calls in the interval, where $\bar{h}$ is the average of all h_i. Let T be the length of the considered interval, then

$$\frac{1}{T}\sum_{i=1}^{x} h_i$$

gives the average number of occupations.

This can also be written

$$\frac{x \cdot \bar{h}}{T}$$

and if $y = \frac{x}{T}$ is the average number of calls per time unit, we can consequently write

$$\boxed{A = y \cdot \bar{h}} \qquad \text{(TFC 1.1)}$$

which is the average number of simultaneous occupations in the interval (0, T).

The average number of simultaneous occupations is generally called *traffic intensity*, or the *traffic flow*.

Traffic units

The traffic intensity is frequently simply called traffic.

The unit for traffic is erlang. Consequently, the number of erlangs is simply the average number of simultaneous occupations for a defined time internal.

The number of erlangs (A) is usually calculated as:

$$A = y \cdot h$$

where

y = the number of calls per time unit or the number of new occupations per time unit;

h = the average duration of these occupations, expressed in the same time unit.

Example:

y =	3600 calls/hour	= 60 calls/min	= 1 call/sec
h =	$\frac{1}{60}$ hour	= 1 min	= 60 secs
$A = 3600 \times \frac{1}{60} = 60 \text{ x } 1$		= 1 x 60	= 60 erl.

The traffic can also be calculated as the sum of all occupation times divided by the length of the period concerned, as shown above.

Example:

The sum of all holding times in a group was 225 minutes during a 15 minutes interval, consequently:

$$A = \frac{225}{15} = 15 \text{ erl.}$$

A third but somewhat artificial way of calculating the traffic is to define how long a time there is one, two, three, etc., occupations in the group. Assume the group has n devices and t_p is the part of the total time interval T, that there is exactly p occupations, then the traffic can be calculated as:

$$A = \sum_{p=1}^{n} p \cdot \frac{t_p}{T}$$

where:

$$\sum_{p=o}^{n} t_p = T$$

A practical method of measuring traffic is to scan the group at regular intervals. The average of these samples then provides an estimate of the observed traffic, i.e.

$$A = \frac{1}{N} \cdot \sum_{\nu=1}^{N} f_\nu$$

where N scans are made in the observation interval and the number of occupied devices found at the ν : th scan is fν.

This method has a certain statistical uncertainty that depends on the scanning interval, T/N, the average holding time and the traffic carried in the observed group.

The traffic unit erlang is an internationally accepted unit. In the USA, however, the unit CCS (Hundred Call Seconds) is generally used in the practical traffic work. From the definition follows that:

1 erlang (erl) = 36 CCS or
100 CCS = 2.778 erl.

Traffic volume

To express the sum of all holding times carried by a group during a given period, the unit erlang-hours is used. These holding times are then expressed in hours.

Example:

A group carried 1200 calls over a 24 hours' period. The average holding time was 2 minutes = 1/30 hour.

$$1200 \times \frac{1}{30} = 40 \text{ erlang hours}$$

A frequently used unit for traffic volumes is the number of paid call minutes (pcm) per month, or per year. It is evident that pcm divided by 60 gives the number of paid erlang-hours.

The traffic volumes are mainly of interest for revenue considerations.

When dealing with the traffic concept, a clear distinction must be made between:

TRAFFIC DEMAND
TRAFFIC OFFERED
TRAFFIC CARRIED

which all are expressed in erlangs, but only the traffic carried can be measured.

2. Generation of traffic

Possible outcome of a call intent, repeated attempts

When a subscriber A wishes to speak to a subscriber B, the result is either the establishment of a conversation or that A gives up. See Figure TFC 2/1:

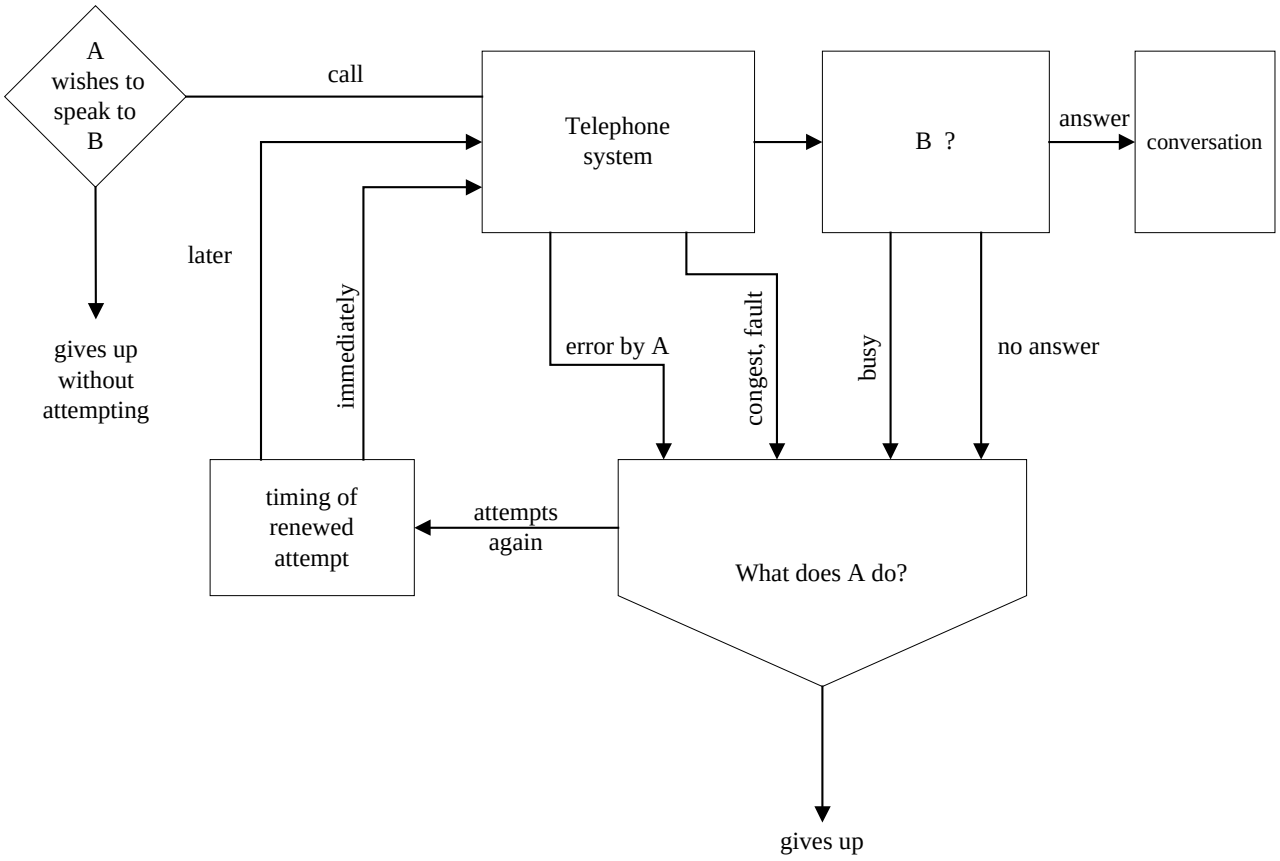

Figure TFC 2/1 : Diagram of events for establishment of conversation between A and B

It will be seen from Figure TFC 2/1 that a conversation can be established either directly or after a number of failures. A's decision to make a renewed attempt may depend on:

1) number of previous unsuccessful attempts;

2) A knows that he has dialled the wrong number or an incomplete number;

3) A is uncertain whether he has dialled correctly;

4) A thinks that there may have been some fault in the telephone system;

5) knowledge of B's telephone habits if no reply or busy signal;

6) degree of urgency.

The time of a renewed attempt is chosen having regard to:

a) A's assessment of the cause of the failure;

b) A's assessment of when it will be worthwhile to try again;

c) A's possibility of making a renewed attempt;

d) degree of urgency.

Reasons for A giving up:

α) A considers a new attempt useless;

β) degree of urgency is low.

If A gives up without making any attempt, this may mean that he has a very low opinion of the operational quality of the telephone system or of the possibility of getting hold of B. The urgency may also be low.

It will be seen from Figure TFC 2/1 that, of the calls offered to a telephone system, not all will result in conversations.

Statistics from a telephone system with reasonably good service quality may have the following appearance:

Incomplete or wrong dialling (mistake by A)	5 - 10 %
Congestion and technical faults	1 - 5 %
B engaged	10 - 20 %
B does not answer before A has replaced	10 - 15%
No conversation	30 - 50 %
Conversation	70 - 50 %
Number of attempts per successful call	1.4

For a network with less good service and high traffic per subscriber, the statistics may be as follows:

Incomplete and wrong dialling (mistake by A)	13 %
Congestion and technical faults	25 %
B engaged	20 %
B does not answer before A has replaced	10 %
No conversation	68 %
Conversation	32 %
Number of attempts per successful call	4.8

The number of dialling errors increases with the number of digits!

Duration of different types of calls

- Unsuccessful attempts usually occupy the switching equipment in the telephone system for a shorter time than conversations. The average holding time of switches is therefore shorter than the average conversation time.
- Attempts at connection can be interrupted at any moment. Therefore, the average holding time of switches at the beginning of the switching chain is shorter than that of switches at the end of the chain (see Figure TFC 2/2).

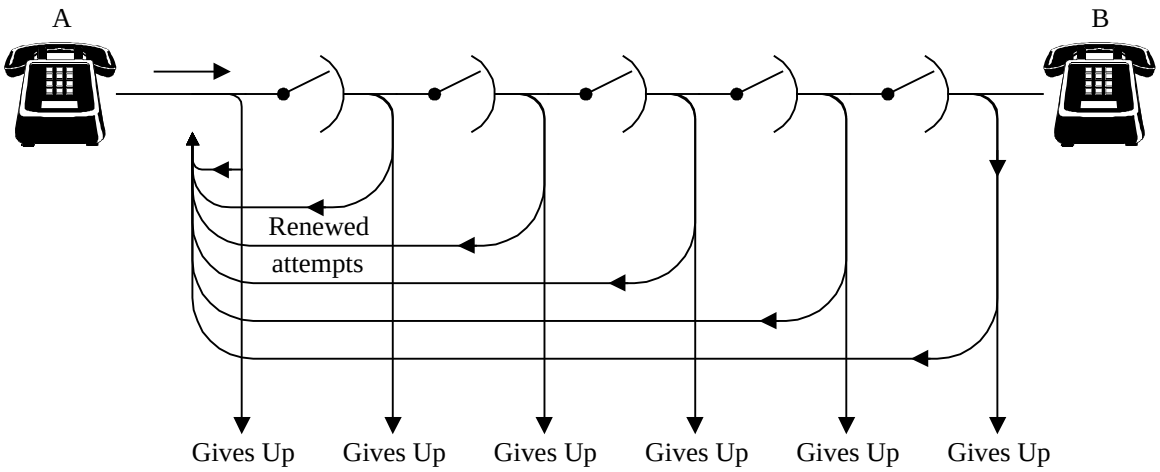

Figure TFC 2/2 : An attempted connection can be interrupted anywhere in the switching chain from A to B, but a renewed attempt always starts from the beginning again

The duration of the various types of occupations of the sector stages vary considerably as follows from this table:

Type of occupation	Typical Average holding time secs.
Conversation, local call	45 - 120
Conversation, long distance call	180 - 300
Conversation, international call	300 - 600
A subscriber, listening on ringing tone when no answer	20
A subscriber, listening on busy tone	5
Time between dialling tone and start dialling	2
Dialling time per digit	1.5 (dial) 0.6 (key-set)
Time to next attempt	30

It follows that the various switching stages carry a mixture of different types of calls and that their holding times and their frequencies may vary considerably.

The change of the average holding time in the connection chain from calling to called subscriber can be illustrated by the following values obtained in a local exchange:

Stage	Average holding time, secs
Subscriber stage, originating traffic	80
First group selector	100
Second group selector	120
Subscriber stage, terminating traffic	150
Conversation time	220

If the average holding times for the various types of unsuccessful calls and their proportions (“the call mix”) are known, the number of attempts per successful calls can be estimated as well as the completion rate, i.e. the percentage of the attempts that lead to conversations.

Traffic Concepts

It is necessary to have a clear distinction between the following four traffic concepts:

TRAFFIC DEMAND
OFFERED TRAFFIC
TRAFFIC CARRIED
CHARGED TRAFFIC

The Traffic Demand (A_D) is the traffic that the subscribers would like to realize if no obstacles were at hand, such as congestion, technical faults or busy B-subscriber. The traffic demand can also vary with the cost of the calls, so one has also to distinguish between traffic demand at fixed rate and when the telephone rate is varied. The traffic demand is a hypothetical quantity that only can be estimated but not measured.

The Offered Traffic (A_O) is the traffic offered to a group in accordance with a defined theoretical description of the traffic case. It is consequently also a hypothetical quantity and it becomes meaningful only if it is referred to a specific theoretical model.

The Traffic Carried (A_C) is the traffic that is handled by a group. It refers both to a defined theoretical description and to reality, the latter since it can be measured.

The Charged Traffic is the part of the traffic carried that is charged to the subscriber. The traffic carried and the charged traffic differ since in most cases the subscriber only pays for calls that are answered. For long distance calls and for international calls, the charging may start a little earlier, but not before the call reaches the first transit exchange. Consequently, statistics of charged traffic will only give a part of the real carried traffic.

The following expressions can be used for A_D, A_O and A_C:

$$\left.\begin{aligned} A_D &= y_D \cdot h_D \\ A_O &= y_O \cdot h_O \\ A_C &= y_C \cdot h_C \end{aligned}\right\} \qquad \text{(TFC 2.1)}$$

where y_D, y_O and y_C are the call intensities and h_D, h_O, h_C are the average holding times. Their relative order of magnitude is:

$$\begin{aligned} A_D &<> A_O \geq A_C \\ y_D &<> y_O \geq y_C \\ h_D &<> h_O \geq h_C \end{aligned}$$

The traffic carried, A_C, can be divided into two parts:

$$A_C = A_{CC} + A_{CF} \qquad \text{(TFC 2.2)}$$

where A_{CC} = conversation traffic

A_{CF} = traffic caused by unsuccessful attempts.

We can then express A_{CC} and A_{CF} as follows:

$$\left.\begin{aligned} A_{CC} &= y_{CC} \cdot h_{CC} \\ A_{CF} &= y_{CF} \cdot h_{CF} \end{aligned}\right\} \qquad \text{(TFC 2.3)}$$

where generally $h_{CF} < h_{CC}$ or $h_{CF} << h_{CC}$

while the relationship between y_{CF} and y_{CC} depends on the completion rate.

Congestion

If a call requests to be served by a group which already has all its devices occupied, congestion occurs. This means that the call cannot be accepted for the moment. Depending on the system used, such a call may either be rejected (Loss System) or be allowed to wait (Delay System). In the former case, the calling subscriber receives busy tone and must make a new trial. In the latter case, delayed calls will be served as soon as any of the devices become free. Different technical methods of handling waiting calls exist as regards the choice of next to be served. They have a certain impact on the theoretical calculation of the waiting times.

As regards the congestion in Loss System, a distinction is made between:

TIME CONGESTION
and
CALL CONGESTION

The Time Congestion is the part of the time additional calls cannot be served.

The Call Congestion is the part of the calls that are rejected or forced to wait in a delay system.

Both are quantities that can be measured. They can also be referred to a defined theoretical description of a traffic case.

3. Traffic variations

The traffic varies in step with the activity in the community. Therefore, the following types of variations are at hand:

A. *Variations during the day*

a) Traffic peaks occur once, twice or three times per day during normal working days. (Traffic is usually lower on Saturdays and public holidays.)

b) The maximum traffic does not always occur at the same time every day.

B. *Variations during the week*

Certain weekdays may have systematically higher traffic than other weekdays.

C. *Seasonal variations*

High traffic during certain parts of the year and low traffic during other parts. Often high traffic before a major public holiday (New Year, Christmas, etc.), and low traffic afterward.

D. *Growth trend*

The traffic shows generally a consistent tendency to increase. This increase is not uniform. Therefore, high and low values may have different relative increase. It is generally rather difficult to distinguish between growth and seasonal variations.

E. *Unpredictable peaks*

The highest traffic during a year may occur during a busy season, but can occasionally occur at any other part of the year.

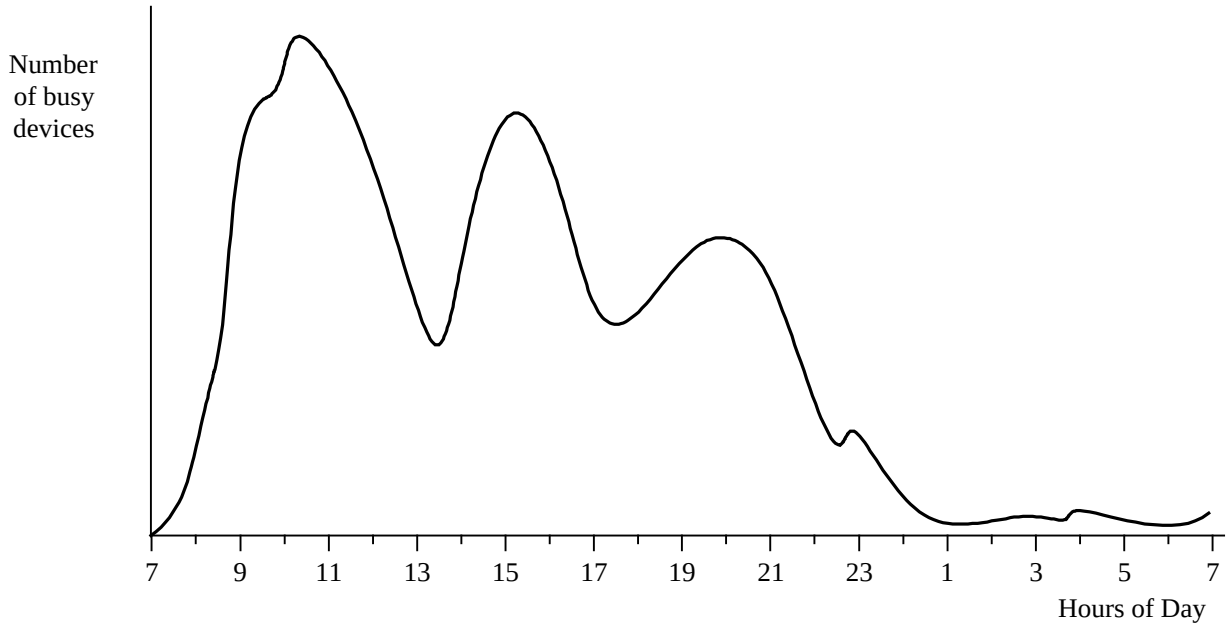

Figure TFC 3/1 : Variations of traffic during a day

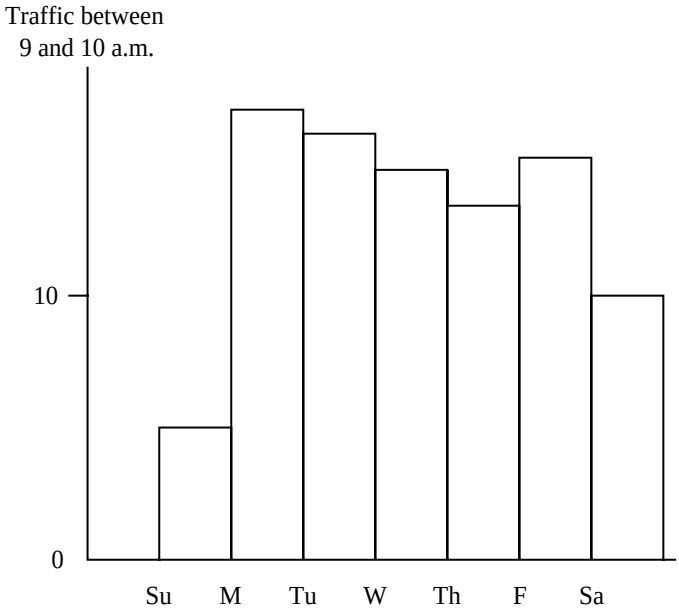

Figure TFC 3/2 : Variations of traffic during one week

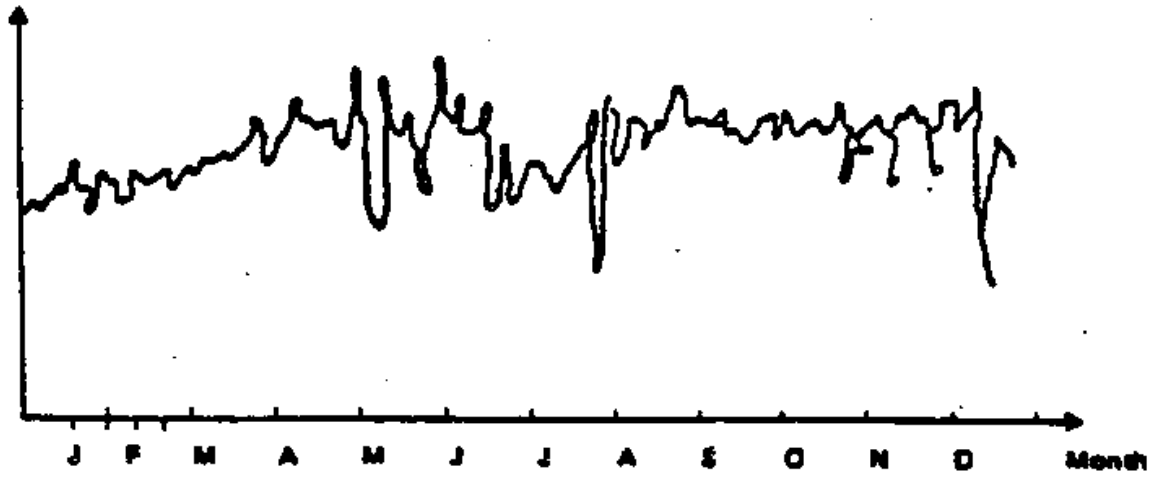

Figure TFC 3/3 : Seasonal variations during a year expressed at busy hour traffic
9-10 a.m., every third normal working day

F. *Variations within short intervals*

If the number of simultaneous occupations in a group is studied for a short interval, such as one hour, it is found that it varies in an irregular way. These variations can be described as random variations. Depending on the time of the day, the number of occupations may show a tendency with time to increase, decrease or remain at approximately constant level.

The traffic theory builds on the conditions during the busy hour, where the variations generally occur around a mean level, with less tendency to increase or decrease than at other parts of the day.

The traffic shows consequently random variations as well as more or less systematical variations. To understand the significance of traffic values given erlangs, it is, therefore, necessary to have information on how they are observed. Samples taken during peak seasons, less high seasons and during slack seasons may have quite different significance.

The BUSY HOUR is the 60-minute period during the day when the traffic is the highest in the LONG RUN (time-consistent busy hour).

The BUSIEST HOUR is the 60-minute period every day when the traffic is highest.

Consequently, the busy hour and the busiest hour may not concur, and the busiest hour may occur at different times of the day.

The variation of the traffic during a lengthy period (e.g. the traffic for each hour during a year) can be described by a histogram (Figure FTC 3/4).

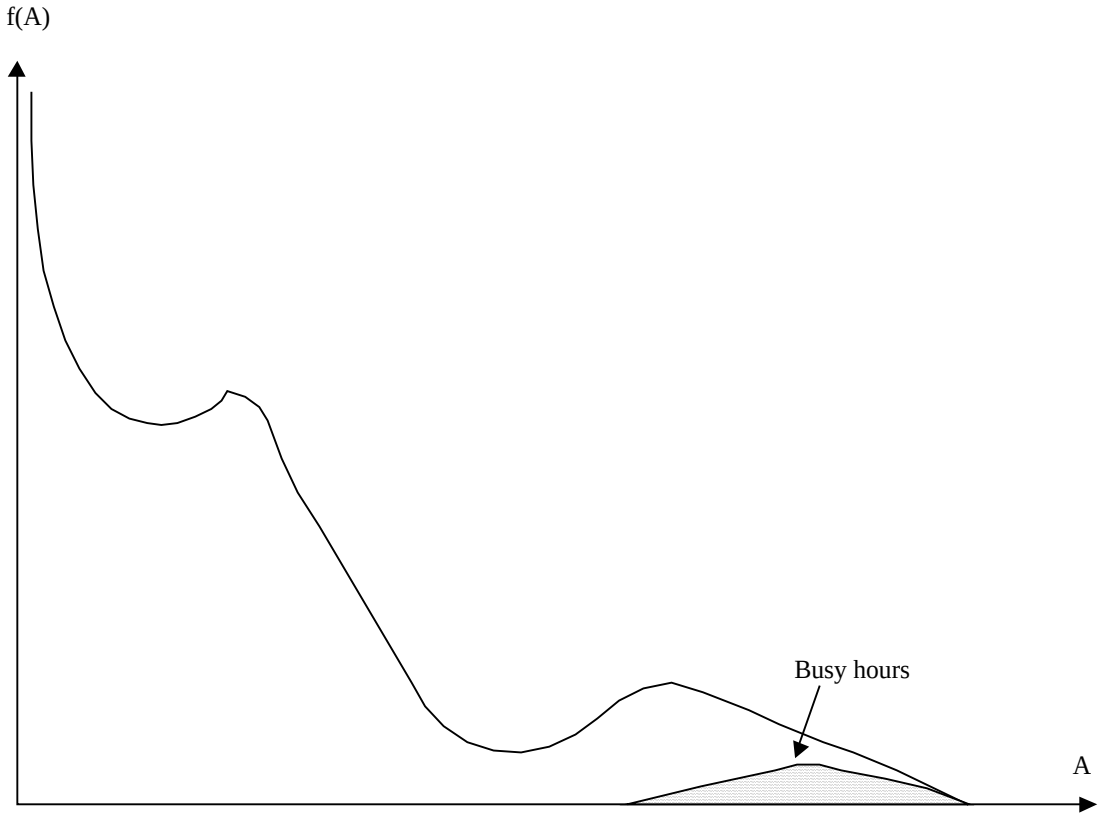

Figure FTC 3/4 : Traffic variation during all 8760 hours of a year.
The shaded area is the contribution from the busy hours

If the traffic during all 8760 hours of a year is represented in the histogram, it is found that there is a considerable number of hours with very low traffic. The highest values derive mostly from the busy hours (shaded area in the diagram) but hours outside the busy hour may also contribute to the upper tail of the distribution.

The average holding times as well may have both daily and seasonal variations, but of lesser magnitude than the traffic variations.

The present traffic theory can only be used for practical calculation of the congestion when the traffic is in equilibrium, which corresponds roughly to the conditions during traffic peaks.

4. Dimensioning principles

Due to the structure and design of telecommunication systems, the equipment - provision cannot be made to match all short time variations of the traffic. The various parts of the system must, therefore, be extended in steps large enough to last for a certain period, usually at least 6 months up till 5 years. The extension steps must be so large that disturbing congestion can be avoided during the whole period.

For such an extension period, a traffic forecast is made which should describe how the traffic will vary and grow. The provision depends on the following facts:

1) The revenue from telephone traffic will depend on the traffic carried, usually on the number of conversations established.

2) The expenditure depends on the grade of service it is desired to give to the subscribers under peak traffic conditions.

3) Most of the carried traffic volume is handled at times when there is practically no congestion at all.

4) Really high congestion occurs only on a few occasions.

5) These occasions of high congestion occur more often at the end of an extension period than at the beginning.

6) It is not economically feasible to dimension a telephone system so that congestion never occurs.

7) An improved grade of service usually results in a certain growth of traffic when the subscribers notice that it has become easier to get through (service improvement jump or traffic stimulus).

8) Human factors always cause more failures than the telephone system. By “human factor”is meant: mistakes by calling subscriber, called subscriber engaged, and no answer.

9) The individual subscriber does not react much to congestion below, say, 10%.

10) The peak traffic is usually comparatively unaffected by minor changes of rates. This is because the bulk of the traffic during peak hours comes from business subscribers and not from private persons.

The dimensioning problem:

How much congestion can be permitted during a few hours of a year?

During a given period, the variations of traffic during all hours of the year can be described as shown in Figure TFC 4/1:

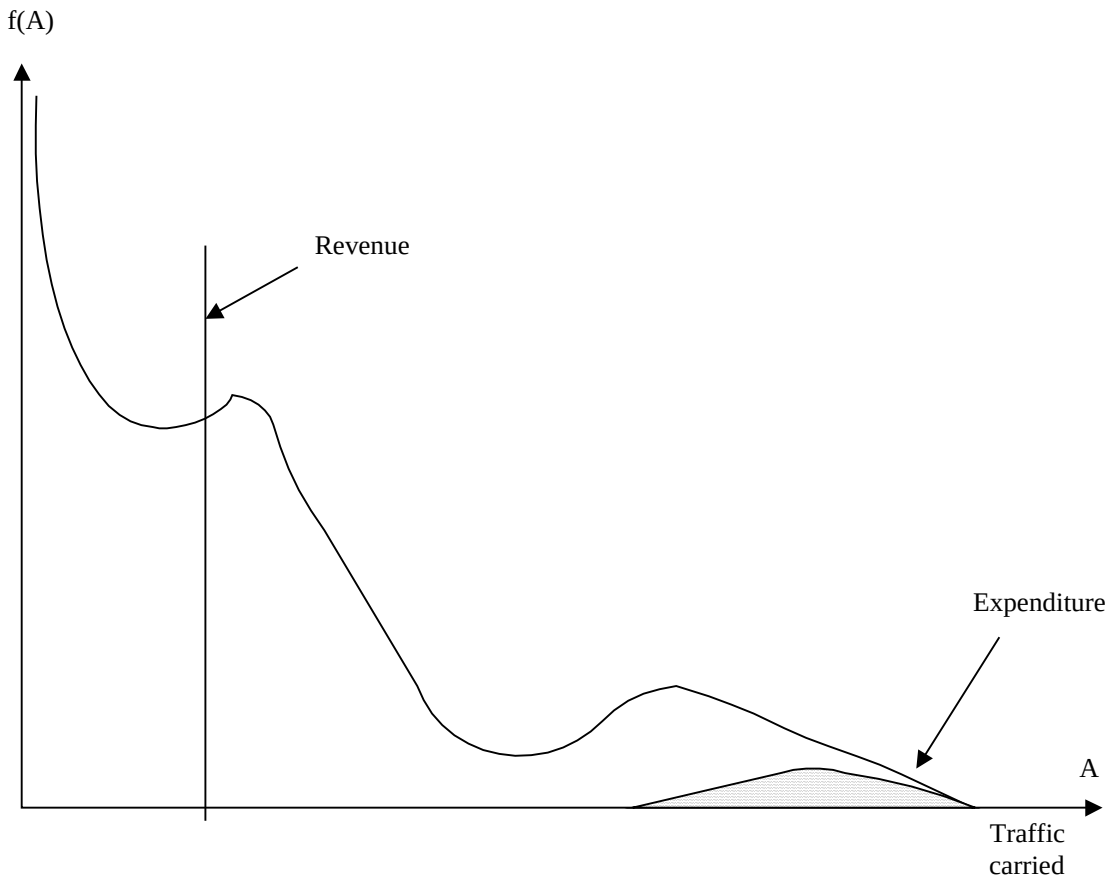

Figure TFC 4/1 : Variations of traffic (all hours) during a period of extension

If a certain number of circuits is used for carrying the above traffic, the congestion will vary as shown in Figure TFC 4/2:

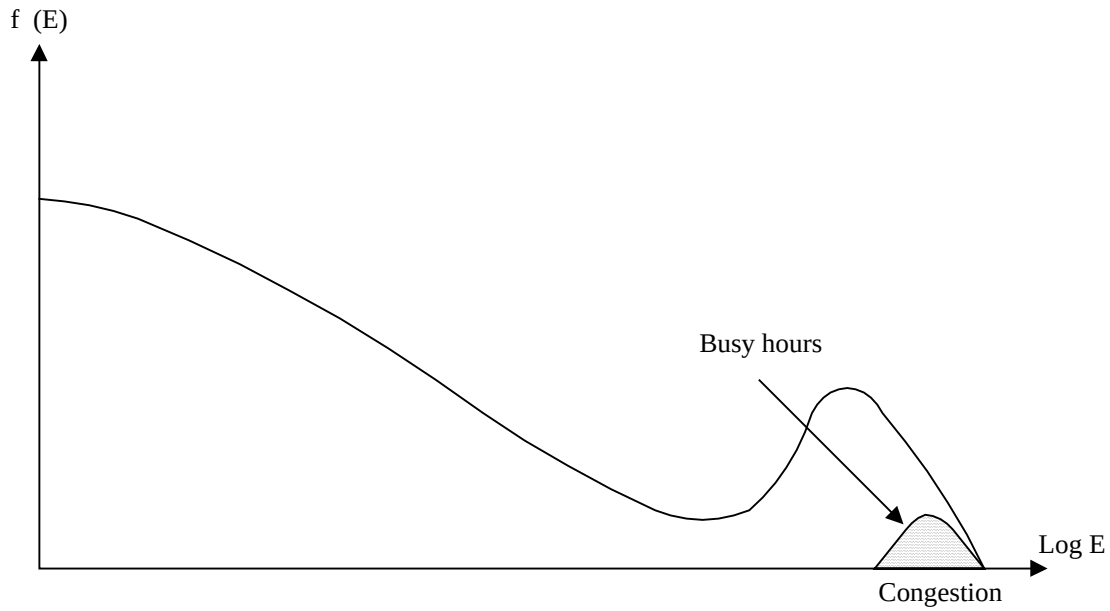

Figure TFC 4/2 : Variation of congestion (all hours) during a period of extension

The revenue is approximately proportional to the traffic carried. This means that the income is proportional to the area between f(A) and the x-axis in Figure TFC 4/1.

The expenditures depend mainly on how often the congestion is permitted to exceed certain given levels.

If more devices are provided in group, measurable congestion will occur less frequently and, of course, disturbingly high congestion will also become a more rare event. It is intuitively understood from the diagrams in Figures TFC 4/1 and TFC 4/2 that a minor change in the number of devices in a group does not cause much change in the revenues, while the expenditures may be more direct proportional to the number of devices provided. The dimensioning, therefore, becomes mainly a question of balancing expenditures to the requirements for the grade of services, where the grade of service is expressed in relation to how frequently the congestion is permitted to exceed certain values.

This is the general statement of the dimensioning principle. To apply it in practice, the telephone administrators have to keep to what can be practically observed. Since the carried traffic and the congestion cannot at present be observed continuously, one has to confine to what is really measured.

A common procedure is to measure the busy hour traffic on 5 or 10 specific days during a high traffic season every year. More measurements may be made on very expensive equipment. Generally, the congestion is also measured simultaneously. To control the daily service performance congestion values exceeding certain given high levels may also be recorded outside these measuring periods.

The annual measurements are generally used as a basis for the traffic forecasts. Since it is known that these measuring periods generally do not catch the very highest traffic peaks of the year, the switch and trunk provision is chosen so as to give a rather low congestion, generally between 0.1 and 5 %. This would then hopefully safeguard against disturbingly high congestion (say 20 % or more) occurring too frequently in the extension period, provided that the forecast is correct.

The mechanism of provisioning periods, forecasted traffic increase and resulting grade of service is illustrated in Figure 4/3 where, of course, the grade of service has to be referred to a defined traffic level.

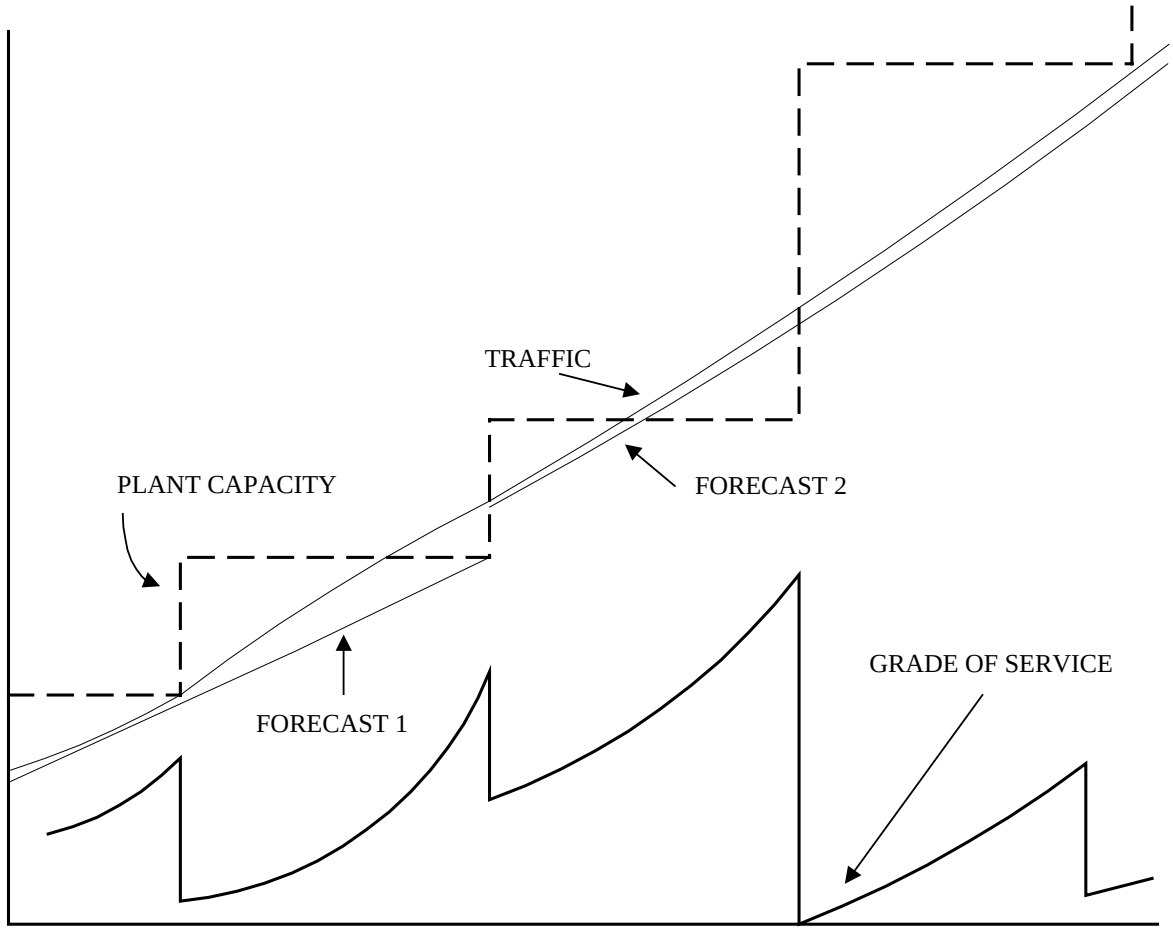

Figure TFC 4/3 : Effect of provisioning period on grade of service

Choice of extension periods

The choice of suitable extension steps and periods is of great importance for the best economical use of available capital for investments. It is so because the telephone networks and the telephone traffic continue to grow.

The choice of suitable extension periods depends ;on a number of factors. The most important ones are:

- growth rate;
- cost of additional equipment;
- technically suitable extension units;
- delivery times;
- time required for installation of an extension;
- coordination with other extensions;
- flexibility towards forecasting errors;
- costs of installation;
- interest on invested capital;
- income from traffic;
- limits for capital to be invested.

It is understood that expensive equipment should be extended in small steps while less expensive equipment can be extended in larger steps. The optimal extension steps are, however, depending on the total extension programme for a whole country and cannot be considered for each single case. Limited resources for investment may, however, sometimes force a telephone administration to use less economic extension steps.

5. Connection cases

A connection through a telephone system takes place in a number of successive selections. For calculation and assessment of the traffic characteristics of each selection step, the following factors must be defined:

a) traffic input to the stage;

b) grouping of the stage;

c) how the stage handles incoming calls (method of hunting);

d) the procedure for dealing with unsuccessful calls.

Equilibrium

The usual assumption in calculations is statistical equilibrium - "stationarity" - the traffic is neither rising nor falling, and the traffic process has been going on for so long that the initial state is of no significance.

Traffic input

The traffic input is defined by the call intensity and the holding time distribution.

Call intensity:

As a rule, the traffic is described according to the number of single sources (a single source can only produce one call at a time). For the call intensity, the usual assumptions are:

BE : Bernoulli or Engset input

Limited number of sources, call intensity falling with the number of occupations:

$$y(p) = (N - p) \cdot \beta \qquad \text{(TFC 5.1)}$$

N = number of sources
p = number of occupations
β = call intensity per source when free (and engaged source has the call intensity 0)
P = Poisson input

Infinite number of sources, call intensity independent of the number of occupations:

$$y(p) = y; \qquad y = lim\,(N\beta) = finite$$

$$N \to \infty$$

$$\beta \to 0 \qquad \text{(TFC 5.2)}$$

NB : Input of negative binomial type

The call intensity is assumed to increase with the number of occupations:

$$y(p) = a + b \cdot p \qquad \text{(TFC 5.3)}$$

$$y(p) = (N - p)\,\beta \qquad a = N \cdot \beta$$

$$b = -\beta$$

i.e. NB is obtained from BE if $\beta < 0$ and $N < 0$.

NB is sometimes used for description of traffic with special characteristics.

Holding distribution:

For the holding time distribution, a negative exponential distribution is usually assumed.

$$f(t) = \frac{1}{s} \cdot e^{-\frac{t}{s}} \qquad \text{(TFC 5.4)}$$

Constant holding time may also exist. Cases involving both constant and exponentially distributed holding times occur frequently in practice.

The description of input traffic should be based on the factual conditions, such as the number of simultaneous calls which can be produced by the inlets.

Grouping

Depending on the availability between inlets and outlets, a grouping arrangement is classified as follows:

1. Full availability group: every inlet has access to every outlet.

2. Grading: an inlet has access only to a limited number of outlets (limited availability). Inlets with access to exactly the same outlets constitute an inlet group. At least some outlets (outgoing circuits) are common to more than one inlet group. No inlet group is independent of the others. Used when more circuits are required than the individual selector has access to.

3. Link system: connection from inlet to outlet takes place over one or more link stages. Conditional selection is applied. Conditional selection = connection effected only if the inlet can find a free path to the desired outlet or outlets. Requires information concerning free and busy links and outlets in the link system and a knowledge of how devices in different stages can be combined, i.e. requires marker or processor control. Used particularly in crossbar systems and programme memory systems.

4. Graded link system: combination of 2 or 3.

Hunting methods

There are chiefly two hunting methods:

1. Sequential hunting: hunting for a free outlet always starts from the same position; the first free outlet is seized.

2. Random hunting: every free outlet has the same probability of being seized.

The method of hunting is of less significance for full availability groups.

Different design of grading and link systems for different hunting methods is usually advisable.

Unsuccessful calls

Depending on whether, on encountering congestion, the subscriber can wait until served or must make a new attempt, telephone systems are classified as:

1. loss systems;
2. Delay systems.

For loss systems, it can be assumed, for example, that:

a) unsuccessful calls give rise to no increase of the call intensity from blocked sources;

b) unsuccessful calls give rise to renewed attempts with a given probability.

Assumption a) leads to simple traffic models.

For delay systems it may, for example, be assumed that:

a) unsuccessful calls wait until served;

b) unsuccessful calls are abandoned with a given probability;

c) unsuccessful calls wait for a given maximum time.

Assumption a) leads to the simplest traffic models.

Queue discipline

Depending on how the waiting calls are picked out of the queue in a delay system, a distinction is made between the following queue-handling methods:

a) ordered queue (first come, first served);
b) random queue : every waiting call has the same probability of being served next, irrespective of how long it has queued;
c) priority queue : every queuing call has a priority number depending on where it comes from or depending on the inlet on which it is waiting; the call with the highest priority is served first.

Method a) usually leads to the simplest mathematical models, whilst c) may be the method most used in practice.