LESSON 8:
TELETRAFFIC, NETWORK TRAFFIC LOAD AND PARAMETERS, GRADE OF SERVICE AND BLOCKING PROBABILITY

Objective

The objective here is to learn about teletraffic, network traffic loads and its parameters, grade of service and blocking probability.

Introduction

Teletraffic theory is defined as the application of probability theory to the solution of problems concerning planning, performance evaluation, operation and maintenance of telecommunication systems. More generally, teletraffic theory can be viewed as a discipline of planning where the tools (stochastic processes, queueing theory and numerical simulation) are taken from the disciplines of operations research.

The term teletraffic covers all kinds of data communication traffic and telecommunication traffic. The theory will primarily be illustrated by examples from telephone and data communication systems. The tools developed are, however, independent of the technology and applicable within other areas such as road traffic, air traffic, manufacturing and assembly belts, distribution, workshop and storage management, and all kinds of service systems.

The objective of teletraffic theory can be formulated as follows:

" to make the traffic measurable in well defined units through mathematical models and to derive the relationship between grade-of-service and system capacity in such a way that the theory becomes a tool by which investments can be planned. "

The task of teletraffic theory is to design systems as cost effectively as possible with a predefined grade of service when we know the future traffic demand and the capacity of system elements. Furthermore, it is the task of teletraffic engineering to specify methods for controlling that the actual grade of service is fulfilling the requirements, and also to specify emergency actions when systems are overloaded or technical faults occur. This requires methods for forecasting the demand (e.g. based on traffic measurements), methods for calculating the capacity of the systems, and specification of quantitative measures for the grade of service.

When applying the theory in practice, a series of decision problems concerning both short term as well as long term arrangements occur.

Short term decisions include a.o. the determination of the number of circuits in a trunk group, the number of operators at switching boards, the number of open lanes in the supermarket, and the allocation of priorities to jobs in a computer system.

Long term decisions include e.g. decisions concerning the development and extension of data and telecommunication networks, the purchase of cable equipment, transmission systems etc. The application of the theory in connection with design of new systems can help in comparing different solutions and thus eliminate non-optimal solutions at an early stage without having to build up prototypes.

Network Traffic, Load & Parameters

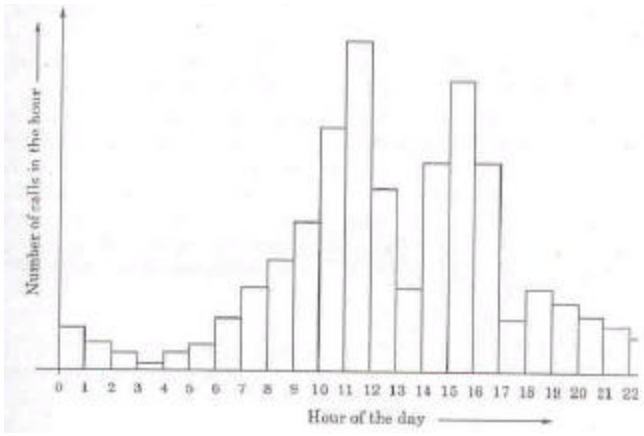

Fig. 8.1 Typical telephone traffic pattern on a working day during tbe 24 hours.

In a telephone network, the traffic load on a typical working day during 24 hours is shown in Fig.8.1. The amplitudes of originating call are relative and the actual values depend on the area where the statistics is collected. The traffic pattern, however, is the same irrespective of the area considered. It is obvious that there is little use of the telephone network during 0 to 6 hours when most of the population is asleep. There is a large peak during 10:00 to 11:00 hour and 15:00 to 16:00 hours. These hours are busy due to office activities. The time duration 15 to 16 hours is, however, slightly smaller. The load is low during the lunch-hour period (12:00-14:00 hours). The period 17:00 – 18:00 hours is characterized by low traffic. During this period the people are on the move from offices to their residences.

Generally, there is a peak of calls around 10.00 hours just before people leave their homes for work and another peak occurs again in the evening.

In a day, the one-hour (60 minute) interval in which the traffic is the highest is known as the Busy Hour (BH). In Fig. (7.1), one-hour period between 11:00 to 12:00 hours is the busy hour. The busy hour may vary from exchange to exchange depending on the location and the community interest of the subscribers.

The busy hour may vary seasonally, weekly or daily. In addition to these variations in busy hour, there are also unpredictable peaks caused by stock market or money market activity, weather, natural disaster, international events, sporting events etc. Busy hours are three types on the basis of fluctuations that are taken

into account while designing switching networks. These busy hours are defined by CCITT in its recommendations:

1. **Busy Hour (BH)**: Continuous l-hour period lying wholly in the time interval concerned, for which the traffic volume or the number of call attempts is greatest.
2. **Peak Busy Hour (PBH)**: The busy hour each day; it usually varies from day to day, or over a number of days.
3. **Time consistent Busy Hour**: The one-hour period starting at the same time each day for which the average traffic volume or the number of calls attempts is greatest over the days under consideration.

For ease of records, the busy hour is taken to commence on the hour or half-hour only. All the call attempts are not materialised into actual conversations for a variety of reasons such as called line busy, no answer from the called line and blocking in the trunk groups or the switching centres. A call attempt is said to be successful if the called party answers. The ratio of number of successful calls to the number of call attempts is known as call completion rata (CCR). The number of call attempts in the busy hour is called Busy Hour Call Attempts (BHCA).

Busy Hour Call Attempts (BHCA) is an important parameter in deciding the processing capacity of a common control or a stored program control (SPC) system of a telephone exchange. The CCR parameter is used in dimensioning the network capacity. Networks are usually designed to provide an overall CCR of over 0.70. A call completion rate (CCR) of 0.75 is considered excellent. Attempts to further improve the value of CCR is generally not cost effective.

Busy Hour Calling Rate (BHCR)

This is a related parameter that is used in traffic engineering calculations. It is defined as the average number of calls originated by a subscriber during the busy hour.

The busy hour calling rate (BHCR) is useful for the size determination of the telephone exchange to handle the peak traffic. The BHCR in rural telephone exchange may be as low as 0.2 but in a business city, it may be as high as 3 or more. We can get another useful information from BHCR that how much of the day's traffic is carried during the busy hour. This is measured in terms of Day-to-Busy Hour traffic ratio. It is the ratio of BHCR to the average calling rate for the day.

Typically, this ratio may be over 20 for a city business area and around 6 or 7 for a rural area. The traffic load on a given network may be on the local switching network. Interoffice trunk lines or other common subsystems. All the common subsystems of a telecommunication network are collectively known as servers. The traffic on the network may then be measured in terms of the occupancy of the-servers in the network. These servers are also called link or trunk.

Traffic Intensity

It is used to measure the traffic on the network. The occupancy of servers or link or trunks is called the traffic on the network. It is called traffic intensity. It is the ratio of period for which a server is occupied to total period of observation. Traffic intensity is denoted by I_T. It is defined as

$$I_T = \frac{\text{Period for which a server is occupied}}{\text{Total period of observation}}$$

Generally, we take period of observation of 1 hour. Obviously, quantity I_T is dimensionless. It is called erlang (E) to honour the Danish telephone engineer A.K. Erlang. The pioneering work in traffic engineering was done by A.K.Erlang. A server said to have 1 erlang of traffic, if it is occupied for the entire period of observation. Traffic intensity may also be specified over a number of servers.

Another Way of Measurement of Traffic Intensity

This measure is known as Centum Call Second (CCS), which represents a call time product. One CCS means that one call for 100 seconds duration or 100 calls for one-second duration each or any other combination. CCS (a number of traffic intensity) is valid only in telephone circuits. At present, telephone networks support voice, data and many other services. So, erlangs is a better measure to use for representing the traffic intensity. Sometimes, Call Seconds (CS) and Call Minutes (CM) are also used as a measure of traffic intensity.

We have following relationship between erlangs (E), Centum Call Second (CCS), Call Seconds (CS) and Call Minutes (CM).

$$1E = 36\ CCS = 3600\ CS = 60\ CM$$

Subscriber traffic or trunk traffic:

Subscriber traffic or trunk traffic is the traffic intensity contributed by subscriber or traffic intensity on a trunk. We already know that traffic intensity is a call-time product. Hence we require two important parameters for finding the estimate of the traffic intensity or network load. These parameters are:

1. Average Call Arrival Rate (R)
2. Average holding time per call, t_h

Now we can express the load offered to the network in terms of above two parameters as

$$A = R * t_h$$

Both R and t_h must be expressed in like time units. For example, if R is in number of calls per minute then t_h must be in minutes per call.

Traffic can be calculated in two ways: One based on the traffic generated by the subscribers and the other based on the observation of busy servers in the network. It is possible that the load generated by the subscribers sometimes exceeds the network capacity. In this situation, the traffic is called overload traffic.

Therefore, overload traffic is handled in two ways:

1. The overload traffic may be rejected without being serviced or held in a queue until the network facilities become available. In the first case, the calls are lost. In the second case, the calls are delayed. Correspondingly, there will be two types of systems. These systems are called loss systems and delay systems.
2. Conventional automatic telephone exchanges behave like loss systems. Under overload traffic conditions a user call is blocked and is not serviced unless the user makes a retry.

On the other hand, operator oriented manual telephone exchanges are the example of delay systems. A good operator

registers the user request and establishes connection as soon as network facilities become available without the user having to make another request.

In data networks, circuit-switched networks behave as loss systems. But store-and-forward (S&F) message and pocket networks behave as delay systems. In the limiting case, delay systems behave as loss systems. For example, in S&F network if the queue buffers become full, then further requests have to be rejected.

The basic performance parameters for a less system are the:

1. Grade of service (GOS)
2. Blocking probability.

The basic performance parameters for a delay system are the service delays. Average delays or probability of delay exceeding a certain limit or variance of delays may be important under different circumstances.

Blocking or congestion models are used for studying loss systems. Queuing models are used for studying delay systems.

Example 8.1: Find the Busy Hour Calling Rate for a telephone exchange which serves 2000 subscribers. For this telephone exchange following parameters are given:

Average BHCA = 10,000

CCR = 60%

Solution:

Average busy hour calls = BHCA * CCR

= 10,000 * 0.60

= 6000 calls

Busy hour calling rate

$$(BHCR) = \frac{\text{Average busy hour calls}}{\text{Total Number of subscribers}}$$

$$= \frac{6000}{2000}$$

$$= 3$$

Example 8.2: Determine the traffic carried by the group. A group of 10 servers and each occupied for half an hour in an observation interval of 2 hours.

Solution:

$$\text{Traffic carried per server} = \frac{\text{Occupied duration}}{\text{Total duration}}$$

$$= \frac{30 \text{ minutes}}{120 \text{ minutes}}$$

$$= 0.25 \text{ erlang or E}$$

Total traffic carried by the group = Traffic carried per server * No. of Servers in group

= 0.25 x 10

= 2.5 erlang or E.

Total traffic carried by the group is 2.5E

Erlang measure indicates the average number of servers occupied. It is useful in deriving the average number of calls put through during the period of observation.

Example 8.3: A group of 20 servers carry traffic of 10 erlangs (E). The average duration of a call is 3 minutes. Determine the number of calls put through by a single server and the group as a whole in 2-hour period.

Solution:

$$\text{Traffic per server} = \frac{\text{Traffic intensity}}{\text{Number of servers in a group}}$$

$$= \frac{10}{20}$$

$$= 0.5 \text{ erlangs}$$

It means that a server is busy for half an-hour *i.e.* 30 minutes in one hour.

$$\text{Number of calls put through by one server} = \frac{\text{Time duration for which a server is busy}}{\text{Average duration of a call}}$$

$$= \frac{30 \text{ minutes}}{3 \text{ minutes}}$$

$$= 10 \text{ calls}$$

Total number of calls put through by the group = Number of calls put through by one server * Number of servers in a group

= 10 * 20

= 200 calls

Number of calls put through by one server is 10 calls

Total number of calls put through by the group is 200 calls

Example 8.4: A subscriber makes four phone calls for duration of 5 minutes, 4 minutes, 3 minutes and 2 minutes in one-hour period. Calculate the subscriber traffic in erlangs, C

CS and CM.

Solution:

$$\text{Busy Period Subscriber traffic in erlangs} = \frac{\text{Busy Period}}{\text{Total Period}}$$

$$= \frac{(5+4+3+2) \text{ minutes}}{60 \text{ minutes}}$$

$$= 0.233 \text{ E}$$

$$\text{Subscriber traffic in CCS} = \frac{\text{Busy Period}}{\text{Total Period}}$$

$$= \frac{(5+4+3+2) \text{ minutes}}{100 \text{ seconds}}$$

$$= \frac{(5+4+3+2) \text{ minutes}}{(100/60) \text{ minutes}}$$

$$= \frac{40 * 60}{100}$$

$$= 8.4 \text{ CCS}$$

$$\text{Subscriber traffic in CM} = \frac{\text{Busy Period}}{\text{Total Period}}$$

$$= \frac{(5+4+3+2) \text{ minutes}}{1 \text{ minute}}$$

$$= 1.4 \text{ CM}$$

Subscriber traffic in erlang is 0.233E

Subscriber traffic in CCS is 8.4 CCS

Subscriber traffic in CM is 1.4 CM

Example 8.5: 40 Number of subscribers initiate calls in 20-minute observation interval. Total duration of the calls is 4800 seconds. Calculate the load offered to the network by the subscribers' and the average subscriber traffic.

Solution

Total number of calls initiated by the subscribers

$$\text{Mean or average Call rate (R)} = \frac{\text{Total number of calls initiated by the subscribers}}{\text{Total Observation interval}}$$

$$R = \frac{40 \text{ calls}}{20 \text{ minutes}}$$

$$R = 2 \text{ calls/minute}$$

$$\text{Mean or average holding time } (t_h) = \frac{\text{Total duration of the calls}}{\text{Total number of calls initiated by the subscribers}}$$

$$t_h = \frac{4800 \text{ seconds}}{40 \text{ calls}}$$

$$t_h = \frac{(4800/60) \text{ minutes}}{40 \text{ calls}}$$

$$t_h = 2 \text{ minutes/call}$$

Load offered to the network by the subscriber = Average call arrival rate (R) * Average holding time per call

$$= 2 * 2$$

$$= 4 \text{ erlangs or } 4E$$

$$\text{Average subscriber traffic} = \frac{\text{Offered Load}}{\text{Total number of subscribers}}$$

$$= \frac{4}{40}$$

$$= 2 \text{ calls/minute}$$

Load offered to the network by the subscriber is 4E

Average subscriber traffic is 2calls/minute

Grade Of Service And Blocking Probability

Grade of Service (GOS)

Grade of Service (GOS) is defined as a number of traffic engineering parameters to provide a measure of adequacy of plant under specified conditions; these GOS parameters may be expressed as probability of blocking, probability of delay, etc. Blocking and delay are caused by the fact that the traffic handling capacity of network or of a network component is finite and the demand traffic is stochastic by nature.

GOS is the traffic related part of network performance (NP), defined as the ability of a network or network portion to provide the functions related to communications between users. Network performance does not only cover GOS (also called traffic ability performance), but also other non-traffic related aspects as dependability, transmission and charging performance.

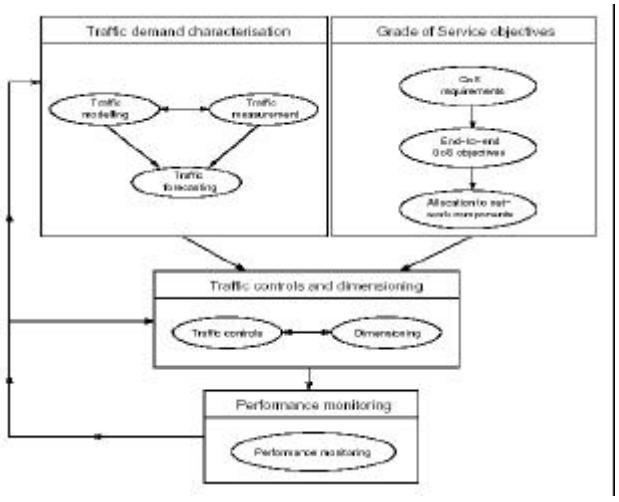

Fig. 8.2 Traffic engineering tasks

NP objectives and in particular GOS objectives are derived from Quality of Service (QOS) requirements, as indicated in Fig. 8.2. QOS is a collective of service performances that determine the degree of satisfaction of a user of a service. QOS parameters are user oriented and are described in network independent terms. NP parameters, while being derived from them, are network oriented, i.e. usable in specifying performance requirements for

particular networks. Although they ultimately determine the (user observed) QOS, they do not necessarily describe that quality in a way that is meaningful to users.

QOS requirements determine end-to-end GOS objectives. From the end-to-end objectives,a partition yields the GOS objectives for each network stage or network component. This partition depends on the network operator strategy. Thus ITU recommendations only specify the partition and allocation of GOS objectives to the different networks that may have to cooperate to establish a call (e.g. originating national network, international network and terminating national network in an international call).

In order to obtain an overview of the network under consideration and to facilitate the partitioning of the GOS, ITU Recommendations provide the so-called reference connections. A reference connection consists of one or more simplified drawings of the path a call (or connection) can take in the network, including appropriate reference points where the interfaces between entities are defined. In some cases a reference point define an interface between two operators.

In loss systems, the traffic carried by the network is generally lower than the actual traffic offered to the network by the subscribers. In these systems, overload traffic is rejected and hence is not carried by the telephone network. The amount of traffic rejected by the telephone network is determined by an index of the quality of the service offered by the network. The index of the quality of the service offered by the telephone network is termed as Grade of Service (GOS).

It is defined as the ratio of lost traffic to offered traffic

$$GOS = \frac{\text{Lost traffic}}{\text{Offered traffic}}$$

Offered traffic is the product of the average number of calls generated by the users and the average holding time per call.

Offered traffic = Average number of calls generated by the users * Average holding limit per call

Actual traffic carried by the telephone network is called the carried traffic.

Grade of Service (GOS) is given by,

$$GOS = \frac{A - A_0}{A}$$

Where,

A = Offered traffic

A_0 = Carried traffic

Therefore, lost traffic is given by

Lost Traffic = Offered traffic - carried traffic = A – A0

If the grade of service (GOS) is smaller then service is better. The recommended value for GOS in India is 0.002. The meaning of GOS = 0.002 that two calls in every 1000 calls or one call in every 500 calls may be lost. Generally, every common subsystem in a network has an associated GOS value. The GOS of the complete or full network is determined by the highest GOS value of the subsystems in a simplistic sense.

A better estimate takes into account the connectivity of the subsystems such as parallel units. Volume of traffic grows as time passes by. But the GOS value of the network deteriorates with time. In order to maintain the value within reasonable limits, initially the telephone network is sized to have a much smaller GOS value than the recommended one so that the GOS value continues to be within limits as the network traffic grows.

Blocking Probability (PB)

The Blocking Probability (PB) is defined as the probability that all the servers in a system are busy. When all the servers or links or trunks are busy), no further traffic can be carried by the system. Therefore, the arriving subscriber traffic is blocked. From first instance point of view, the blocking probability is the same measure as the GOS. The probability that all the servers are busy may well represent the fraction of the calls lost. In a system with equal number of servers and subscribers, the GOS is zero as there is always a server available to a subscriber. On the other hand, there is a definite probability that all the servers are busy at a given instant and hence the blocking probability is non-zero. The fundamental difference between GOS and blocking probability (BH) is that the GOS is a measure from the subscriber point of view whereas the blocking probability is a measure from the telephone network or switching system point of view. GOS is arrived at by observing the number of rejected subscriber calls.

But the blocking probability is arrived at by observing the busy servers in the switching system. Through analysis carried out on the loss systems, we shall see that GOS and P_B may have different values. These values depend upon the traffic characterization model used. In order to distinguish between GOS and P_B terms clearly, GOS is called call congestion or loss probability and the blocking probability is called time congestion.

In the case of delay systems, the traffic carried by the telephone network is the same as the load offered to the network by the subscribers. In delay systems, the overload traffic is queued; all calls are put through the network as and when the network facilities become available. GOS is not meaningful in the case of delay systems because it has a value of zero always.

The probability that a call experiences delay is called delay probability. It is a useful measure for the delay systems. If the offered load (or the input rate of traffic) far exceeds the network capacity then the queue lengths become very large. Therefore, calls experience undesirable long delays. Under such circumstances, the delay systems are said to be unstable as they would never be able to clear the offered load.

In practice, it is possible that there are some spurts of traffic, which tend to take the delay systems to the unstable region of operation. We have an easy way of bringing the system back to stable region of operation is to make it behave like a loss system until the queued up traffic is cleared to an acceptable limit. This technique of maintaining the stable operation is known as flow control.

In more recent times, a new term called Quality of Service (QOS) is being used. It is considered to be more general than the GOS. It includes some other factors like quality of speech,

error-free transmission capability etc. It is essential to recognize that there are two different important performance measures.

One measure is obtained by observing the subscriber traffic and the other obtained by observing the network behaviour.

Busy Hour Calling Rate (BHCR) is 3

Total traffic carried by the group is 2.5E

Number of calls put through by one server is 10 calls

Total number of calls put through by the group is 200 calls

Subscriber traffic in erlang is 0.233E

Subscriber traffic in CCS is 8.4 CCS

Subscriber traffic in CM is 1.4 CM

Load offered to the network by the subscriber is 4E

Average subscriber traffic is 2calls/minute

Notes

.