1

Introduction to switching systems

1.1 Purpose of the book

The world-wide public telephone system is remarkable by virtue of the wide variety of equipment used in its construction from mechanical devices installed in the late 1920s to the most modern miniaturised digital circuitry. It all has to work together to provide an economic and reliable service.

Although it is the most visible switching system, the telephone system is not the only one. Other examples of switching systems are the public telex network allowing dial-up communication between teleprinters, and a range of national and international networks for company, bank and military users. Of increasing importance during the 1980s will be the growth of data switching which many people predict will overtake voice communication (in volume) by the end of the century. There are other systems which at first sight may not be regarded as switching systems but whose design, as this book shows, is based on the same principles. These include telemetry systems, on-line access to central computers from widely dispersed devices such as visual display terminals, card readers and so on.

The purpose of a telecommunications switching system is to provide the means to pass information from any terminal device to any other terminal device selected by the originator. Three components are necessary for such systems:

terminals which are either input or output transducers. They convert the information into an electrical signal at the transmitting end and convert the electrical signal back into a usable form at the receiving end. A further function of a terminal is to generate and transmit control signals to indicate the required destination of the information signal.

transmission links to convey the information and control signals between the terminals and switching centres.

switching centres to receive the control signals and to forward or connect the information signals.

Transmission links are covered in detail in a companion volume [1] and receive only scant mention in this book. Terminals are covered only from the point of view of their capabilities to generate and receive the control signals. The book deals with the design of the individual switching centres and their incorporation in switching networks.

The organisation of this book has two aims:

(1) To show that there is a unified set of principles behind the wide range of superficially different switching centre designs, and to show how these principles may be applied to switching centre design using modern technology.
(2) To give a description of the implementation of the design principles in some of the switching centre types in use today.

This chapter introduces the basic systems concepts and the objectives of a system design. Chapter 2 describes the basic signalling and switching technique used in voice and data switching centres. The design of economic switching centres is covered in Chapter 3. Economic design involves resource sharing and resource sharing implies that there is a probability that a resource (such as a switch or a transmission link) will not be available at the instant it is required. The design of a switching centre therefore involves determining the number of resources required to achieve a particular probability of no resource being available (and possibly a particular length of wait for a resource to become available). This is the subject of traffic theory which is covered in Chapter 4.

Switching centres are organised in networks and these are discussed in Chapter 5. The practical means by which control signals are passed from centre to centre are the subject of Chapter 6.

The next two chapters are more theoretical and in Chapter 7 a coherent approach is given to the design of switch networks. Of particular importance in this chapter is a discussion of the work of Takagi on optimum channel graphs. It is thought that this is the first time that this treatment has been covered in a text book.

Chapter 8 is a theoretical approach to the design of control systems and attempts to show a unified approach to the design of electronic and computer controlled systems and the relationship to signalling systems.

After the theoretical chapters, the remainder of the book deals with practical aspects of telephone switching systems. Chapter 9 describes a wide range of the practical techniques used within electro-mechanical and electronic systems. A selection of electro-mechanical and electronic systems are described in Chapter 10 and the examples are carefully chosen to demonstrate the application of particular principles. Computer controlled systems have Chapter 11 to themselves and some of this chapter requires a basic knowledge of computers.

The long-term future for the technology of switching systems is almost certainly going to be digital and the basic differences to other systems together with an example are discussed in Chapter 12.

Finally, the author has allowed himself some licence and the last chapter is in the form of an editorial, which attempts to predict the future direction of system architectures. This chapter attempts to show that the centralised control systems origins come from the 1960s and are not the way today's systems should be designed. The centralised systems appear to have been designed as an exercise in programming rather than as the design of a reliable telephone switching system using principles described in this book. Time will show who was right.

1.2 Types of switching systems

A system similar to (but smaller than) the telephone network is the public telex network (*tel*egraph *ex*change) which provides world-wide direct interconnection of teleprinters. Both the telephone and telex networks are examples of what are called *circuit switching* since they set up a circuit between two terminals which then interchange information directly.

Another class of system which is more familiar to the business or military user is that of *message switching*. The terminals of message switched systems are usually teleprinters, but unlike the telex network they are not interconnected directly. Instead, when a terminal user types a message destined for some other terminal, the system stores the message and delivers it to the required terminal at some later time. The reason for the delay is that the system is designed to maximise the utilisation of transmission links by queueing messages awaiting the use of a link. In order to set up a direct connection over many links connected end-to-end it is necessary for each link to be simultaneously free. As will be seen later, this implies that the average utilisation of the links must be low if the probability of a direct connection being available on demand is to be high enough to satisfy most users. However, in a message switched system, where messages are queued for each link, a much higher link utilisation is achieved. Another name for this type of system is *store and forward* switching.

The advent of real-time computer systems for airline reservation, banking systems and remote data processing in general has been based on the use of telecommunications networks to carry data between computer-type terminal devices and large real-time computers. Such applications may be served by a purpose-built circuit switched network or by another system such as the public telephone network in conjunction with special signal processing techniques.

More recently, general purpose *packet switching systems* have been developed which take the data from a terminal or a computer and transmit it as short packets of information to the required destination. Such systems are midway between the two extremes of circuit switching and store and forward. The interchange of packets may be made so rapid that a terminal appears to provide a 'conversational' connection while at the same time high transmission link utilisations are obtained through queueing.

1.3 Centralised switching systems

A simple way of structuring a switched network is to arrange that each terminal has a direct transmission link to every other terminal, as shown in Figure 1.1a. Each terminal needs a switch to connect it to the required link and a switch to make connection to a link in order to receive an incoming call. For N terminals this arrangement needs a total of $\frac{1}{2}N(N-1)$ links.

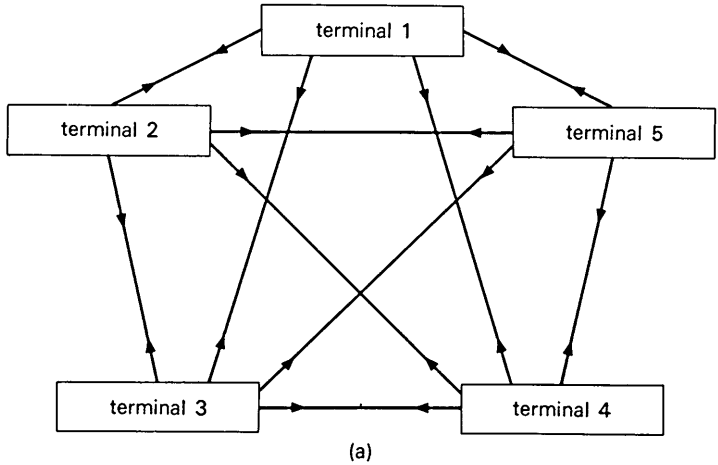
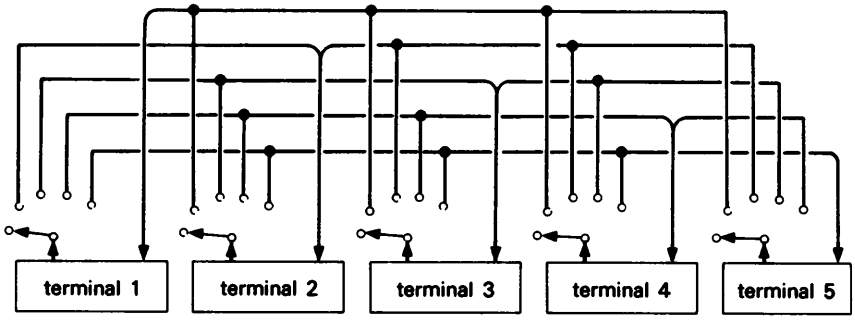
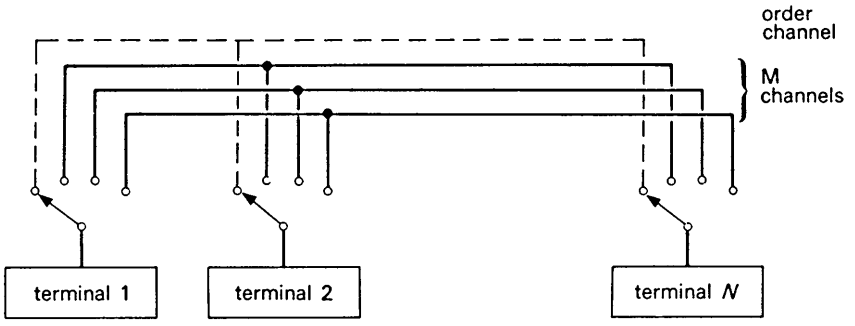


Figure 1.1 (a) Full interconnection for five terminals. Some form of switching is needed within each terminal to connect it to the appropriate link. (b) Use of one channel per terminal. Each terminal is permanently connected to one channel and all other terminals may access a particular terminal by operating a switch which connects it to the appropriate channel. (c) As in (b) but with less than N channels.

An alternative approach which needs only N links is to provide one link per terminal and to arrange that all other terminals have access to it as shown in Figure 1.1b. This simplifies the terminal equipment because it removes the need to connect a terminal to a link for an incoming call. This is a practical arrangement for systems such as house telephones or intercoms where there is a relatively small number of terminals close together. For instance, it is possible to provide an eleven-terminal system with each terminal having ten buttons and eleven pairs of wires going around to each instrument. However, when the number of terminals increases, or their geographical separation increases, the cost of cabling makes this arrangement uneconomic. Another example of the technique illustrated in Figure 1.1b is a radio telephone network in which each terminal is given its own frequency. An originator can set-up a call by tuning his transmitter to the frequency of the called terminal.



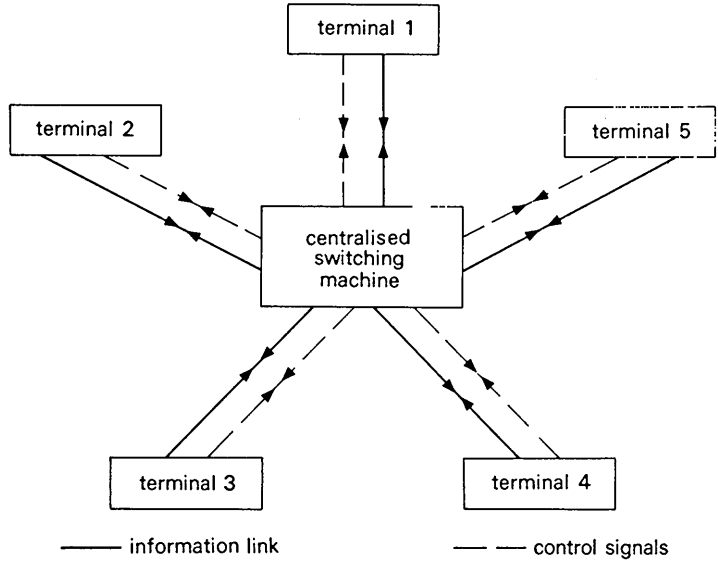
(b)



(c)

It is possible to use an arrangement similar to that just described but with fewer than N links, as shown in Figure 1.1c. With this arrangement it is necessary for the calling terminal to select a free link and for the called terminal to be connected to the same link. One technique used extensively for radio telephones involves a common channel to which all free terminals are connected [2]. A calling signal (which may be coded or verbal) is sent along this link and is received by the terminals. The calling signal informs the called terminal (or its user) of an incoming call and indicates the link to which the terminal should be switched in order to receive the call. The recognition of the calling signal and the switching operations may be performed automatically in a system using coded signals. In simpler systems the users may be required to listen out for, and act upon, verbal calling signals.

At the present time this technique of 'distributed' switching is applied only to small telephone systems and to some radio telephone networks. One of the possibilities of the future is what is called a 'digital ring main' in which a high speed binary digital link is connected to a large number of terminals. Similar techniques to that described above may then be used. For example, a terminal could connect itself to a free time slot within the digital bit stream in order to set up a communication path to the called terminal.



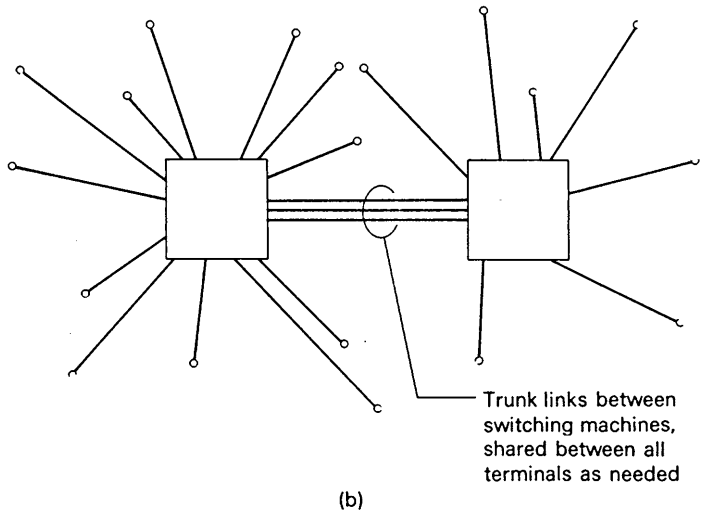


Figure 1.2 The use of centralised switching machines. (a) Single machine which reduces average length of transmission link as compared to Figure 1.1. (b) Use of additional switching machine to reduce transmission costs further (if the terminals have a low utilisation).

Centralised switching centres. In most practical switched networks it is usually more economic to provide a link between each terminal and a central location and to perform all switching operations there, as illustrated in Figure 1.2a. This arrangement significantly reduces the total transmission costs in the network. However, the switching centre must be operated by remote control from the terminal and this tends to increase the total switching costs.

The total transmission costs may be reduced even further if a number of local switching centres are used instead of one national centre because this reduces the average length of the connection between a terminal and its nearest switching centre (Figure 1.2b). The local centres must be interconnected by transmission links which are usually called *trunks*. These trunks are shared by all the terminals connected to each centre, and as will be shown later, the number of trunks connected to a local centre can be very much smaller than the number of terminals.

The use of multiplexing techniques for long distance transmission makes cost per unit distance for a trunk less than the cost per unit distance for the link between a terminal and its local switching centre. Therefore increasing the number of switching centres lowers the total transmission costs. However, as the number of centres is increased the total switching costs tend to increase for two reasons. First, the local centres become more complex because they must be able to decide on a suitable routing to another centre and because the centres involved in a call must be able to exchange information. Secondly, economy of scale is lost with an increased number of local centres because two half-size centres, plus their buildings and power supplies, cost more than one full-size centre. In general there is an optimum number of local centres for minimum total cost of transmission and switching. This optimum number of local centres depends upon the relative costs of switching equipment and transmission equipment and the geographical distribution of terminals.

If certain assumptions are made about the geographical distribution of terminals, their traffic characteristics, and the costs of switching and transmission, it is possible to make mathematical analyses of the minimum cost for a total system [3] (see Appendix A). However, these analyses have limited value because the practical details of a given situation invalidate the generalised assumptions. For instance, in the telephone field, detailed costings of possible network plans are needed in order to decide how many switching centres should be installed to cover a particular area, or, more usually, to decide whether it is better to install a new switching centre in the suburbs of a growing town, or to extend the main centre [4]. These costings are simplified by using detailed computer models of the area under consideration to facilitate rapid estimation of the costs of alternative arrangements.

Hierarchical systems. As the number of separate switching centres increases the number of different trunk routes between them increases. Above about ten centres the number of trunk routes becomes very large and routes tend to contain too

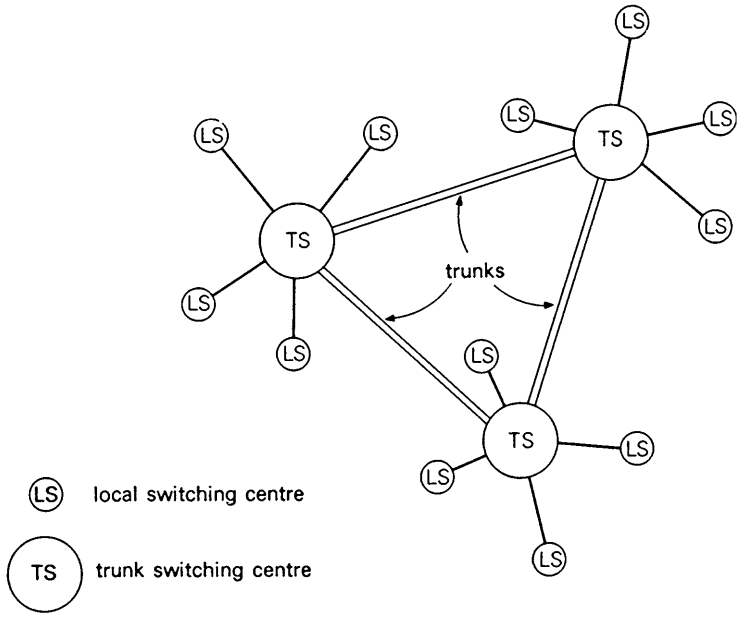


Figure 1.3 Mesh of trunk routes.

few circuits to make the network economic. The same argument that led to centralised local switching now applies to the trunk routes, so trunk switching centres are introduced to switch between trunks. Figure 1.3 shows how the trunk centres reduce the total length of trunk circuits. However, this arrangement introduces the additional cost of trunk switching centres and also necessitates three switching points (rather than one or two) on some connections.

The process of centralising switching centres can occur at several levels leading to what is called a *hierarchical network*. This is best explained by reference to the public telephone network. In a country there is normally a number of local switching centres, each serving anything from 20 to 10 000 terminals (or even up to 100 000 in special cases). These local centres are gathered into groups and each group is served by a trunk centre which can connect calls between local centres within the group. For calls between terminals of one group of local centres, and those of another, the trunk centres themselves have to be interconnected by 'super' trunk centres each covering an area consisting of many local groups. Usually there is a number of these wider areas and these too will be interconnected by higher level trunk centres. This principle is extended to the level where the number of trunk centres at that level is small enough for it to be practical to interconnect them fully with a mesh of trunk routes similar to those shown in Figure 1.3. The result is as shown in Figure 1.4a.

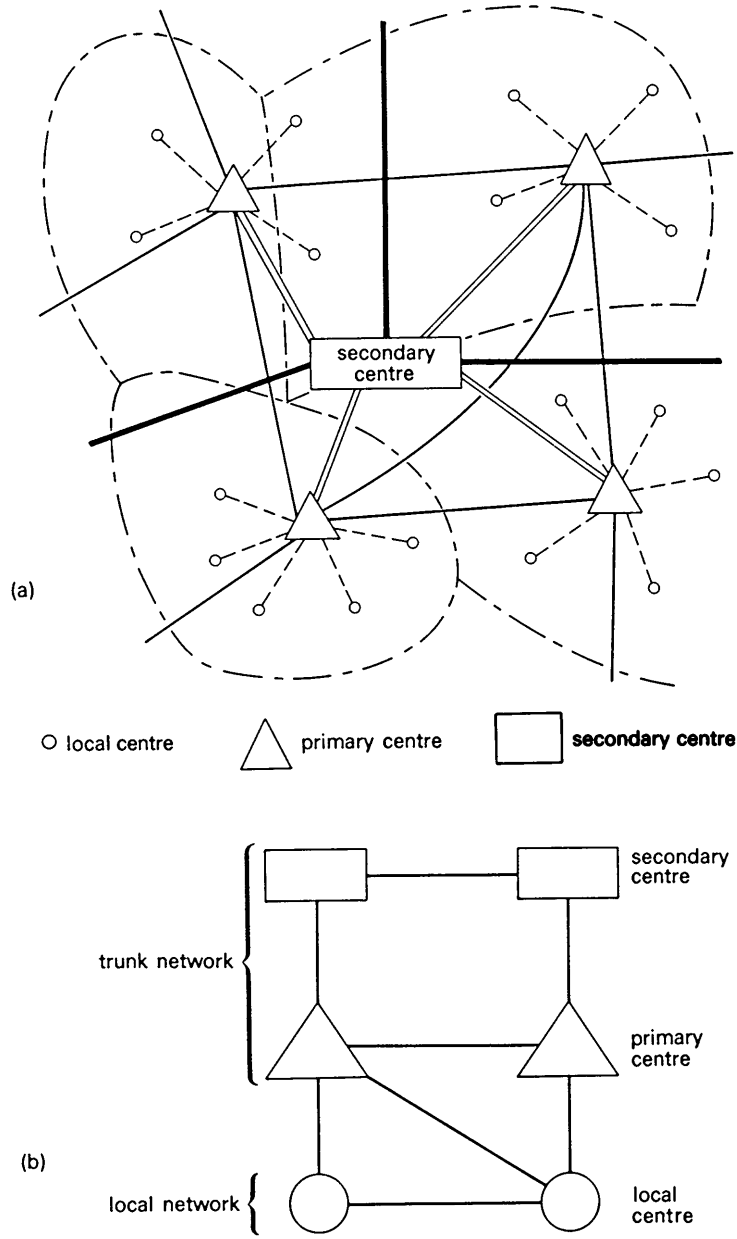


Figure 1.4 (a) Hierarchial network of switching centres. (b) Conventional representation.

Table 1.1 Names of different level switching centres in telephone networks (The numbers in brackets indicate approximate numbers in each country (1976).)

Function	*International name*	*U.K. name*	*North American name*
Local switching centre directly connected to terminals	Local Exchange	Local exchange (6200)	End office or Class 5 office (18 000)
First level trunk switching centre	Primary centre	Group switching centre (370)	Toll centre or Class 4 office (1500)
Second level trunk switching centre	Secondary centre	District switching centre (27)	Primary centre or Class 3 office (250)
Third level trunk switching centre	Tertiary centre	Main switching centre (9)	Sectional centre of Class 2 office (65)
Fourth level trunk switching centre	Quaternary centre	(not required in U.K.)	Regional centre or Class 1 office (18 in U.S.A., 2 in Canada)
Centre which provides connections only between local centres (i.e. no access to trunk network)	Tandem exchange	Tandem exchange	Tandem office
Switching centre which provides access to international network.		Centre du transit 3 (or CT3)	
International transit centre		Centre du transit 2 (or CT2)	
Top level international transit switching centre (fully interconnected)		Centre du transit 1 (or CT 1) (there are 7 of these in the international network)	

In the telephone system, different countries have adopted different names for these different levels of centre. So to simplify discussion a set of names has been internationally agreed [5]. Table 1.1 shows these names together with equivalent U.K. and North American terms.

Figure 1.4b shows a conventional way of illustrating a hierarchical network plan. It can be seen from this that a hierarchical network, as described above,

guarantees that a connection between any two terminals will be possible. Also it sets a limit to the number of links required in the worst case. In practice the situation is more complex. Large numbers of direct routes are provided between switches if the traffic level is justified. So, for example, there may be direct routes between a tertiary centre in one area and a secondary centre in an area served by a different tertiary centre.

Functions of telephone switching systems. Some of the functions of telephone switching will now be defined. The local switching centre must react to a calling signal from a terminal and must be able to receive information to identify the required destination terminal. It must be able to decide from the input information whether the required terminal is connected to the same local centre or whether a trunk connection is necessary via one or more intermediate trunk centres. If an intermediate trunk centre is needed the local centre must find a free trunk on the required trunk routes and connect the terminal to it. Further information must then be forwarded to the intermediate trunk centre or centres to progress the call to its destination.

Once a path has been set up from the originating centre to the terminating centre, the called terminal must be rung; and once the called party has answered, a speech path must be established between the two terminals for as long as the call lasts.

Since public telephone systems must make money, at some stage it is necessary to extract charging information for billing purposes.

A further requirement which is not obvious from what has been said so far is that a telephone system must be very reliable. In the language of reliability mathematics, a telephone system must have a *high availability*. Most switching systems are required to give uninterrupted service for many years. Telephone systems, for example, have design lives of from 20 to 40 years. Present technology is such that no system can be guaranteed to be completely free of faults for this length of time, but it is nevertheless possible to design a system to provide an adequate service even in the presence of faults or malfunctions.

System reliability can be expressed mathematically in terms of *availability* defined as:

$$A = \frac{\text{up-time}}{\text{up-time + down-time}}$$

where the up-time is the total time that the system is operating satisfactorily and the down-time the total time that it is not.

An alternative and equivalent definition of availability is in terms of the mean time between failures (m.t.b.f.) and the mean time to repair (m.t.t.r.):

$$A = \frac{\text{m.t.b.f.}}{\text{m.t.b.f. + m.t.t.r.}}$$

Table 1.2 Some typical availability objectives for public telephone systems

For faults causing complete loss of service for more than 3 minutes and

affecting only a single terminal	– m.t.b.f. $\geqslant$ 10 years
affecting 10% of terminals	– m.t.b.f. $\geqslant$ 20 years
affecting complete switching centre	– m.t.b.f. $\geqslant$ 50 years

Total down-time (due to switching centre failures) $\leqslant$ 2 hours in 40 years i.e. overall availability $\geqslant$ 99·9994%

Availability objectives are difficult to define. A fault which upsets the service of only one or a small number of users is less troublesome than a fault which makes a complete switching centre inoperative. The period of time that service is denied is also important. For telephone users, a break in service of only a few seconds (if it occurs only rarely) is not too troublesome, but breaks of fifteen minutes or more are very troublesome. Hence the availability figures must be split into a number of different categories. Some examples are shown in Table 1.2.

It should be noted that the origin of these objectives is largely historical; early systems (step-by-step or Strowger switching systems to be described later) actually achieved availabilities of this order.

1.4 Basic switching centre model

Most of the fundamental principles of switching system design can be understood by considering a single centralised switching centre like that of Figure 1.2a. In addition to a channel for transfer of information, there is also a two-way path between each terminal and the centre for interchange of the system control signals. In fact, in most of the existing systems the same physical channel is used for both purposes. One of the practical problems of system design is the separation of the control signals from the information at the switching centre.

The general realisation of this centralised machine (for circuit switching) is shown in Figure 1.5 where each terminal has its own switch and control signal path to its own control unit. Each switch has outlets to each of the other terminals. Also each control system has access to each of the other control units; this is necessary so that the control unit associated with a calling terminal may test whether the control unit associated with a called terminal is free and, if it is free, busy it.

Signal exchange diagrams. The first step in any system design is to consider the range of control signals that has to be interchanged between a terminal and the system. This information is conveyed in the form of *signals* and many different ways are used to code these signals. In telephone systems a commonly used form of signal is the changing of the value of an analogue quantity, such as

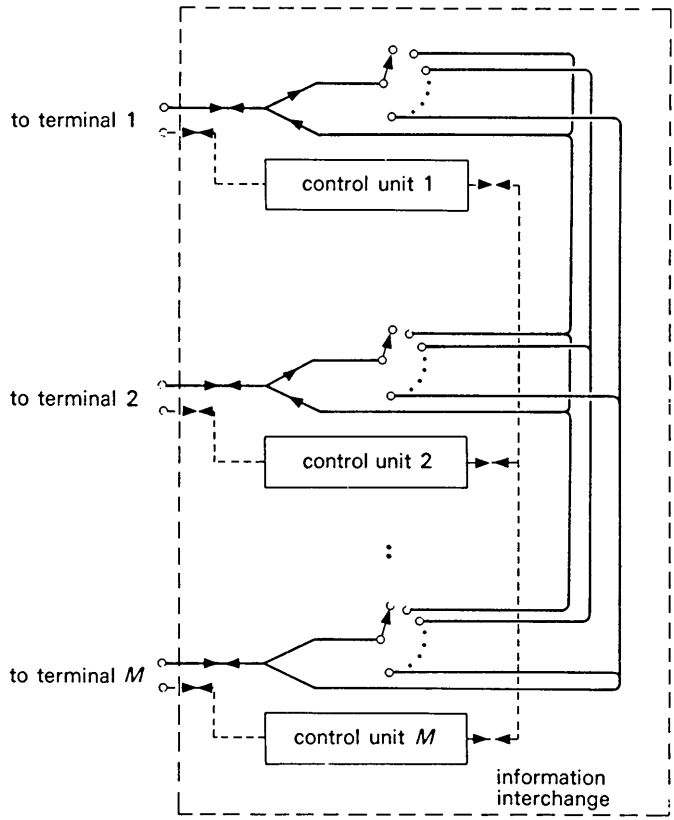


Figure 1.5 Simple model of a centralised switching machine.

the d.c. resistance across a pair of wires, the amplitude of a sinusoidal voltage etc. In message and data systems, special bit patterns are used as signals. Specific techniques are discussed later.

Whatever the techniques used, the basic signal types remain the same. Figure 1.6 (called a signal exchange diagram) shows a basic set of signals for a terminal that can be used in either a calling or a called mode. In the calling mode the first action that the user of the terminal must take is to transmit a *seize* signal to the system to indicate that the terminal wishes to make a connection or pass a message. The system generally responds to the *seize* by an *accept* signal (for reasons which will become apparent later). The terminal then transmits *routing* signals and the system responds with a variety of *status* signals, for example:

- *line busy*
- *line free*
- *number invalid*
- *line answered*

Once the need for the connection is over, the terminal sends a *clear forward* signal. 'Forward' and 'back' here refer to the direction of traffic which is deemed

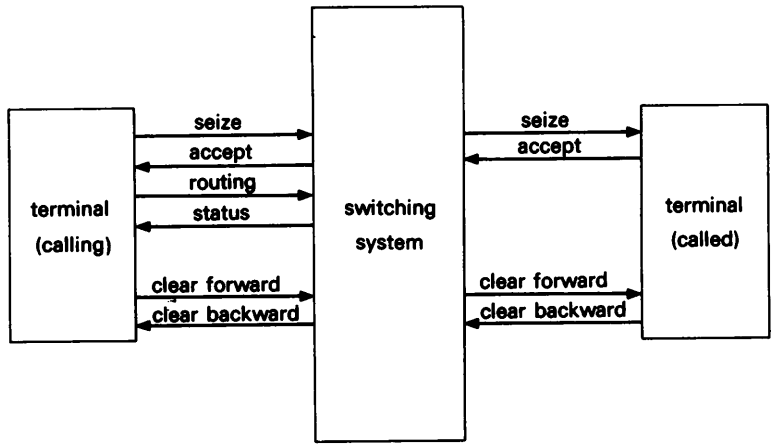


Figure 1.6 Basic signal exchange diagram.

to flow from the originating terminal to the terminating terminal.

In the called mode the terminal is sent a *seize* signal from the system and responds with an *accept* signal. At the end of the period of communication the terminal may be sent a *clear forward* signal to indicate the end of the connection.

In some systems a *clear back* signal is sent from the called terminal to the system and possibly also from the system to the calling terminal if the user of the called terminal is the first to indicate that the communication is finished. For example, the signal sent from a telephone terminal to the local centre when the handset is replaced at the end of a call is regarded as a *clear forward* signal if the user initiated the call, but a *clear back* signal if the user was the recipient of the call, even though the signals may take the same electrical form in practice.

State transition diagrams. The signal diagram gives the 'alphabet' of the valid signals between two devices. However, it does not indicate what sequences (or 'sentences') are possible or what they mean. The valid sequences and their meaning can be expressed conveniently in what is called a *state transition diagram* (s.t.d.). This is such a useful specification and design tool that a set of international standards have been produced [6] for use in telecommunications systems.

The concepts underlying an s.t.d. can be explained by reference to the control units in Figure 1.5. These control units may be thought of as existing in a number of stable states such as:

- *idle*
- *waiting for routing information*
- *waiting for answer*

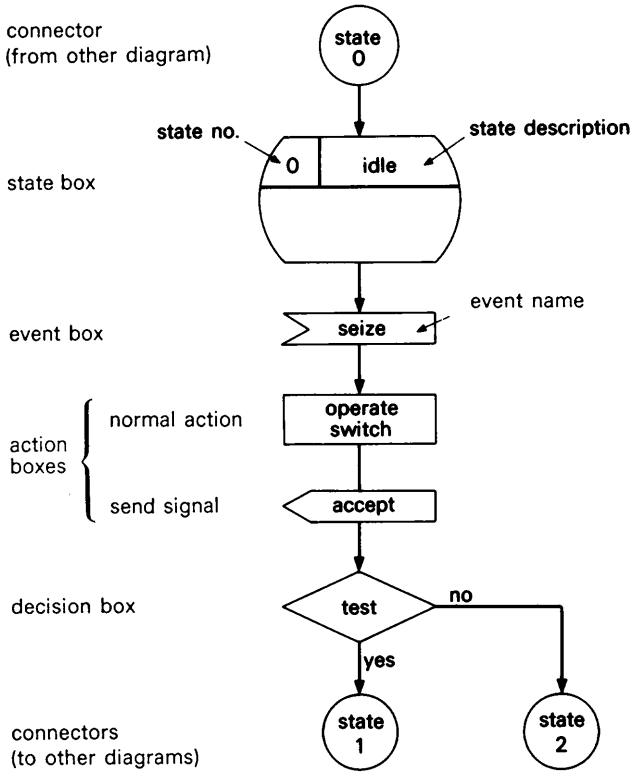


Figure 1.7 Example of state transition diagram symbols. (a) Basic symbols. (b) Multiple decision box.

In general a control unit moves from one state to another only because of the arrival of a signal from a terminal or another control unit. The arrival of a signal is described as an *event*. When the unit does move from one state to another it may perform some action such as operating a switch or sending a signal to

another control unit or terminal. The actions performed between one stable state and another are collectively called a *task*.

In many cases the combination of current state and new event defines the task and new state, but in some cases there is a choice of next states. Such a choice depends upon information external to an individual control unit. The most obvious example of this is the choice of next state after the *routing* signal has been received. A different state is necessary depending upon whether the called terminal is busy or free.

Thus there are four components of a state transition diagram, the symbols for them are shown in Figure 1.7:

(a) *State boxes.* These are labelled with a number and descriptive title. In some cases it is useful to place a pictorial representation of the state in box.
(b) *Event boxes.* The possible (or permitted) events are each drawn as arrow-indented boxes, the paths to which come from the state box in question.
(c) *Action boxes.* Actions are shown by rectangular boxes except for the action of sending a signal to another control unit or terminal which is given the special symbol of an arrowed box.
(d) *Decision boxes.* For binary decisions the symbol is the diamond-shaped box as normally used in computer flowcharting. For multiple decisions the extended version shown is used.

S.t.d. for the calling states of a simple control system. Figure 1.8 shows a simplified s.t.d. for the states of a control unit involved on an originating call. In the *idle* state the only valid event is *seize*. This is acknowledged by an *accept* signal and the control system moves into state 1, *waiting for routing*. The event of reception of *routing* leads to a test of the state of the required terminal. (The mechanisms for performing this test are discussed later.) If the called terminal is free, the calling terminal is informed, a path is set-up, and the unit moves into state 2, *waiting for answer* The *accept* signal from the called terminal causes an *answer* signal to be sent to the calling terminal, and the control unit moves to state 3, the *talking* state.

In this simple example it is assumed that the connection is under the control of only the calling terminal. This arrangement is referred to as calling party release. Any *clear back* signals from the called terminal are therefore irrelevant. There are three other possible call clear down techniques:

Called party release;
First party release;
Last party release.

In states 1 and 2, it is possible for the user of the calling terminal to abandon the call. If this happens a *clear forward* signal is sent to the system. The s.t.d. in Figure 1.8 shows the action to be taken in these cases.

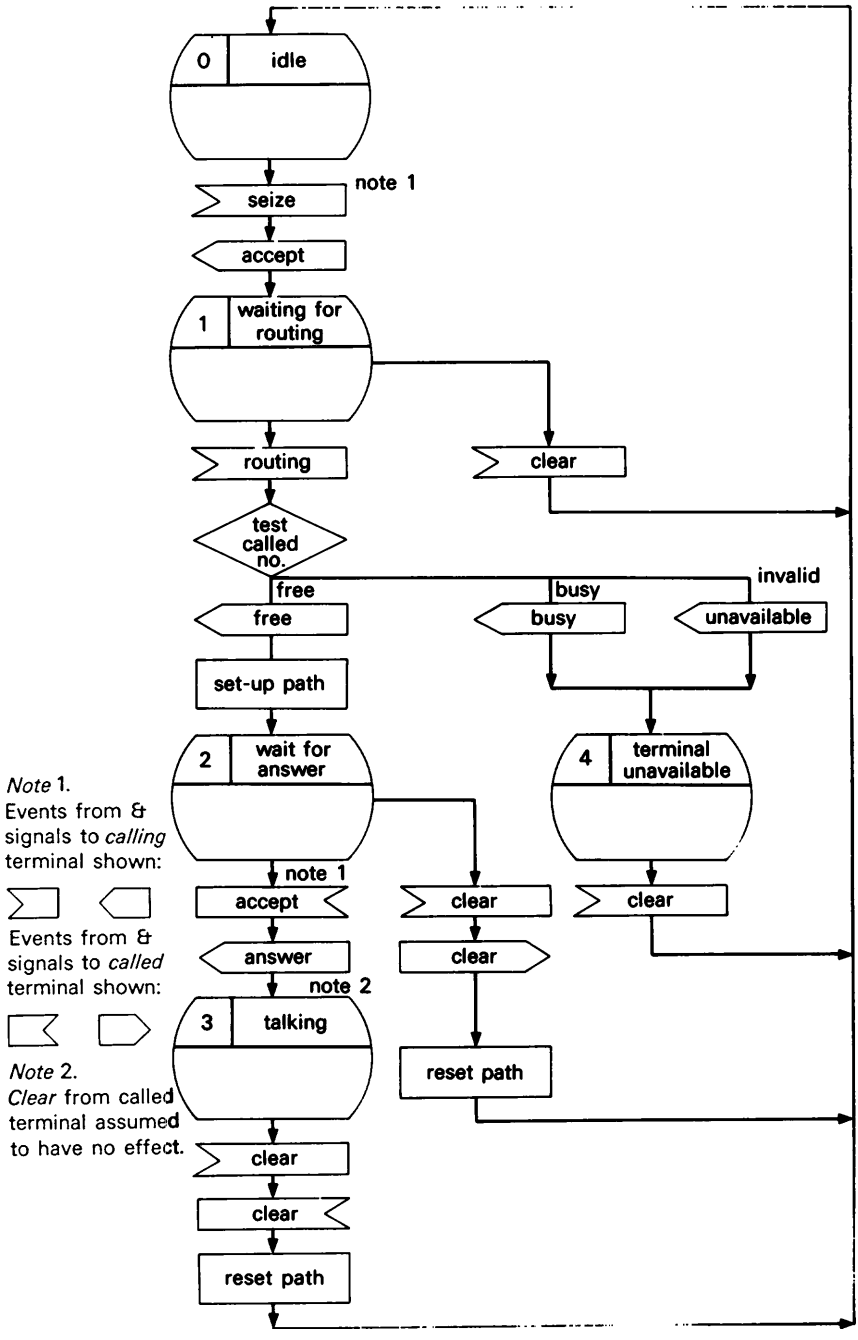


Figure 1.8 Simplified s.t.d. for originating calls only.

The complete s.t.d. for originating and terminating calls are described in Chapter 8.

1.5 Resource sharing

A system like the Basic Switching System described above has two inherent advantages:

(a) It can be made with very high system availability, because if it is properly designed, any fault in a control unit or its associated switch should affect at most, only two terminals, that is its own and any one which happens to be connected to it.
(b) In the absence of faults, the only possible reason for an on-demand connection to another terminal not being achieved is that the required terminal is already busy.

With present day technology, however, a system like this is generally uneconomic. The art of switching systems' designers over the last century has been to find techniques to reduce the cost of a switching system, while maintaining high levels of system availability. Practical system designs take advantage of the

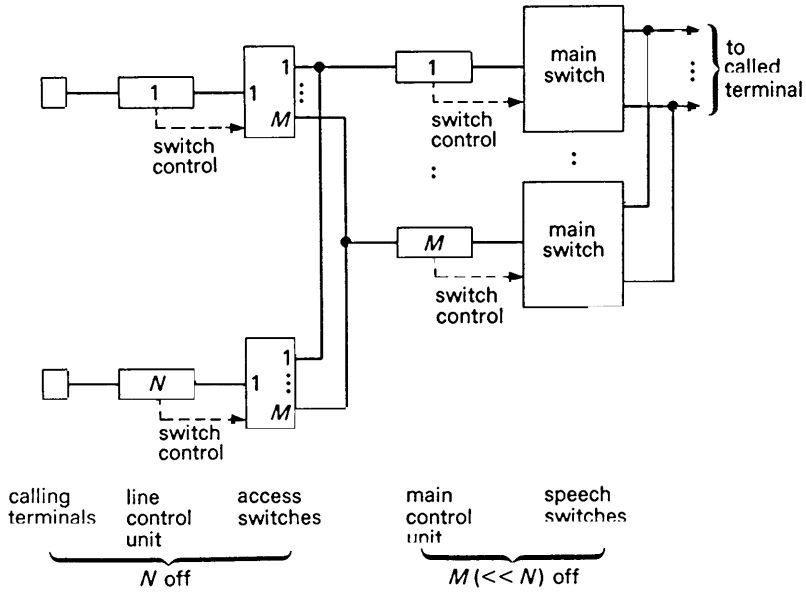


Figure 1.9 Example of functional subdivision. *Note:* lines between blocks carry both control signals and information.

fact that not all parts of a complete system are required all the time, either because terminals are not in use or because they can operate much faster than needed. These factors permit the possibility of *resource sharing* as a means of cost reduction. In switching systems there are three techniques of resource sharing.

(a) Functional subdivision. It is not necessary for all the functions of a control unit to be available all the time. For instance, while it is in the *idle* state a control system for one terminal need only have the function of detecting a *seize* from its terminal or from another terminal that is its calling terminal. Also, when a terminal is in the *conversation* state its control system does not need the functions involved in receiving and decoding the *routing* signals. The technique of functional subdivision involves partitioning a control unit into a number of units, each of which provides only some of the functions of the complete control system. Each terminal needs a permanently associated unit to detect *seize* signals, but other units may be pooled and switching arrangements introduced so that they are associated with a terminal only while they are needed. This technique is shown in Figure 1.9 where the control unit has been partitioned into a line control unit and a main control unit. There is one line control unit per terminal and these have access to a lesser number of main control units.

This technique can reduce the total amount of control equipment, but it introduces the need for extra switching and its control. It also introduces the possibility of a terminal having to wait for service if a common control unit is not immediately available.

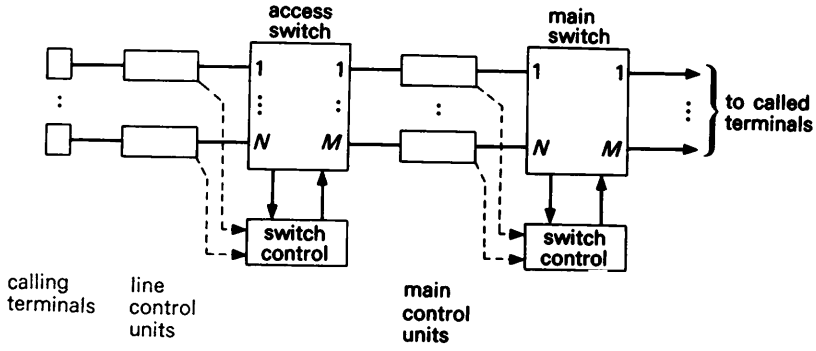


Figure 1.10 Use of common switch networks and controls.

(b) Common switch network and control. The second technique of resource sharing is replacing a number of one-by-*N* switches by a single multiple-input, multiple-output network. Figure 1.10 shows this applied to the system of Figure

1.9, where the M one-by-N main switches are replaced by a single M-by-N network, and the N one-by-M access switches are replaced by an N-by-M network. It will be shown in Chapter 7 that this considerably reduces the total number of switch elements required in a system. There are two problems with this technique:

(i) In most designs of switch network there is a particular probability that no path will be found between an inlet and outlet, even though the terminals themselves are free.
(ii) It is usually necessary to have a common control for the switch network, and if this goes wrong it will affect the complete system.

(c) Time-Shared Decision Making. Each control system exists in a number of states and changes its state only when some event occurs. When an event occurs its control system must decide on the necessary actions and the new state. The third method of resource sharing is possible because the time between occurrence of events at a control system is much greater than the time required for making decisions. For instance, in a telephone system the shortest time between events for a single terminal is typically 30 ms or more, whereas electronic circuits can make a decision in 1 μs or less. The decision-making component may therefore be time-shared between a large number of control units of a given type. Each unit needs a means of storing its current state, buffering the received event, and storing the required actions.

The method by which these techniques of resource sharing are achieved are discussed in Chapter 3.

Objectives of System Design. All practical switching systems achieve their economic objectives by different combinations of the above three techniques of resource sharing. They all introduce the possibility that a resource, such as a control unit, or a path through the network, may not be available when needed by a terminal. When this occurs, the terminal must either abandon its call or wait until the resource that it needs becomes available. A switching system must provide sufficient resources so that the probability of a resource not being available when it is needed, is kept below some design maximum. A significant part of switching system design is therefore concerned with statistics.

It will be seen later that the implementation of these three techniques introduces system availability problems. The Basic Switching System can be the basis of a highly reliable system structure, but when resource sharing is introduced the possibility arises that a single fault may affect more than one terminal and in some cases, the complete system. One aspect of the art of system design is to develop techniques of resource sharing that minimise the effects of the faults that will inevitably occur. Because a switching system must be economic, the problem of system design is ultimately an economic one.

The problem of switching system design may be summarised as follows:

To develop the optimum level and arrangement of resource sharing which satisfies the functional requirements of the systems and keeps below a design objective the probability of a terminal not being able to set-up a call due to

(a) lack of resource when needed, or
(b) faults within a sub-system.

1.6 Introduction to traffic and queueing theory

The introduction of resource sharing into a telecommunications switching system means that sometimes when a call request is made, the resources required will not be immediately available. In these circumstances a call is said to be *blocked*. For some applications it is assumed that, if a call is blocked, the call is abandoned or 'lost'. In other applications the call request is queued in some way until the required resources become available, that is the call 'waits' or 'queues'.

In practice most systems have a mixture of call loss and call queueing. In a telephone system the caller must wait for dial tone but thereafter, if a required resource such as a trunk line is not immediately available, the call is lost and the user receives a busy tone. Usually, even if the resource subsequently becomes available, the busy tone will not be removed.

Traffic theory and queueing theory are used to estimate the probability of occurrence of call blocking and, if queueing is involved, to estimate the statistical distribution of the waiting times of blocked calls. In fact theory is usually used for design rather than analysis, that is to estimate the quantity of resources required in order for a system to meet particular probabilities of loss and queueing.

A theoretical analysis of the performance of a switching system needs two pieces of information as a starting point:

(a) a statistical description of the traffic demands from terminals, and
(b) a set of performance objectives.

Traffic statistics. The traffic demands from terminals may be characterised by the following:

(a) Calling rate. This is the average number of requests for connection that are made per unit time. If the instant in time that a call request arises is a random variable, the calling rate may be stated as the probability that a call request will occur in a certain short interval of time.

(b) Holding time. This may be characterised most simply by the mean time that calls last. In some circumstances (particularly where there is call queueing) the statistical distribution of the holding times is also important. The most commonly used distribution is the negative exponential distribution in which the probability of a call lasting at least t seconds is given by:

$$P(t) = \exp(-t/h)$$

where h is the average holding time in seconds. An illustration of this function is given in Figure 1.11 for $h = 100$ seconds. It shows that with this mean holding time there is:

50% probability that a call lasts longer than 70 seconds;
36% probability that a call lasts longer than 100 seconds;
13% probability that a call lasts longer than 200 seconds.

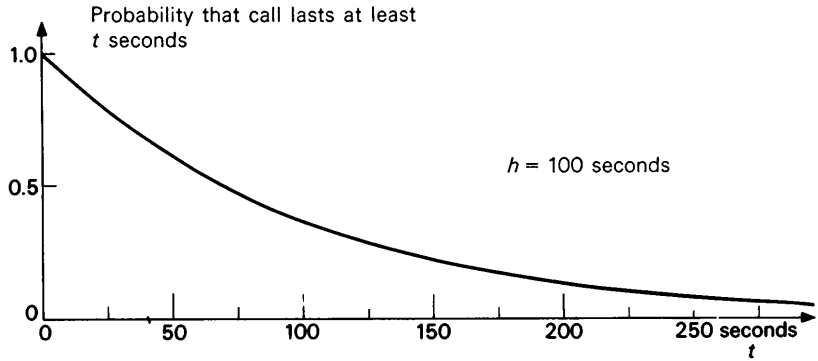


Figure 1.11 The negative exponential function, $\exp(-t/h)$, for h=100s.

This distribution is found to fit a large number of practical cases such as local calls in a telephone network. In many cases the holding time for long distance calls also follows this distribution but with a different mean. This distribution has the advantage that it leads to convenient analytic equations. The main reason for this is that the probability per unit time of a call ending is constant, and therefore independent of the length of time that the call has already been in progress.

In the analysis of control systems it is sometimes more appropriate to assume a constant holding time.

(c) Distribution of destinations. This is described as the probability of a call request being for particular destinations. This is obviously of importance in a hierarchical, many centre system, since it determines the number of trunks needed between individual centres.

(d) User behaviour when no resource is immediately available. The statistical properties of the switching system are a function of the behaviour of users who encounter call blocking. For instance, the system will behave differently if the user

- abandons the request,
- makes repeated attempts to set up the call,
- waits until the resources become available.

Macroscopic traffic statistics: the erlang. As far as the switching elements of a switching machine are concerned, it makes no difference whether a terminal makes ten two-minute calls or one twenty-minute call. It will of course make a difference to the control because to set up ten calls means more work. For this reason studies of switching networks and their interconnecting trunks are normally made on the basis of total occupancy rather than the two parameters of calling rate and holding time. The internationally agreed parameters used are as follows [7]:

The *amount of traffic carried* (by a group of circuits or a group of switches) during any period is the sum of the holding times expressed in hours.

The *traffic flow* (on a group of circuits or a group of switches) equals the amount of traffic carried divided by the duration of the observation, provided that the period of observation and the holding times are expressed in the same units. Traffic flow calculated in this way is expressed in *erlangs*.

Thus the amount of traffic has the units of time and traffic flow is dimensionless.

The erlang is named after A. K. Erlang, a mathematician who developed much of the early theory of telephone traffic. Applying the definition to an individual terminal, if the average number of calls arising in time T is n and the average holding time of calls is h, the amount of traffic carried in time $T = nh$ units of time and the traffic flow $A = nh/T$ erlangs.

In the case of a single terminal the traffic in erlangs is equal to the average occupancy of the terminal, where occupancy is defined as the proportion of the time that a terminal is busy. It can be easily shown that the traffic in erlangs from a group of terminals is numerically equal to both of the following:

(a) The average number of concurrent calls.
(b) The average number of calls which originate during the average holding time.

An alternative unit for traffic measurement in common use in North America is the *hundred call seconds* (ccs). This is used as a measure of the amount of traffic expressed in units of 100 seconds. The number of ccs per hour is also used as a measure of traffic flow. Since 1 erlang may be regarded as an average of one call for one hour, then:

$$1 \text{ erlang} = 36 \text{ ccs h}^{-1}$$

Busy hour traffic. The actual values of all parameters of traffic, and in particular the calling rate, vary with time. In a telephone system the number of

calls made per hour varies throughout the day. For instance, there is a peak in the calling rate in the morning for a telephone system with mainly business users. For a system in a largely residential area the peak may occur in the early evenings. Although the evening peak may be partly due to the operation of a 'cheap rate' tariff, the morning peak is largely due to the habits of business users and is affected only temporarily by the introduction of higher 'peak rate' tariffs. So, cheap rate tariffs may be said to stimulate use of the telephone whereas peak rate tariffs are simply a way of maximising revenue in an acceptable way. For long distance traffic the peaks for different parts of the network may occur at different times due to time-zone differences. Other significant traffic patterns may arise from regular commercial events, for example traffic peaks between fishing ports and London may follow the tides because the fish market follows the landing of the catch.

In addition to changes within the day, the calling rate may vary with the season of the year, for instance it may be higher during the summer season in a holiday resort. There are very large peaks at holiday times such as Christmas, or, in the U.S.A., Mothers' Day. Most switching systems grow with time, in terms of the number of terminals, so the total traffic increases. It is interesting to note that, as a system such as the telephone system grows, there is a tendency for calling rates (per terminal) to decrease during certain stages of this growth, presumably because an increased number of terminals means fewer people sharing each terminal. At other stages, this effect is outweighed by increased use per person, due to the increasing number of easily contractable correspondents.

In order to design a suitable switching system some traffic figures are needed which represent the average demands made by users on the system over a planning period of the order of six months. In the telephone field, the so-called *busy hour* traffic figures are used for planning purposes. An hour is chosen because it is long compared with the average holding time of around 3 minutes. If the parameters are measured for the busiest hour on a number of the busiest days of the year, the mean of these measurements gives a useful measure of the traffic.

Once the statistical properties of the traffic from or to a set of terminals are known, it is necessary to state an objective for the performance of a switching system. This is done by specifying a *grade of service* (g.o.s.). For a system designed on a loss basis, a suitable grade of service is the percentage of calls which are lost because no equipment is available at the instant of the call request. In a waiting system a grade of service objective could be either the percentage of calls which are delayed or the percentage which are delayed more than a certain length of time.

This grade of service is applied to a terminal-to-terminal connection, but in a system containing many switching centres it is usually more convenient to break the objectives down into component parts such as the grades of service for:

- an internal call,
- an outgoing call to the trunk network,

– the trunk network itself,
– a terminating call.

The reasons behind the choice of objectives are not clear cut. They must be a balance between economics and user satisfaction. The economics can be computed, but the user's satisfaction cannot. One objective approach for a commercial system is to find the grade of service for which the cost of adding the extra equipment is equal to the revenue that is being lost by calls being lost [8]. Although easy to state, this computation is difficult to perform and often gives results which, although 'economic', are unacceptable to the user. For instance, if the grade of service for a long distance route is greater than 10% the user will experience considerable annoyance.

Typical objectives for overall grades of service for a commercial telephone system are 3–5% for local calls and (because the cost of equipment provision is higher) up to 6 or 7% for long distance calls. These figures are comparable with the probability that the called user is already busy.

Typical objectives for component parts of a connection are:

Internal Calls 3%	Trunk Calls 1–3%
Outgoing Calls 2%	Incoming Calls 2%

For the above, the overall grade of service is in fact approximately the sum of the component grades of service.

However, this is not a complete specification because it relates only the grades of service for a mean busy hour. In order to ensure that the grade of service does not deteriorate disastrously if the actual busy hour traffic exceeds the mean (or if some of the equipment is out of service), additional grades of service are specified relating to traffic loads of 10% or 20% above the mean, or with certain percentages of the system not operational.

Traffic design objectives. Traffic design objectives for a switching system are generally expressed in terms of a set of traffic flows to be carried by the system with a performance no worse than a set of grades of service for the different types of traffic. In addition, the performance of the system must not deteriorate beyond a certain set of grades of service under specified overload or fault conditions, such as 10% overload or only 90% of the equipment being available because of faults.

The grades of service that are relevant for a message switching system relate to delay. All communications are delayed to a certain extent so a system might have an objective of the form: 99% probability of delivering a message within 24 hours of transmission. This may be too slow for certain types of message, so a range of message priorities is normally used to allow the more urgent messages to 'jump' the queue and be delivered ahead of previously transmitted, lower

priority messages. The existence of these levels of priorities complicates the analysis.

A data system may be either circuit switched or packet switched. In a circuit switched system conventional loss probabilities may be used. The critical factor in a packet switching system is usually response time, for example the time taken for a packet to travel from one terminal to a main computer, for that packet to be processed, and for response to be returned to the originating terminal. If there is a human user, acceptable performance standards are typically a mean response time of 1·5 seconds and 90% of responses obtained within 3 seconds [9].

References

1. Hills, M. T. and Evans, B. G. (1973) *Telecommunications Systems*. Allen & Unwin.
2. Beck, I. H. (1972) Mobile radio systems, *Post Office Elect. Eng. J.*, 64, p. 238.
3. Rapp. Y. (1950) The economic optimum in urban telephone network problems, *Ericsson Technics*, 49.
4. For example, see Back, R. E. G. (1975) Network planning, in *Telecommunications Networks*. J. E. Flood, ed., Peter Peregrinus.
5. *CCITT (1964) National Telephone Networks for Automatic Service*. ITU.
6. *CCITT Orange Book* (1977) Vol. VI, Recommendation Z101.
7. *CCITT Orange Book* (1977) Vol. II, Recommendation E160.
8. Jenson, A. (1950) *Moes Principle. An Economic Investigation Intended as an Aid in Dimensioning and Managing Telephone Plant*. Copenhagen Telephone Company.
9. Martin J. (1972) *Systems Analysis for Data Transmission* (Chapter 7). Prentice Hall.