GRADE OF SERVICE OF A TELEPHONE EXCHANGE

by

Lansun Lee, B.E., M.Eng.

Submitted in partial fulfilment of
the requirements for the degree of
Doctor of Philosophy

Department of Electrical Engineering
Faculty of Pure and Applied Science
University of Ottawa
Ottawa, Ontario
July, 1967

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Approved for the Department
of Electrical Engineering

Supervisor

Chairman of the Examining
Committee

Chairman of the Department

ABSTRACT

This project is concerned with the estimation of the grade of service given to the subscribers in a telephone exchange. The analytical models used are (a) the probability linear graph for lost calls and (b) the traffic flow graph for delayed calls. Both of these models are expressed in terms of the congestion functions of the sub-units of the switching system in the exchange. When these models are used, the engineering procedure of determining the quantity of equipment for the switching system in the exchange to satisfy the objective grade of service is simple.

The switching system in the exchange may be divided into three major sub-units: the trunk groups, the switching network, and the common control. The congestion functions for these sub-units are discussed. The structure and the mode of operation for these sub-units are usually very complex, so that the congestion functions for these units can only be approximated. Thus, the overall grade of service estimated this way is only an approximation.

Finally, the procedure of estimating the overall grade of service of an exchange is demonstrated by considering a simple hypothetical telephone switching system.

The main contributions in this thesis are:

1. The traffic flow graph

   This graph is used as the model for estimating the grade of service for delayed calls in an exchange. It can also be used for approximating the waiting time distribution for a given waiting system.

2. The equivalent trunk technique

   This technique is shown to be reasonably accurate and simple for use in computing the trunk group congestion distribution when the trunk group is arranged in grading.

3. A method of approximating network blocking probability

   This method can be used to approximate the blocking probability of any complex switching network. It is simple and inexpensive and produces reasonably accurate results.

4. Single linking stage network loaded with multitype of traffic

   When some types of traffic require the service of more than one link at a time, the network for these types of traffic is blocked even if there are some idle links in the network. Extremely poor approximation will result if the usual assumption of independent busy links is used for calculating the blocking probability. In this thesis, such a network is analysed and a better approximation for computing the blocking probability is given.

5. General case approach for determining waiting time distribution for a single server delay system.

Such a single server delay system is analysed. This aids one to the understanding of the congestion processes producing delays.

ACKNOWLEDGEMENTS

The writer wishes to express his appreciation for the direction and advice given by Professor J. A. Brzozowski during the course of this research project. He would like to thank Professor G. S. Glinski, Chairman of the Department of Electrical Engineering, for his interest and encouragement. He wishes also to acknowledge the many helpful discussions with members of the Faculty. Special appreciation is given by the writer to his wife, Susie (Shew York), for her patience and encouragement.

The writer is indebted to the Northern Electric Company, Canada, for the financial support through which this thesis was made possible. He is grateful to the Company and to the Electrical Engineering Department, University of Ottawa, for making possible this cooperative research. He is also indebted to K. S. Hoyle, J. C. Kennedy, W. C. Chan, L. E. N. Delbrouck, L. Durr, G. R. Richard and many other members of staff of the Company who offered encouragement, and valuable and helpful suggestions and criticism, and to Miss C. A. Matthews for typing this thesis.

TABLE OF CONTENTS

* Fig. 4 is not used in this thesis.

Chapter I

Introduction

1. Telephone Exchange

A telephone exchange is a central switching point which provides connections among a group of subscriber-lines when signals of requests are placed by the telephone subscribers. The system of equipment capable of recognizing the signals to provide proper connection in a telephone exchange is called the telephone switching system. The connecting devices in such a system are usually called the trunks, links, junctors or service circuits. If there are N subscriber-lines in a telephone exchange, ideally there should be $\frac{1}{2}N(N-1)$ connecting devices in the switching system so that all subscribers can talk simultaneously in pairs when required. However, usually, an exchange serves a large number of subscriber-lines so that this ideal system becomes impracticable because of its enormous size, very high cost, and maintenance difficulties. Therefore, it is necessary to equip the system with a smaller number of connecting devices provided that the subscribers will be confronted with a small possibility that some of their telephone calls may be unsuccessful. This can be done since usually it is only at infrequent intervals that many subscribers make simultaneous calls.

Based on this fact that a smaller number of connecting devices can be provided in a switching system without too much deterioration of the grade of service given to the subscribers in a telephone exchange, a modern switching system in an exchange usually has the form as shown in Fig. 1. It consists of (a) terminals of subscriber-lines,

(b) trunk (junctors and service circuits) groups,

(c) switching network, and

(d) common control equipment.

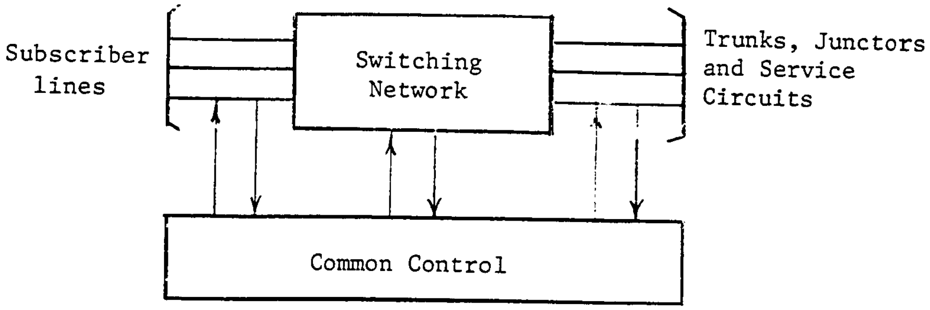


FIG. 1 - A TELEPHONE SWITCHING SYSTEM

The number of trunks (junctors or service circuits) in each group is usually much smaller than the number of possible pairs of subscribers. It is usually engineered in such a way that the trunks (junctors or service circuits) are all busy only for a small portion of time. Similarly, the number of paths in the switching network and the number of common control units are also small. The terminals of subscriber-lines may be connected to the input terminals of the switching network, and the trunks, junctors, and service circuits to the output terminals. The common control has access to the terminals of the lines, trunks, junctors and service circuits, and to the switching network. Of course, the modes of operations for various types of calls (eg. dial tone calls, outgoing calls, incoming calls, intra-office calls, etc.) vary from one switching system to another.

2. Grade of Service

Generally, the grade of service is a measure of inconvenience to the subscribers [1-3]. However, in this research, the grade of service is taken as a measure of the service given in an exchange from the point of view of sufficiency of switching equipment. More specifically, it represents the numerical magnitude of the probability of congestion in an exchange. Thus, the smaller the numerical magnitude, the better the grade of service is given.

The objective or desired grade of service given for an exchange varies for different telephone administrations. For a modern telephone exchange such as shown in Fig. 1, the objective grade of service may be expressed in terms of the probability of delay for dial tone calls and probability of lost calls for overflow conditions. Although the switching system in the exchange is very complex, as an approximation, the system may be considered as composed of isolated parts or sub-units. For simple engineering purposes, each of these sub-units is usually assigned the maximum permissible blocking probability. The overall grade of service of the exchange is then taken as the sum of those of the sub-units. Of course, the overall grade of service of the exchange should be equal or less than the objective grade of service. This way of estimating the overall grade of service for an exchange may be a good approximation for small blocking probabilities but is obviously false in general [1]. It is inaccurate because a telephone call may not necessarily flow through the sub-units consecutively, but rather flows in a manner in accordance with the mode of operation of the system in the exchange (to be described later), and moreover, the dependence of the sub-units has not been taken into account.

In this research, an attempt is made to estimate the overall grade of service for an exchange in a more systematic manner. For simple engineering purposes, the switching system in an exchange may still be considered as composed of sub-units. Thus, the congestion function for each sub-unit is still required to be known. However these sub-units are not considered as isolated parts connected in series but rather related to each other according to the system mode of operations.

A modern switching system as shown in Fig. 1, may be divided into three major sub-units: (a) trunk (junctors, and service circuits) groups,
(b) switching network, and
(c) common control.

In general, these three major sub-units have different congestion functions. It should be noted that each sub-unit may be divided further into smaller units as required.

3. Summary

In this research, the grade of service of an exchange is investigated as follows:

Chapter II outlines the method of estimating the overall grade of service, given the grades of service of the sub-units.

Chapter III reviews some commonly used congestion functions for trunk groups.

Chapter IV studies blocking probability of the switching network.

Chapter V investigates methods of determining the waiting time distribution.

Chapter VI describes a simple hypothetical telephone switching system and shows how the method proposed here can be used for estimating the overall grade of service of the system.

In conclusion, in Chapter VII, suggestions for further research in this area are given.

Chapter II

Grade of Service

1. Introduction

The grade of service is a measure of service given in an exchange from the point of view of sufficiency of switching equipment. If follows that the grade of service is deteriorated if congestion conditions in an exchange are permissible. For a modern switching system in an exchange, this deterioration of grade of service may be due to all or part of the three major sub-units, namely: the trunk (junctors and service circuits) groups, the switching network, and the common control. As examples, consider a switching system in an exchange having the modes of operations for dial tone calls and intra-office calls as follows:

(a) Dial Tone Call

When the common control detects an off-hook signal (dial tone request) on a subscriber-line through the line terminal, it searches for an idle (not busy) dial tone service circuit. If there is no idle dial tone service circuit, the common control may serve requests from other lines and eventually return to serve this line again. If there is an idle dial tone service circuit, the common control sets up a path between them. If all paths between them are busy, the common control may again serve requests from other lines and regard this particular request as a new request signal. When the line is connected to a dial tone service circuit, the subscriber receives a dial tone. This indicates to the subscriber that the exchange is ready to serve him.

(b) Intra-office Call

Assuming that a dial tone has been received, the subscriber dials (or keys) the number assigned to the called subscriber-line. With the accessibility to the dial tone service circuit, the common control can analyse the number to determine the called line. Then, it tests whether the called subscriber-line is busy. If busy, it sends a busy tone to the calling subscriber through the dial tone service circuit. If not busy, it searches for an idle intra-office trunk. If all intra-office trunks are busy, an overflow tone is sent to the calling subscriber through the service circuit and the call is lost. Only if the called line is not busy and there is an idle intra-office trunk, the common control attempts to set up a path between the calling line and one port of the intra-office trunk, and another path between the called line and the other port of the trunk. If this is successful, the common control disconnects itself and the subscribers can proceed with the conversation. If this is unsuccessful, an overflow tone is sent to the calling subscriber through the dial tone service circuit.

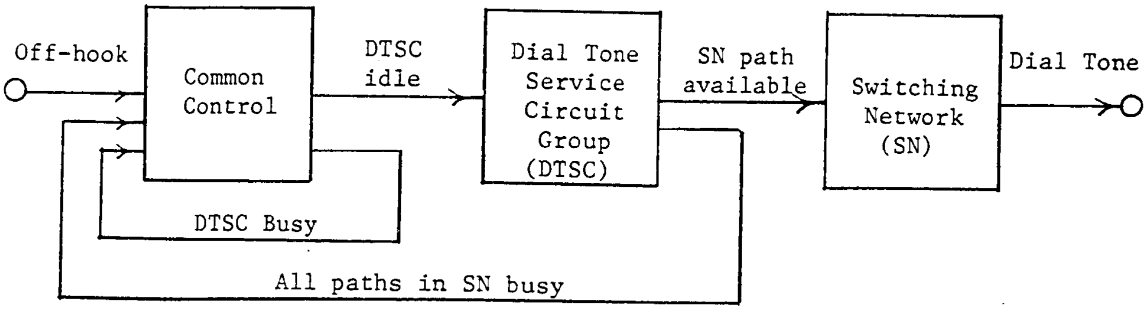


FIG. 2 - A MODE OF OPERATION FOR A DIAL TONE CALL

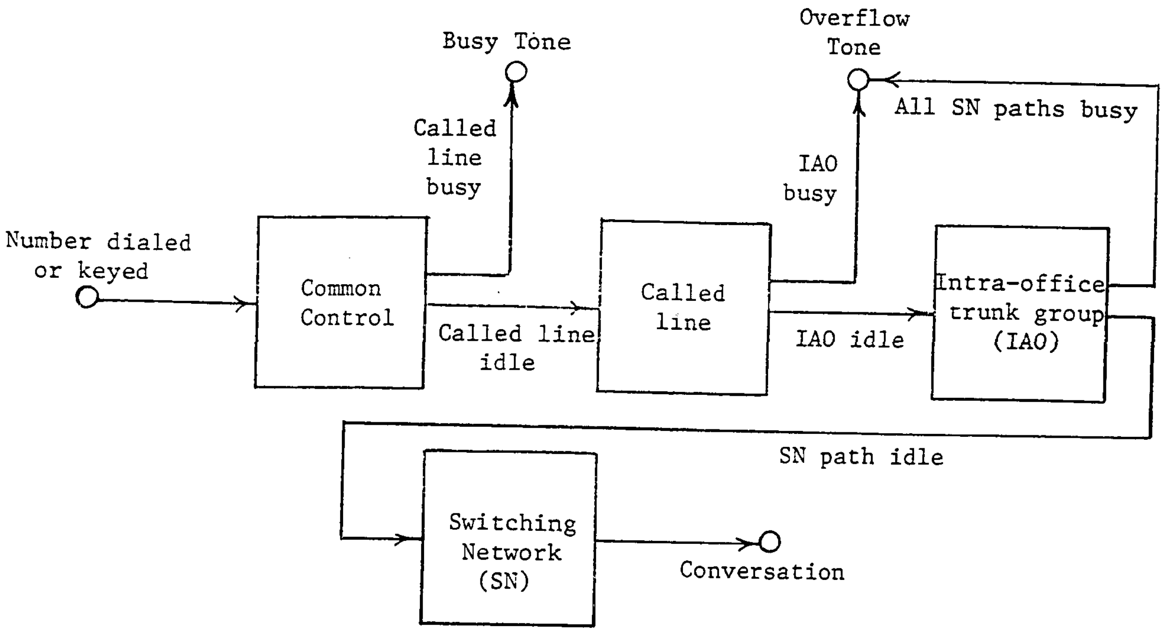


FIG. 3 - A MODE OF OPERATION FOR AN INTRA-OFFICE CALL

These modes of operation for dial tone calls and intra-office calls may be displayed graphically as shown in Figs. 2 and 3 respectively. It is clear that a dial tone call will eventually be served if the subscriber is willing to wait. On the other hand, it is possible for an intra-office call to be unsuccessful even if the subscriber is willing to wait. Thus, an administration may establish an objective grade of service for dial tone calls on a delay basis, whereas that for intra-office calls on a lost-calls basis. For dial tone calls, it is seen that the grade of service may be deteriorated because of insufficient equipment in the common control, a small number of dial tone service circuits, and an insufficient number of paths in the network, all of which cause a delay of dial tone to the subscriber. For intra-office calls, it is also seen that the grade of service may be deteriorated by too small a number of intra-office trunks and by an insufficient number of paths in the network which results in unsuccessful calls.

Therefore, to estimate the overall grade of service of an exchange, two models are required: one for lost calls, and another for delayed calls. In this research, the model for the former is the probability linear graph [4], and for the latter the traffic flowgraph [5].

2. Probability Linear Graph

A probability linear graph is a finite, oriented, connected, cycle-free linear graph with at least two nodes such that:

(a) there exists a pair of nodes, called originating and terminating nodes of the linear graph, this assignment being determined initially,

(b) with each link connecting two nodes of the linear graph is associated a random variable X_i, the subscript i running over all links of the linear graph such that the X_i's are mutually independent and have a probability $P\{X_i = 0\} = P_i$ and a probability $P\{X_i = 1\} = q_i$ where $p_i + q_i = 1$.

Let R be the set of all directed paths between the originating and terminating nodes of a probability linear graph with N links. Then, each directed path $\gamma_j \varepsilon R$ is composed of series links of the linear graph with associated random variables $X_{j_1}, X_{j_2}, \ldots, X_{j_n}$. To each $\gamma_j \varepsilon R$ assign a new random variable Y_j as a function of X_{j_1}, X_{j_2}, ..., X_{j_n} such that $Y_j = 1$ if and only if $X_{j_k} = 1$, $k = 1, 2, \ldots, n$ and otherwise $Y_j = 0$.

Physically, $X_i = 0$ represents that the associated link i of the linear graph is busy (blocked or congested) and $X_i = 1$ idle (linked). Therefore, the events

$$\{Y_j = 1\} = \{X_{j_1} = 1\} \cap \{X_{j_2} = 1\} \cap \ldots \cap \{X_{j_n} = 1\} \tag{2.1}$$

and

$$\{Y_j = 0\} = \{X_{j_1} = 0\} \cup \{X_{j_2} = 0\} \cup \ldots \cup \{X_{j_n} = 0\}, \tag{2.2}$$

represent that the directed path γ_j is linked or blocked respectively, where U and ∩ denote union and intersection, respectively. By definition, the originating and terminating nodes of a probability linear graph are linked if there is at least one idle path between them and otherwise blocked. The probability L that the originating and terminating nodes are linked is called the linking probability; and the probability B that the two nodes are blocked is called the blocking probability. Therefore, it follows that

$$L = P\{\bigcup_{\substack{\text{Associated} \\ \gamma_j \in R}} \{Y_j = 1\}\} \tag{2.3}$$

and

$$B = P\{\bigcap_{\substack{\text{Associated} \\ \gamma_j \in R}} \{Y_j = 0\}\} = 1 - L \tag{2.4}$$

where $P\{\cdot\}$ is the probability of the occurrence of event $\{\cdot\}$.

Either L or B may be used to measure the overall grade of service of an exchange on the lost-calls basis. In order to provide good service to the subscribers, it is necessary for the numerical magnitude of B to be very small or that of L to be very large (almost unity) during the heavy traffic periods of the exchange. In practice, B rather than L is usually used to express the overall grade of service of an exchange.

3. Traffic Flow Graph

A traffic flow graph is defined as a graph consisting of a set of n nodes, D_1, D_2, ..., D_n where $2 \leq n \leq \infty$, interconnected by directed links in some fashion, such that there is at most one link from D_i to D_j and such that

1) Node D_1, called the source, has only outgoing links, node D_n, called the sink, has only incoming links;

2) Each directed link $L_{i,j}$ connecting node D_i to its adjacent node D_j, where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, n$, has associated with it two quantities:

(a) The link delay $W_{i,j}$ which is independent of other link delays and has a distribution giving $P\{W_{i,j} \leq t\}$, the probability that $W_{i,j}$ does not exceed time t;

(b) The link selection binary variable $A_{i,j}$ which is independent of the link delays and takes values from $\{0,1\}$ with $P\{A_{i,j}=1\} = P_{i,j}$ representing the probability that the link $L_{i,j}$ is selected. Also

$$\sum_{j=1}^{n} P_{i,j} = 1; \tag{2.5}$$

otherwise the link selection variables are mutually independent. The link selection probability $P_{i,j}$ for a non-existing link $L_{i,j}$ may be equated to zero.

In a traffic flow graph, the link selection probabilities $P_{i,j}$, and the link delay distributions $P\{W_{i,j} \leq t\}$, are usually assumed known. The problem is to find the distribution of the overall delay time involved in reaching D_n from D_1 through one of the possible paths, where a path is defined as any continuous succession of links traversed in the indicated link directions. A path consisting of k directed links between the source and the sink is called a k-step path.

Between D_1 and D_n in a given traffic flow graph, there may exist a number of paths consisting of k directed links. Let this number be $Z^{(k)}$. Note that $Z^{(1)}$ is, at most, equal to 1. Consider one of $Z^{(k)}$ paths, say the r^{th} path, as shown in Figure 5. In this paths, there are a total of k appearances of directed links, the first being L_{1,r_1}, the 2nd L_{r_1,r_2}, ..., the u^{th} L_{r_{u-1},r_u} ..., and k^{th} (the last) $L_{r_{k-1},n}$ where each subscript r_u takes values from $\{2, 3, \ldots, n-1\}$.

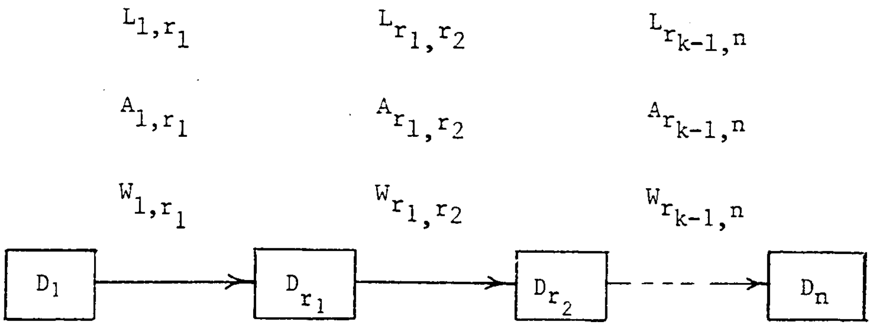


FIG. 5 - THE rth k-STEP PATH

Note that the same directed link may appear in more than one position in the path. An example of this is a path involving feedback loops. However, if the path is selected, it is necessary that the link selection variable associated with the directed link in each of the positions equals unity. In order to distinguish the link selection variables associated with the same directed link L_{r_{u-1},r_u} appearing in more than one position in the path, a superscript is given to each of these link selection variables if ambiguity arises. Thus, $A^{(i)}_{r_{u-1},r_u}$ may designate the link selection variable A_{r_{u-1},r_u} associated with the directed link L_{r_{u-1},r_u} appearing in the i^{th} position in the path. The same superscript, if required, will also be given to the corresponding link delay. The total delay time for this path is, of course,

$$W_{1,r_1} + W_{r_1,r_2} + \ldots + W_{r_{u-1},r_u} + \ldots + W_{r_{k-1},n}.$$

If $\{W_r^{(k)} \leq t\}$ is the event that the total delay time does not exceed t and that the r^{th} one of $Z^{(k)}$ paths is selected, then the probability that the total delay time does not exceed t for this path is

$$P\{W_r^{(k)} \leq t\} = P\{\{(W_{1,r_1} + W_{r_1,r_2} + \ldots + W_{r_{k-1},n}) \leq t\} \cap$$

$$\{A_{1,r_1} = 1\} \cap \{A_{r_1,r_2} = 1\} \cap \ldots \cap \{A_{r_{k-1},n} = 1\}\}$$

$$= \int_{t_n=0}^{t} \int_{t_{r_{k-1}}}^{t-t_n} \int_{t_{r_{k-2}}=0}^{t-t_n-t_{r_{k-1}}} \cdots \int_{t_{r_1}=0}^{t-t_n-\sum_{m=1}^{k-2} t_{r_{k-m}}} \alpha, \quad (2.6)$$

where $\alpha = P_{1,r_1} \cdot dP\{W_{1,r_1} \leq t_{r_1}\} \cdot P_{r_1,r_2} \cdot dP\{W_{r_1,r_2} \leq t_{r_2}\}$

$$\ldots P_{r_{k-1},n} \cdot dP\{W_{r_{k-1},n} \leq t_n\} \quad (2.7)$$

The probability that the overall delay time W does not exceed t when a specific path between D_1 and D_n in a given traffic flow graph is chosen is:

$$P\{W \leq t\} = P\{\bigcup_{k=1}^{\infty} \bigcup_{r=1}^{Z^{(k)}} \{W_r^{(k)} \leq t\}$$

$$= \sum_{k=1}^{\infty} \sum_{r=1}^{Z^{(k)}} P\{W_r^{(k)} \leq t\} \quad (2.8)$$

$1 - P\{W \leq t\}$ rather than $P\{W \leq t\}$ is usually used to express the overall grade of service of an exchange for delayed calls.

To use these models to estimate the overall grade of service given to the subscribers in an exchange, the probabilities associated with the links represent the congestion functions of the sub-units of the switching system in the exchange. Thus, the accuracy of the estimation depends greatly on the accuracy of the congestion functions describing the congestion conditions of the sub-units. Therefore, it is important to investigate these congestion functions in detail. In the next three chapters, the congestion functions for the three major sub-units of a switching system are discussed.

Chapter III

Trunk Group Congestion Distribution

1. Introduction

In a telephone switching system, if there is no blocking anywhere but in the trunk (junctors or service circuits) groups, then the grade of service equals the trunk group blocking probability. Such an idealized system has been studied fairly extensively [1,6,7]. However, except for the full availability trunk group, a large number of problems still remains to be solved due to the complexity of interconnections in the trunking arrangement.

In this chapter, some of the commonly used full availability and grading trunking formulae are first briefly reviewed. Then, a simple method called the equivalent trunk technique for approximating the grade of service of a giving trunking arrangement is introduced.

2. Full Availability

When a trunk group is connected so that all subscribers have access to any trunk in the group, it is called a full availability trunk group. The generalized formula for such a trunk group is [8]

$$B(C,A,j) = \frac{\dfrac{A^C}{C!} + \displaystyle\sum_{x=C+1}^{\infty} \frac{A^x}{C! \displaystyle\prod_{n=1}^{x-C} (C+nj)}}{\displaystyle\sum_{x=0}^{C} \frac{A^x}{x!} + \displaystyle\sum_{x=C+1}^{\infty} \frac{A^x}{C! \displaystyle\prod_{n=1}^{x-C} (C+nj)}} \qquad (3.1)$$

where $B(C,A,j)$ is the probability of a call encountering congestion, which is equivalent to the probability of a call having a delay greater than zero units of time;

C is the number of trunks in the trunk group;

A is the average offered traffic load in erlangs [1];

j = h/H, where

h is the average trunk service time (usually called the average holding time per call); and

H is the average waiting time of a call.

In deriving (3.1), the distribution of the trunk service times is assumed exponential and steady-state conditions are also assumed.

To apply (3.1) for estimating the trunk group blocking probability the factor j must be known besides the traffic A and the number of trunks C in the group. This factor j may be determined from the mode of operations of the system or from direct measurements. The physical interpretation of factor j in (3.1) is that waiting calls are cleared out at a rate j times the rate at which calls are terminated when served by the trunks.

In particular, when calls are cleared out of the group immediately (ie. no call waits), we have $j = \lim_{H \to 0} \frac{h}{H} = \infty$. Then (3.1) is the Erlang B formula,

$$B(C,A,\infty\} = \frac{A^C/_{C!}}{\sum_{x=0}^{C} \frac{A^x}{x!}} \tag{3.2}$$

If calls are cleared out at a rate equal to that with which calls are terminated when served by the trunks, we have j = 1. Then (3.1) becomes,

$$B(C,A,1) = \frac{\frac{A^C}{C!} + \sum_{x=C+1}^{\infty} \frac{A^x}{C! \prod_{n=1}^{x-C} (C+n)}}{\sum_{x=0}^{C} \frac{A^x}{x!} + \sum_{x=C+1}^{\infty} \frac{A^x}{C! \prod_{n=1}^{x-C} (C+n)}}$$

$$= \frac{\sum_{x=C}^{\infty} \frac{A^x}{x!}}{\sum_{x=0}^{\infty} \frac{A^x}{x!}}$$

$$= e^{-A} \sum_{x=C}^{\infty} \frac{A^x}{x!} \qquad (3.3)$$

which is the well-known Poisson formula. When $H \to \infty$ which signifies that calls wait until served, we have j = 0. Then (3.1) becomes

$$B(C,A,0) = \frac{\frac{A^C}{C!} + \sum_{x=C+1}^{\infty} \frac{A^x}{C!\, C^{x-C}}}{\sum_{x=0}^{C} \frac{A^x}{x!} + \sum_{x=C+1}^{\infty} \frac{A^x}{C!\, C^{x-C}}}$$

$$= \frac{\frac{A^C}{C!}\left(1 + \sum_{x=C+1}^{\infty} \frac{A^{x-C}}{C^{x-C}}\right)}{\sum_{x=0}^{C-1} \frac{A^x}{x!} + \frac{A^C}{C!}\left(1 + \sum_{x=C+1}^{\infty} \frac{A^{x-C}}{C^{x-C}}\right)}$$

$$= \frac{\frac{A^C}{C!}\left(1 + \sum_{x=1}^{\infty} \left(\frac{A}{C}\right)^x\right)}{\sum_{x=0}^{C-1} \frac{A^x}{x!} + \frac{A^C}{C!}\left(1 + \sum_{x=1}^{\infty} \left(\frac{A}{C}\right)^x\right)}$$

$$= \frac{\frac{A^C}{C!}\left(\frac{1}{1-A/C}\right)}{\sum_{x=0}^{C-1} \frac{A^x}{x!} + \frac{A^C}{C!}\left(\frac{1}{1-A/C}\right)}$$

$$= \frac{\frac{A^C}{C!}\,\frac{C}{C-A}}{\frac{A^C}{C!}\,\frac{C}{C-A} + \sum_{x=0}^{C-1} \frac{A^x}{x!}} \qquad (3.4)$$

which is known as the Erlang C formula. These three special formulae are frequently used in practice to describe the congestions of the trunk group.

3. Grading

For economic reasons, it is often impossible for subscribers to have access to all trunks in a trunk group. In this case, a special interconnection method known as grading may be used in order to secure the most efficient flow of calls. This interconnection method is as follows: a group of subscribers is given access to individual trunks on the early choices, but on the later choices it shares access to trunks with other groups of subscribers.

Consider an ideal trunking arrangement in which a group of subscribers is given access to only k trunks out of the total $R(>k)$ trunks in the trunk group. It is assumed that the R trunks are equally occupied. Then, for small value of call congestion probability B, Erlang [7] has derived the following approximation:

$$B = \left(\frac{A}{R}\right)^k \qquad \text{for } R,\ A \gg k \qquad (3.5a)$$

$$= A^k \frac{(R-k)!}{k!} \qquad \text{for } A \ll 1 \qquad (3.5b)$$

where A is the traffic offered.

As (3.5) applies theoretically only when A and R are very large or when A is extremely small, a modification is necessary to meet working conditions. O'Dell[9] suggested the following: For a given call congestion probability, the traffic capacity A_g of the trunk group is

$$A_g = A + r(R-k) \tag{3.6}$$

where A is determined from (3.2) with k = C (ie. from $B(k,A,\infty)$ = the specified trunk group congestion probability) and the coefficient r depends on whether the traffic is "smooth" or "pure chance". By smooth traffic, it is meant that the traffic has already passed several stages of equipment and has become smoothed out and lost its pure chance (random) character. For smooth traffic,

$$r = (B(k,A,\infty))^{1}/k \tag{3.7}$$

and for pure chance traffic

$$r = 0.53\ (B(k,A,\infty))^{1}/k + 0.47\ \frac{A}{k} \tag{3.8}$$

where the constants were obtained experimentally [9]. It should be noted that O'Dell took into account neither the mode of operation nor the structure of grading. However, it is usually assumed that the selection of trunks is arranged in a definite order, namely individual trunks first and common trunks last, and that the grading is arranged in the "best" manner. The best, or optimum, grading is considered to be that in which trunk selection from individual trunks to common trunks is smooth [1].

Because of the complexity of the interconnections of trunks and of the associated mode of operation, the more exact solutions to grading are not as simple as the Erlang's and O'Dell's approximations. Since 1920, there has been a huge number of publications in this field [10]. However, because of its simplicty, Erlang's and O'Dell's approximations have been used widely in the telephone administrations [1]. Note that their approximations also assume that every group of subscribers has access to k trunks out of the total R(>k) trunks in the trunk group. In the past, practically all trunking arrangement were done in this way. However, in a modern telephone switching system, this value of k may not be constant for every group of subscribers. A more exact solution to such grading may be found, but any simple approximation is worth considering for practical purposes. The simple approach for approximation introduced here is called the equivalent trunk technique which is outlined in the next section.

4. Equivalent Trunk Technique

Physically, the quantity of trunks in a trunk group can have only positive integral value. However, the equivalent trunk technique assumes that the quantity of trunks in a trunk group can have any positive value. By this assumption, it is possible to find a theoretical full availability trunk group such that its congestion probability equals to the average congestion probabilities of separate full availability trunk groups. The modes of operation of these trunk groups are all assumed on lost calls cleared basis (ie. (3.2) is applicable).

The principle of the equivalent trunk technique is as follows: Let there be m full availability trunk groups. The traffic offered to the ith trunk group consisting of C_i trunks is A_i with average trunk service time h_i. The congestion probability of this ith trunk group, in accordance with (3.2), is

$$B(C_i,A_i,\infty) = \frac{\dfrac{{A_i}^{C_i}}{C_i!}}{\displaystyle\sum_{x=0}^{C_i} \frac{{A_i}^{x}}{x!}} \qquad (3.9)$$

The total traffic offered to these m trunk groups is $\sum_{i=1}^{m} A_i$. If the average congestion probabilities of these m trunk groups is $B(T,A_g,\infty)$, then

$$B(T,A_g,\infty) = \frac{\displaystyle\sum_{i=1}^{m} \frac{A_i}{h_i} B(C_i,A_i,\infty)}{\displaystyle\sum_{i=1}^{m} \frac{A_i}{h_i}} \qquad (3.10)$$

By assumptions, there exists a full availability trunk group consisting of T trunks such that its congestion probability is $B(T,A_g,\infty)$ with the offered traffic being

$$A_g = \sum_{i=1}^{m} A_i. \tag{3.11}$$

Knowing the numerical magnitudes of A_g and $B(T,A_g,\infty)$, it is possible, using (3.2), to find an integer C such that

$$B(C+1,A_g,\infty) \leqslant B(T,A_g,\infty) \leqslant B(C,A_g,\infty). \tag{3.12}$$

It follows, then,

$$C \leqslant T \leqslant C+1. \tag{3.13}$$

Since only an approximation is required, T may be determined from the following linear relation:

$$T = C + \frac{B(C,A_g,\infty) - B(T,A_g,\infty)}{B(C,A_g,\infty) - B(C+1,A_g,\infty)} \tag{3.14}$$

To apply this equivalent trunk technique to grading, a further assumption, that the average trunk service times for all the trunk groups are identical, may be needed. Based on this assumption, (3.10) combined with (3.11) results in

$$A_g.B(T,A_g,\infty) = \sum_{i=1}^{m} A_i.B(C_i,A_i,\infty). \tag{3.15}$$

In a full availability trunk group consisting of T or more trunks, $A_g.B(T,A_g,\infty)$ may be regarded as the traffic passed on from the first T trunks, if the offered traffic is A_g. Based on this fact, the equivalent trunk technique may be used for obtaining the average congestion probability of a given trunk group.

To illustrate this, suppose there are m groups of subscribers. Each group i offers traffic A_i and has access to C_i + x trunks in a trunk group consisting of $x + \sum_{i=1}^{m} C_i$ trunks. The C_i trunks are accessible to only group i subscribers whereas the x trunks are accessible (common) to all subscribers. A call from a subscriber in group, uses one of the x trunks only if all C_i trunks are busy. In other words, the order of trunk selection is as shown in Fig. 6.

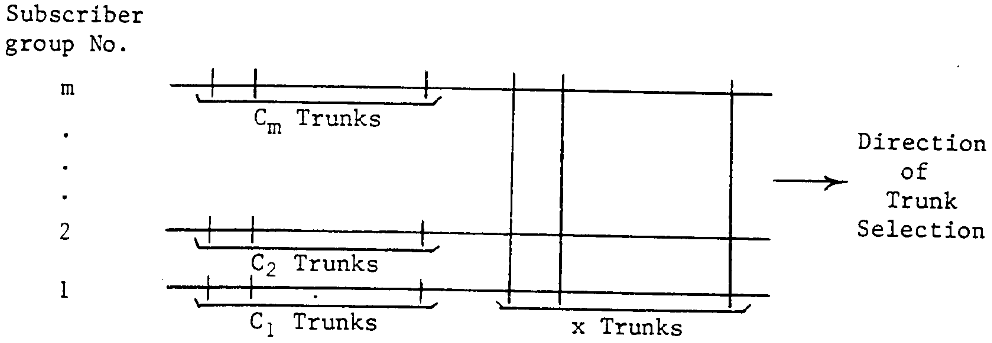


FIG. 6 - A SIMPLE GRADING

Using the equivalent trunk technique, the average congestion probability for this trunk group may be found as follows:

First, consider the C_i trunks associated with group i subscribers. The traffic passed on from these trunks is $A_i.B(C_i,A_i,\infty)$. For all m groups, the total traffic passed on is (3.15). Thus the numerical magnitude of T, the equivalent trunks, can be determined from (3.11), (3.12) (3.14) and (3.15).

Then, construct a theoretical full availability trunk group consisting of T' trunks such that

$$T' = T + x. \tag{3.16}$$

Thus, T is part of the trunks in the new full availability trunk group. If the traffic offered to this new trunk group is A_g, (3.15) is the traffic passed on from the first T trunks, and $A_g.B(T',A_g,\infty)$ is the traffic passed on from all T + x trunks. Therefore, the average congestion probability required is $B(T',A_g,\infty)$ which can be obtained approximately from (3.12) to (3.14).

It is of interest to compare the approximate results obtained by means of the equivalent trunk technique on grading with those obtained using O'Dell's approximation. Let the simple grading to be considered be as shown in Fig. 6. In order to obtain agreement with O'Dell's grading, it is necessary that

$$C_1 = C_2 = \ldots = C_i = \ldots = C_m \tag{3.17}$$

$$A_1 = A_2 = \ldots = A_i = \ldots = A_m \tag{3.18}$$

and

$$C_i + x = k. \tag{3.19}$$

Consider the case where $C_i = x = 10$ and $m = 2$. If the trunk group congestion probability is 0.002, the traffic capacity according to O'Dell (ie. from (3.6), (3.7) and (3.8)) is

$$A_g = 10.07 + 10r$$

where, for smooth traffic,

$$r = (0.002)^{\frac{1}{20}} = 0.733$$

and for pure chance traffic,

$$\begin{aligned} r &= 0.53 \times 0.733 + 0.47 \left(\frac{10.07}{20}\right) \\ &= 0.625 \end{aligned}$$

Therefore,

$$\begin{aligned} A_g &= 17.40 \quad \text{for smooth traffic} \\ &= 16.32 \quad \text{for pure chance traffic.} \end{aligned}$$

Hence, it follows, from (3.11) and (3.18),

$$\begin{aligned} A_i &= 8.70 \quad \text{for smooth traffic} \\ &= 8.16 \quad \text{for pure chance traffic.} \end{aligned}$$

If these types of traffic are offered to the trunk group considered, the congestion probability approximated by means of the equivalent trunk technique is as follows:

First, let $A_i = 8.70$. Therefore, from (3.15), $\boxed{11}$,

$$\begin{aligned} B(T,\ 17.40,\ \infty) &= B(10,\ 8.70,\ \infty) \\ &= 0.154. \end{aligned}$$

By (3.14)

$$\begin{aligned} T &= 17 + \frac{0.182 - 0.154}{0.182 - 0.150} \\ &= 17.88 \end{aligned}$$

By (3.16)

$$T' = 17.88 + 10 = 28.88$$

Therefore, the congestion probability required is, from (3.14),

$$B(28.88,\ 17.40,\ \infty) = 0.005 - 0.88\ (0.005 - 0.003) \doteq .0033$$

which is slightly greater than O'Dell's (0.002) for smooth traffic.

Now, let A = 8.16. Similarly, one obtains $B(T, 16.32, \infty) = 0.130$. Therefore, $T = 17.67$ and $T' = 27.67$. Thus, $B(27.67, 16.32, \infty) \doteq 0.0030$. Again, the approximation by the equivalent trunk technique results in a slightly higher congestion probability than that of O'Dell (0.002). However, for practical purposes, these differences are insignificant. Indeed, this technique has been used successfully by the Canadian telephone administrations for the first Canadian-developed switching system - the SA-1.

Chapter IV

Network Blocking Probability

1. Introduction

For the convenience of engineering, the switching network in a telephone switching system is usually considered as a major sub-unit. To estimate the overall grade of service of an exchange, it is necessary to know the congestion distribution function of every sub-unit of the switching system in the exchange. Therefore, it is essential to study the blocking probability of a given network.

In general, the structure of the network may have the form as shown in Fig. 7. It has z switching stages. In each switching stage k, there are W_k switches.

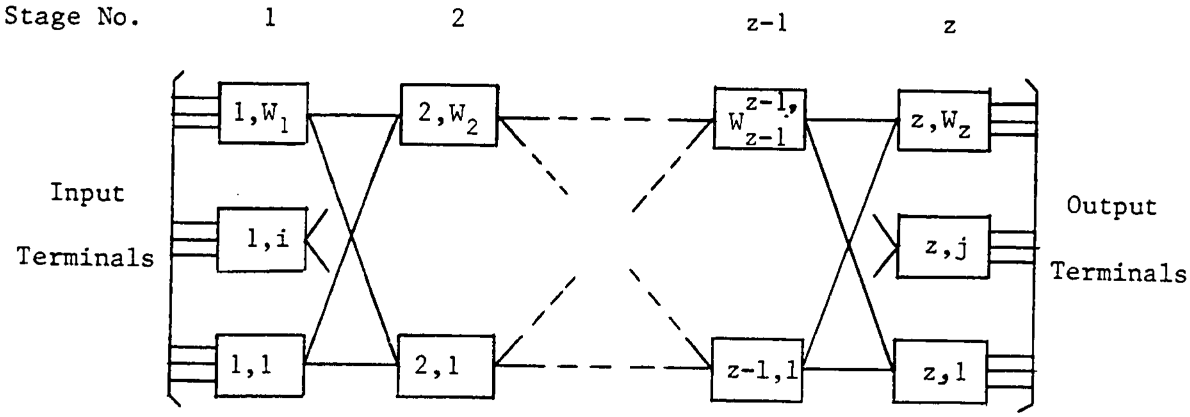


FIG. 7 - A GENERAL TELEPHONE SWITCHING NETWORK

Each switch has a certain number of inlets and outlets. The number of inlets and outlets in a switch may or may not be equal, but each inlet has full access to the outlets and vice versa. Between switching stages, the links (junctors) are joining the outlets of the switches in a stage to the inlets of the switches in the following stage forming a linking stage. Thus, for switching network of z stages there are z-1 linking stages. The inlets of the switches in switching stage 1 are called the input terminals, and the outlets of switches in stage z the output terminals. Each input terminal may be connected to a subscriber line whereas each output terminal to a port of a trunk or of a service circuit. A trunk or a service circuit may have one or more ports. Thus, it is possible for a trunk or a service circuit to connect to more than one output terminal, one port per terminal. As examples, an intra-office trunk is connected to two output terminals, and a conference circuit to three or more output terminals.

An ideal switching network is the non-blocking network [12]. In such a network it is always possible to establish a connection from an idle input terminal to an output terminal regardless of the nubmer of calls served by the network. However, practical switching networks are usually designed on the basis of limited paths between input and output terminals. Consequently, a demand for service may be unsatisfied when all the paths capable of connecting a pair of input and output terminals in the network are busy. The probability that all the paths between a pair of input and output terminals in the network are busy is known as the network blocking probability. However, it will be shown later that such a definition for network blocking probability is not sufficient when the network is loaded with multitype traffic.

The blocking probability for the telephone switching network has been studied fairly extensively [4,13-15]. However, in this chapter, the widely used Lee - LeGall [1] analytical model for evaluating the blocking probability of a given switching network is first briefly reviewed. Then an approximate method [16] to supplement this model for computing complex network blocking probability is presented. Finally, a single linking-stage network loaded with multitype traffic is investigated.

2. Lee - LeGall Method

The Lee - LeGall analytical model for evaluating the blocking probability of a given switching network is essentially the probability linear graph which has been outlined previously in Chapter II. In this model, blocking probability is easily computable when the network consists of a series-parallel arrangement of links. However, in more complicated (and more realistic) networks, the computation may be very involved. For this reason, Grantges and Sinowitz [13] have proposed a simulation method to avoid the computational difficulties associated with complex networks. But the problems with the simulation are as follows: It simulates only an approximated model, it provides only numerical results and hence lacks the insight of an analytical method, and it usually requires a large amount of programming effort.

Therefore, it is convenient to have an approximate formula for estimating the blocking probability of a complex network readily. This will not only enable one to evaluate whether simulation or complex computational effort is worth spending on a given switching network, but also will provide a tool for checking the results from computation or simulation. Indeed, for some purposes, blocking probability of a given network obtained by means of an easily computable approximate formula is sufficient.

Of course, if the switching network is relatively simple, the blocking probability can be easily calculated. Hence, in this case, there is no need for simulation or approximation. Therefore, the approximate formula developed in the following will be useful only when the switching network has four or more stages and is not a simple series-parallel structure.

2.1 BASIC MODEL FOR APPROXIMATION

Consider a general L-stage switching network, as shown in Fig. 8. It is assumed that there is always one and only one junctor between any pair of switches in adjacent stages.

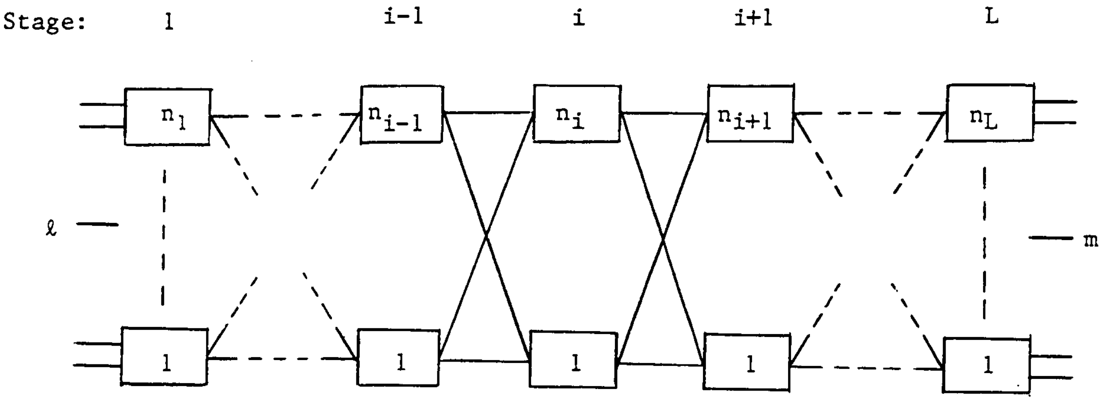


FIG. 8 - GENERAL SWITCHING NETWORK

There are no junctors between switches within a stage nor between switches which are not in adjacent stages. It should be noted that when there are more than one junctors between a pair of switches in the adjacent stages, this model is still applicable, if these multiple junctors are reduced to an equivalent single junctor. Therefore, between any input terminal ℓ and any output terminal m, a linear graph can be drawn, as shown in Fig. 9. In the linear graph, each switch is represented by a node and each junctor by a directed link.

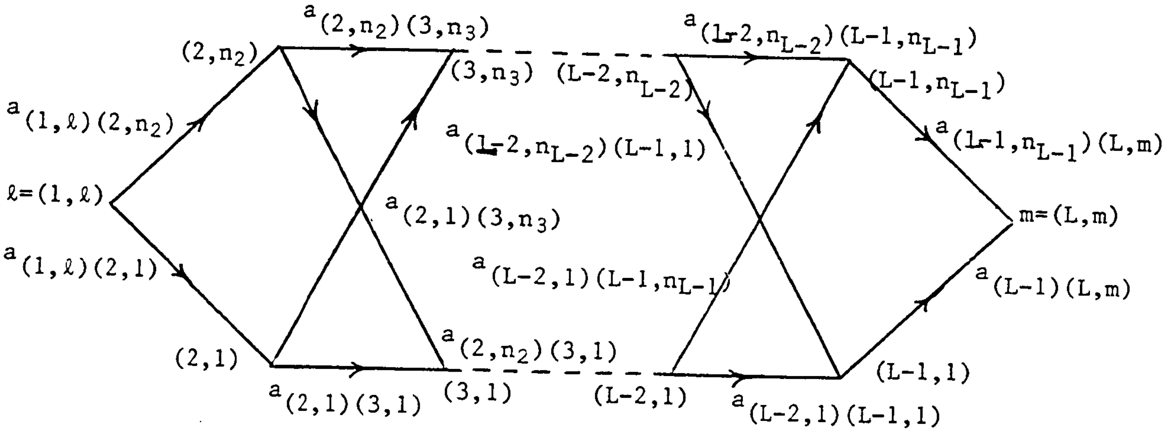


FIG. 9 - LINEAR GRAPH FOR THE NETWORK

The notation used is as follows:

(i,j) is the jth node in the ith stage of the switching network,

n_i is the number of nodes (switches in stage i, and

$a_{(i,j)(i+1,k)}$ is the link occupancy of the link between nodes (i,j) and (i+1,k) i.e., the probability that the link between nodes (i,j) and (i+1,k) is busy.

The general formula for calculating the blocking probability B between a pair of nodes ℓ and m in a switching network is (2.4).

The blocking probability B may also be found by using its generating function [4]. Suppose that there are M directed paths in the linear graph. Each directed path is composed of links

$$X_{(1,\ell)(2,i_1)},\ X_{(2,i_1)(3,i_2)},\ \cdots,\ X_{(L-1,i_{L-2})(L,m)}$$

in series. Denote the directed path by the formal product f_i, i.e.,

$$f_i = X_{(1,\ell)(2,i_1)} \cdot X_{(2,i_1)(3,i_2)} \cdots X_{(L-1,i_{L-2})(L,m)}. \tag{4.1}$$

In this formal product, the X's are considered as undefined real numbers and are manipulated as such except for the reduction rule

$$X_{(i,j)(i+1,k)} \cdot X_{(i',j')(i'+1,k')} = X_{(i,j)(i+1,k)} \tag{4.2}$$

if i = i', j = j', and k = k'. Otherwise the operation is ordinary multiplication. Then, the generating function B(t) of the blocking probability B is defined as

$$B(t) = \Pi^* (1 - f_i t) \tag{4.3}$$

where Π^* denotes formal product such that the reduction rule is always carried out. After B(t) is expanded into a sum of products, $X_{(i,j)(i+1,k)}$ is replaced by $(1 - a_{(i,j)(i+1,k)})$. Then the blocking probability is equal to the generating function with t unity, i.e., B = B(1).

In the case when the occupancies of the links between two adjacent stages are identical, that is

$$a_{(i,1)(i+1,1)} = a_{(i,1)(i+1,2)} = \cdots = a_{(i,n_i)(i+1,n_{i+1})} \tag{4.4}$$

but $a_{(i,1)(i+1,k)}$ is not necessarily equal to $a_{(i',j')(i'+1,k')}$; the blocking probability B sometimes can be found more easily by means of combinatorial analysis [14] than by means of the above two methods. However, when the network is relatively complex, even in this special case (when all link occupancies are the same), computational difficulaty still arises.

Therefore, a general approximate formula for this model is desired.

2.2 DERIVATION OF APPROXIMATE FORMULA

A. The Blocking Probability Bounds

Pick any two adjacent stages, i and i+1, from the linear graph of the general L-stage switching network, as shown in Fig. 10. An upper and a lower bound of the network blocking probability can be found by disturbing only this section, as follows:

1. Lower Bound: In the general L-stage switching network, there are no junctors between switches within a stage. Therefore, there is no link between nodes within a stage in the section (Fig. 10) considered. Since the blocking probability of the switching network will be calculated in accordance with Lee - LeGall model, it will not be affected by the addition of extra links whose occupancy is unity. Suppose, to each stage in the section considered, extra links are added between switches within each stage. Then the resulting network blocking probability will be smaller than the general network blocking probability if the occupancies of these extra links are not all unity. This can be easily seen from the generating function (4.3). It is convenient to let all the occupancies of these extra links be zero; then the network section considered (Fig. 10) takes the form of Fig. 11. For the entire network with only section (i,i+1) designed in this way, the network structure is simplified, and the network blocking probability will be smaller than that of the actual general network.

2. Upper Bound: On the other hand, if the occupancies of some junctors of the general L-stage switching network are increased, the resulting blocking probability will obviously be greater.

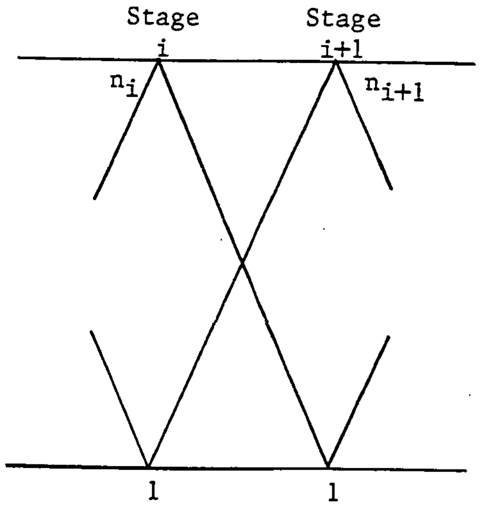


FIG. 10 - A SECTION OF THE NETWORK

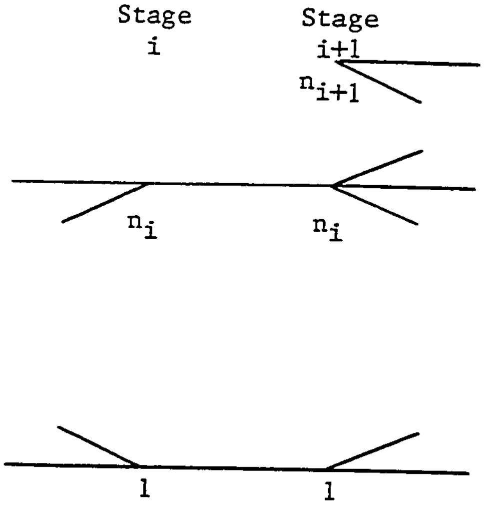


FIG. 12 - MODIFICATION FOR EVALUATING B_1.

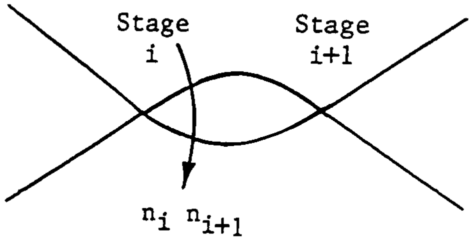


FIG. 11 - MODIFICATION FOR EVALUATING B_0.

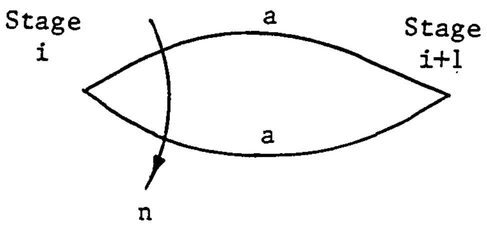


FIG. 13 - MODEL FOR $F_{i,i+1}$.

Again, consider the section of adjacent stages i and i+1 (Fig. 10) and assume $n_i \leq n_{i+1}$. If occupancies of all the links except $a_{(i,j)(i+1,j)}$'s are equated to unity, the section considered will then take the form of Fig. 12. Thus the network structure with only section (i,i+1) designed in this way is simplified, and the resulting network blocking probability will be greater than that of the actual network. It should be noted that there are many choices of retaining $a_{(i,j)(i+1,j)}$'s for an upper bound network structure. One choice may give a better result than another. Usually, keeping the smaller occupancies will result in a better approximation.

Summary: To sum up, let

B be the blocking probability between terminals ℓ and m in the general L-stage switching network;

B_0 be the blocking probability between the terminals ℓ and m with the occupancies of the additional inserted junctors within stage i and those within stage i+1 all zero; and

B_1 be the blocking probability between terminals ℓ and m when all link occupancies $a_{(i,j)(i+1,k)}$'s except $j = k$ between adjacent stages i and i+1 are equated to unity; then,
$B_0 \leq B \leq B_1$.

B. Approximate Formula

In order to get a simple (but a nontrivial) approximation, assume that the approixmate blocking probability B_a is of the form:

$$B_a = B_0 + (B_1 - B_0)\, F_{i,i+1}, \qquad (4.5)$$

where the value of $F_{i,i+1}$ must be in the range of $0 \leq F_{i,i+1} \leq 1$ and $F_{i,i+1}$ depends on the actual occupancies and the number of links between stages i and i+1. To determine a suitable factor $F_{i,i+1}$ for a given switching network, consider the following facts:

1) As the link occupancies $a_{(i,j)(i+1,k)}$'s between stages i and i+1 approach zero, the network blocking probability B approaches B_0. Hence, the value of $F_{i,i+1}$ should decrease as the link occupancies decrease, and should reach zero when the occupancies become zero.

2) Equating a link occupancy to unity is equivalent to removing that link from the network. The larger the number of links between stages i and i+1, the larger the number of links removed in order to obtain B_1. Hence, the larger the number of links, the larger the difference $B_1 - B$. Therefore, to achieve a better approximation with formula (4.5), $F_{i,i+1}$ should decrease as the number of links between stages i and i+1 increases.

To satisfy the above two conditions, many functions can be constructed. However, one must remember that the approximate formula must be simple or else it will lose its significance. Suppose that the factor $F_{i,i+1}$ will take the form analagous to the blocking probability of a simple network which has only two nodes connected by n links, as shown in Fig. 13. This blocking probability is a^n (where a is the occupancy of each link) and thus satisfies the conditions. Then to account for condition 1) let the occupancy of every link in Fig. 13 be proportional to the arithemtic average of the occupancies of the links between the two adjacent stages i and i+1 considered. To account for condition 2), let the number of links in Fig. 13 equals the geometric average of the number of links per node of the two adjacent stages i and i+1.

The arithmetic average of the occupancies of the links between the two adjacent stages i and i+1 is defined as usual, ie.:

$$\overline{a}_{i,i+1} = \frac{\sum_{j=1}^{n_i} \sum_{k=1}^{n_i+1} a_{(i,j)(i+1,k)}}{n_i n_{i+1}} . \tag{4.6}$$

The occupancy of a link in Fig. 13 is proportional to $\overline{a}_{i,i+1}$. When the occupancies of the links between stages i and i+1 are identical to those for obtaining B_1, the value of $F_{i,i+1}$ must be unity. Therefore, the occupancy of a link in Fig. 13 is $\overline{a}_{i,i+1}/\overline{a}'_{i,i+1}$ where, assuming $n_i \leq n_{i+1}$,

$$\overline{a}'_{i,i+1} = \frac{n_i n_{i+1} - n_i + \sum_{j=1}^{n_i} a_{(i,j)(i+1,j)}}{n_i n_{i+1}} \tag{4.7}$$

which is the arithmetic average of the occupancies of the links between stages i and i+1 of the linear graph for evaluating B_i.

Also, as usual, the geometric average of the number of links between stages i and i+1 per node is $\sqrt{n_i n_{i+1}}$. Therefore, in Fig. 13 each link has an occupancy of $\overline{a}_{i,i+1}/\overline{a}'_{i,i+1}$, and there are $\sqrt{n_i n_{i+1}}$ links. Thus, its blocking probability and hence, the value of factor $F_{i,i+1}$ is

$$F_{i,i+1} = \left(\frac{\bar{a}_{i,i+1}}{\bar{a}'_{i,j+1}}\right)^{\sqrt{n_i n_{i+1}}} \tag{4.8}$$

which satisfies the requirements.

C. Choice of Starting Point

The result of the approximate blocking probability between an input and an output terminal is not unique. It depends not only on the choice of retaining $a_{(i,j)(i+1,j)}$'s for obtaining the upper bound network structure as pointed out previously, but it also depends on the choice of the two adjacent stages to be disturbed for obtaining the approximation.

In retaining $a_{(i,j)(i+1,j)}$'s for the upper bound structure, it has been pointed out that better approximation results by keeping the smaller occupancies. For the choice of the two adjacent stages to be disturbed for obtaining the approximation, the best result occurs when $(B_1 - B)$ is a minimum. Since $(B_1 - B_a)$ is not known until it is computed, it is suggested to use those stages which result in the greatest simplification. As a general rule, this may be achieved by disturbing the two adjacent stages having the largest number of actual links. This will be illustrated in the 2.3 below.

2.3 APPLICATION OF APPROXIMATE FORMULA

A. Four-Stage Switching Network

The linear graph between terminal $(1,\ell)$ and $(4,m)$ in a four-stage switching network is as shown in Fig. 14(a). The largest number of links in the linear-graph occurs between stages 2 and 3. Therefore, these stages are to be disturbed. To evaluate B_0, imagine that there are links between switches within stage 2 and within stage 3 and the occupancies of these links are all zero. Thus, the linear graph for obtaining B_0 is as shown in Fig. 14(b). The linear graph for evaluating B_1 is obtained by equating all $a_{(2,i)(3,j)}$'s except $i = j$ to unity, as shown in Fig. 14(c), where $n_2 \leq n_3$ is assumed. As the linear graphs of Figs. 14(b) and 14(c) involve only the parallel and series forms of directed paths, their blocking probabilities can be easily evaluated using the Lee - LeGall model. Thus,

$$B_0 = 1 - \left(1 - \prod_{i=1}^{n_2} a_{(1,\ell)(2,i)}\right)\left(1 - \prod_{i=1}^{n_2}\prod_{j=1}^{n_3} a_{(2,i)(3,j)}\right) \cdot \left(1 - \prod_{j=1}^{n_3} a_{(3,j)(4,m)}\right), \qquad (4.9)$$

and

$$B_1 = \prod_{i=1}^{n_2}\left[1 - (1 - a_{(1,\ell)(2,i)})(1 - a_{(2,i)(3,i)})(1 - a_{(3,i)(4,m)})\right]. \qquad (4.10)$$

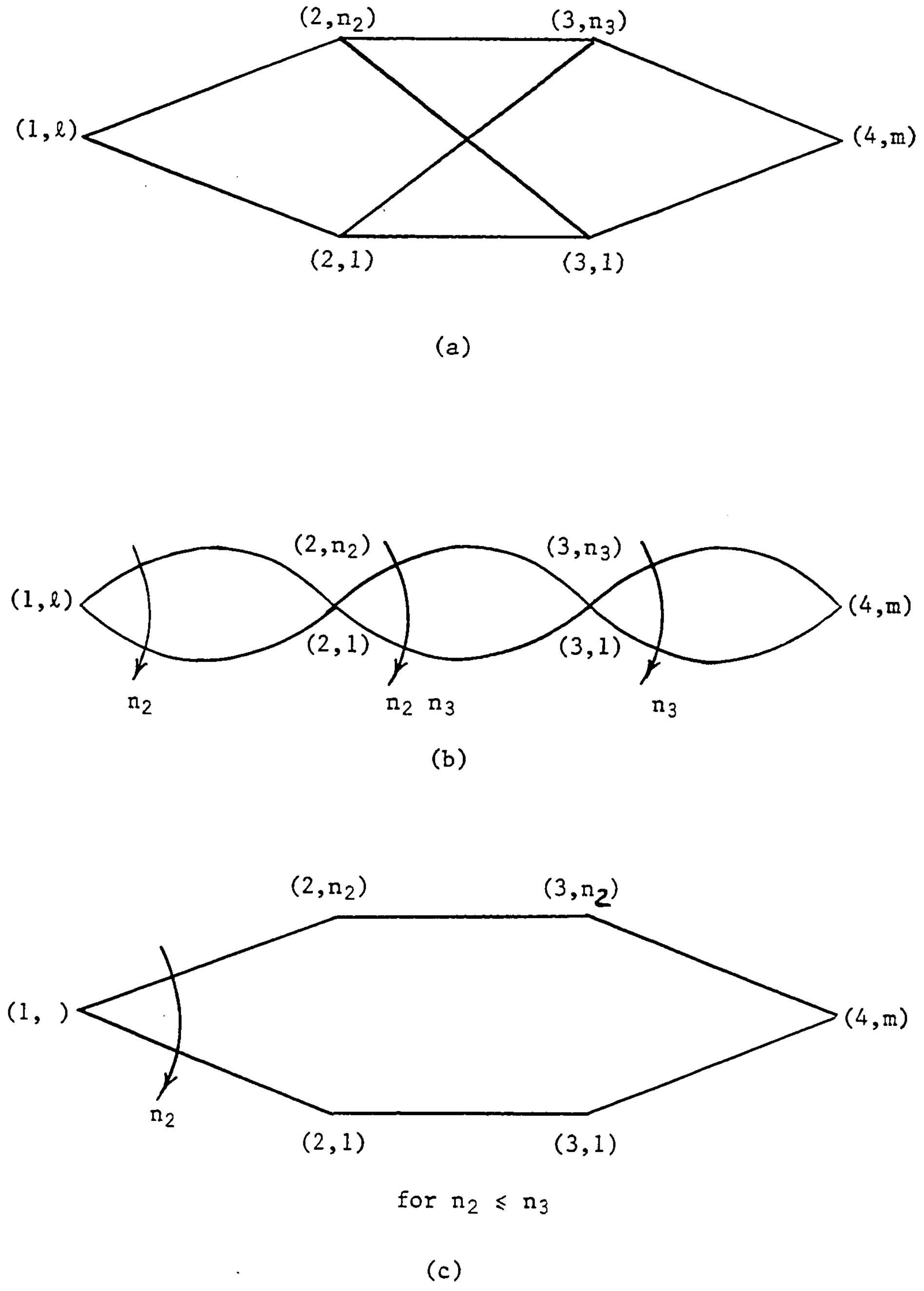


FIG. 14 - (a) 4-STAGE NETWORK - (b) GRAPH FOR B_0 - (c) GRAPH FOR B_1.

Replacing i by 2 in (4.6) - (4.8), one obtains

$$F_{2,3} = \left(\frac{\bar{a}_{2,3}}{\bar{a}'_{2,3}}\right)^{\sqrt{n_2 n_3}} \qquad (4.11)$$

Hence, substituting (4.9) - (4.11) into (4.5) yields the approximate blocking probability.

B. Switching Network with More Than Four States

1. Step-by-Step Method: Consider a six-stage switching network having the linear graph shown in Fig. 15(a). The occupancies of the links between stages 4 and 5 of the linear graph can be disturbed resulting in Figs. 15(b) and 15(c). The blocking probability between terminals (1,ℓ) and (6,m) in Fig. 15(b) is B_0, and that in Fig. 15(c) is B_1 for that in Fig. 15(a). The $F_{i,i+1}$ factor here is, of course,

$$F_{4,5} = \left(\frac{\bar{a}_{4,5}}{\bar{a}'_{4,5}}\right)^{\sqrt{n_4 n_5}} \qquad (4.12)$$

To find B_0, the linear graph in Fig. 15(b) may be divided into two subgraphs in series. One subgraph consists of stages 1 to 4, another consists of stages 4 to 6. If the blocking probability between terminals (1,ℓ) and (4,i) and that between (4,i) and (6,m) are known, then the blocking probability between terminals $(1,\ell)$ and $(6,m)$ can be easily calculated in the standard manner. For the subgraph consisting of stages 4 to 6, the blocking probability is

$$\prod_{i=1}^{n_5} a_{(5,i)(6,m)} + \left(1 - \prod_{i=1}^{n_5} a_{(5,i)(6,m)}\right) \prod_{j=1}^{n_4} \prod_{k=1}^{n_5} a_{(4,j)(5,k)}.$$

For the subgraph consisting of stages 1 to 4, the blocking probability may be calculated in accordance with 2.3 - A above.

To find B_1, let

$$a_{(4,i)(6,m)} = a_{(4,i)(5,i)} + \left(1 - a_{(4,i)(5,i)}\right) a_{(5,i)(6,m)} \tag{4.13}$$

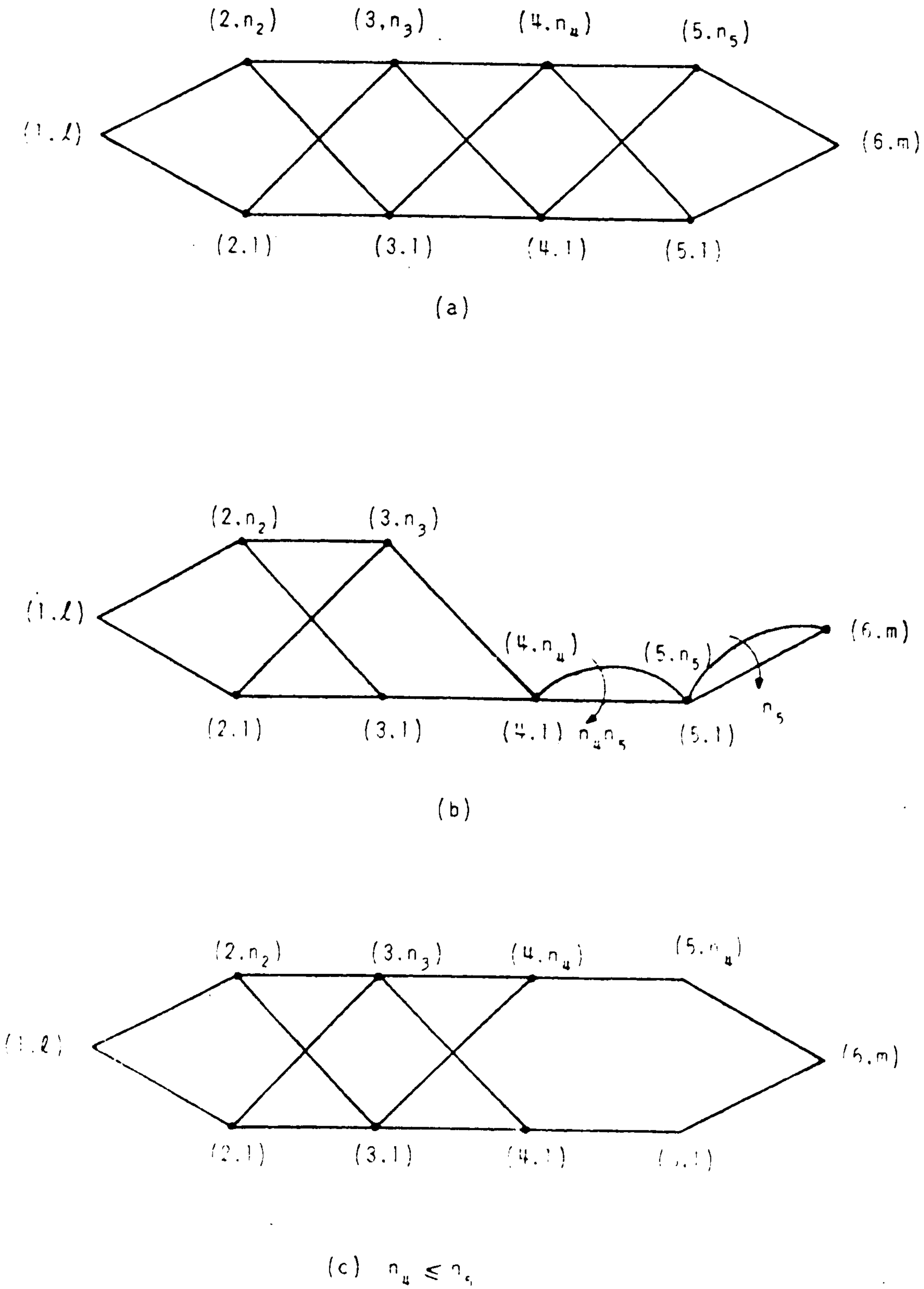


Fig. 15 - (a) 6-stage network. (b) Graph for B_4.
(c) Graph for B [illegible].

in Fig. 15(c). Thus, Fig. 15(c) becomes essentially a five-stage network shown in Fig. 16(a). In this linear graph Fig. 16(a), another pair of adjacent stages can be disturbed resulting in Figs. 16(b) and 16(c). The blocking probability between terminals $(1,\ell)$ and $(6,m)$ in Fig. 16(b) and that in Fig. 16(c) are B_0 and B_1, respectively, for that in Fig. 16(a). Since the linear graph for Fig. 16(b) is a series-parallel type, the blocking probability can be easily calculated. In Fig. 16(c), if further let

$$a_{(3,i)(6,m)} = a_{(3,i)(4,i)} + (1 - a_{(3,i)(4,i)})\ a_{(4,i)(6,m)} \tag{4.14}$$

it becomes a 4-stage network. Thus the blocking probability can be calculated in accordance with 2.3 - A.

To summarize at this point: The blocking probability of 4-stage graph Fig. 16(c) is used as B_1 (together with B_0 obtained from Fig. 16(b) and with $F_{3,4}$) to calculate the blocking probability of the 5-stage network Fig. 16(a). In turn, this blocking probability is used as B_1 (together with B_0 obtained from Fig. 15(b) and with $F_{4,5}$) to calculate the blocking probability of the 6-stage network Fig. 15(a).

While the above concerns only the six-stage network, this method of obtaining the blocking probability between two terminals by reducing one stage at a time may be extended to any number of stages. It should be noted that details of computation are affected by the choice of stages to be disturbed. Suppose, in the previously considered six stages, stage 3 and 4 are disturbed instead of 4 and 5. One obtains the linear graphs of Fig. 17(b) and 17(c) for its B_0 and B_1, respectively. Fig. 17(b) is a series-parallel type of network, and hence its blocking probability can be easily found. Although the blocking probability of Fig. 17(c) cannot be easily evaluated in its present form, one can apply this step-by-step technique described above.

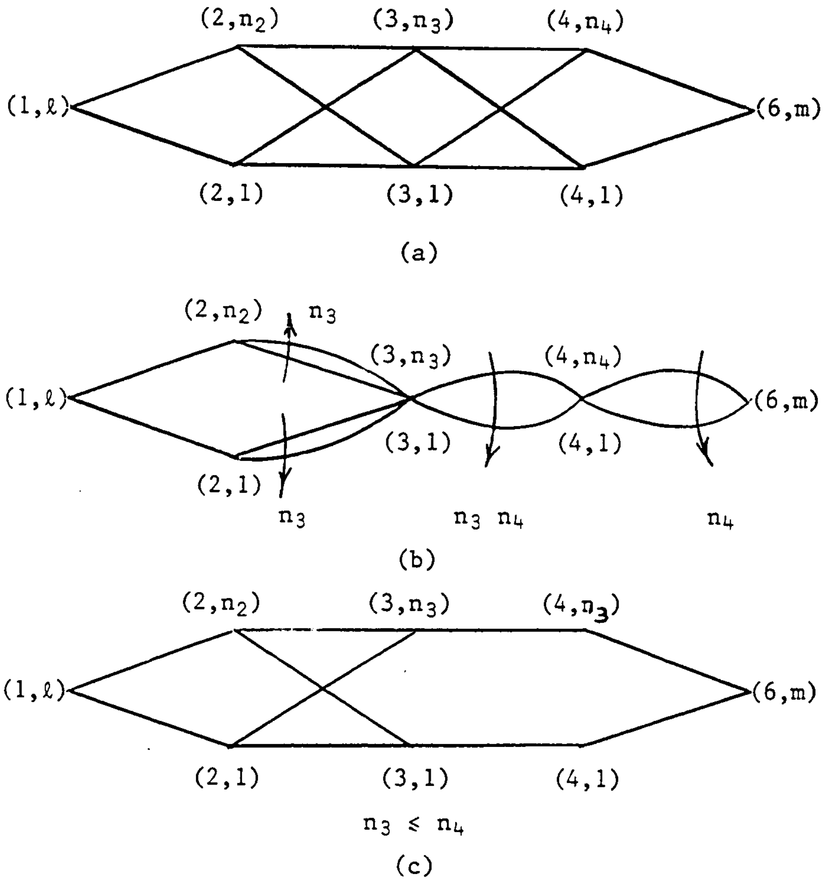


FIG. 16 - SECOND STEP IN SIMPLIFYING THE NETWORK OF FIG. 15(a).

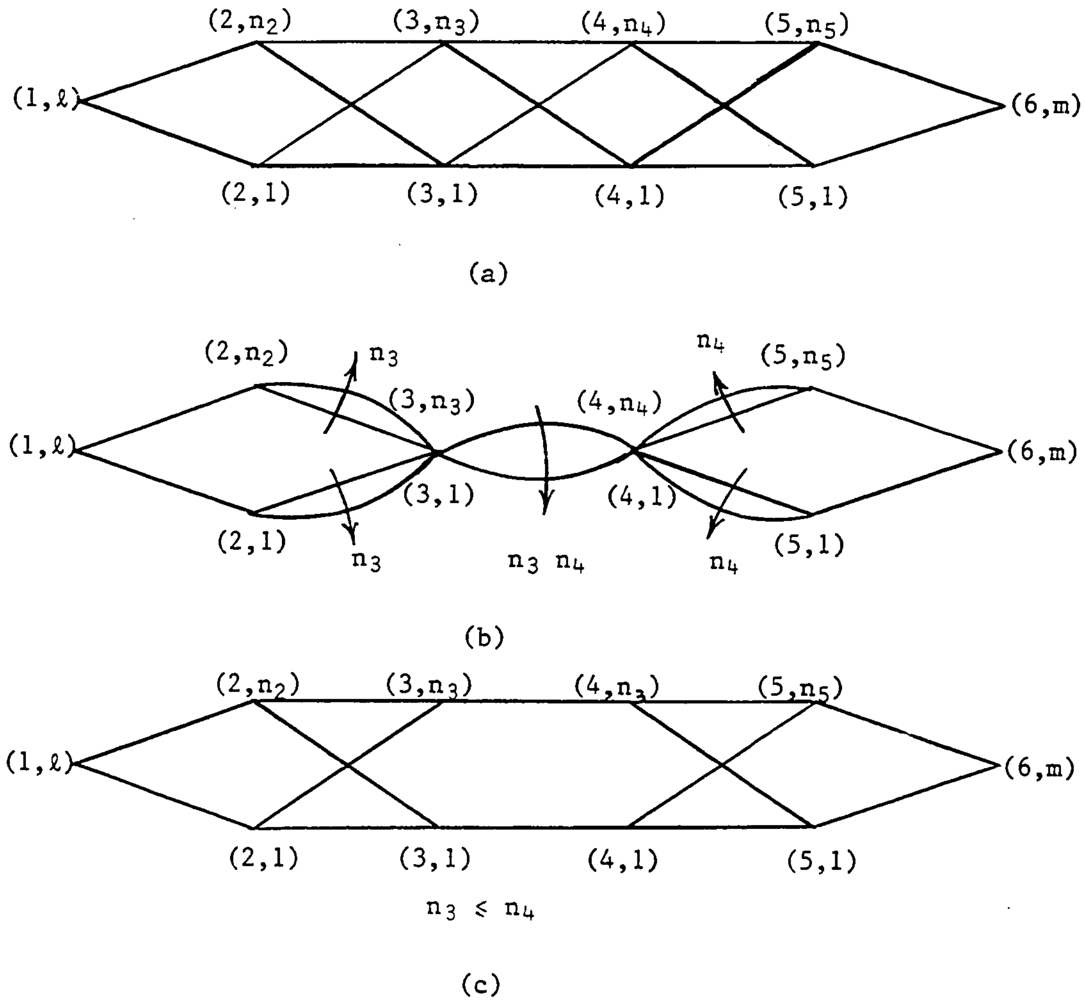


FIG. 17 - A DIFFERENT REDUCTION OF THE NETWORK OF FIG. 15(a).

In general, these two approximated blocking probabilities may not be the same; but hopefully both should be reasonably close to the exact figure.

2. Average F-Factor Method: When the number of stages in a switching network is large, the step-by-step method described above becomes very laborious. In this case, one may obtain the approximate blocking probability as follows: First, redefine:

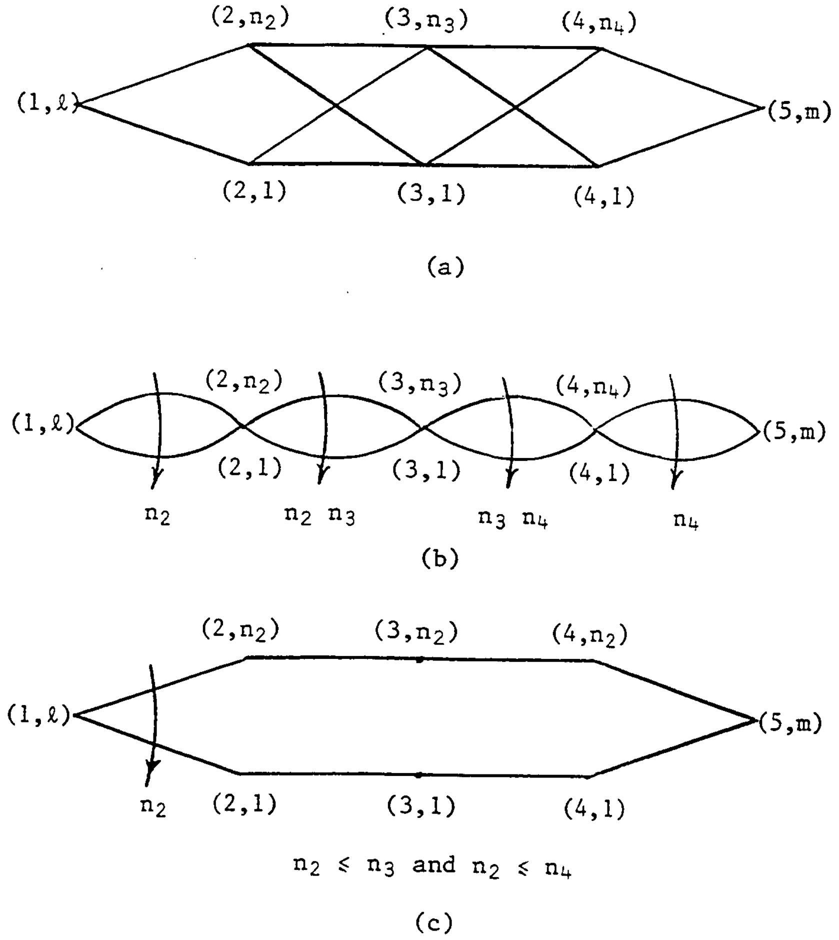


FIG. 18 - ILLUSTRATING THE AVERAGE F-FACTOR METHOD.

B_0 as the blocking probability between the terminals ℓ and m with the occupancies of the imaginary links between switches within each stage all zero.

B_1 as the blocking probability between terminals ℓ and m when all link occupancies $a_{(i,j)(i+1,k)}$'s between stages i and i+1 except j = k are equated to unity for i = 2, 3, ... L-2.

Next, let

$$F = \frac{1}{L-3} \sum_{i=2}^{L-2} F_{i,i+1}, \qquad (4.15)$$

where $F_{i,i+1}$ is the $F_{i,i+1}$ of (4.8), that is

$$F_{i,i+1} = \left(\frac{\bar{a}_{i,i+1}}{\bar{a}'_{i,i+1}}\right)^{\sqrt{n_i n_{i+1}}}$$

Finally, since the relation $B_0 \leq B \leq B_1$ still holds, similar to (4.5), assume

$$B_a = B_0 + (B_1 - B_0)F. \qquad (4.16)$$

As an example, consider a five-stage switching network having a linear graph as shown in Fig. 18(a). Then, the linear graphs for B_0 and B_1 become as shown in Figs. 18(b) and 18(c) respectively. Since these linear graphs involve only parallel and series links, one can easily evaluate the blocking probabilities. As has been noted before, the factor F is quite easy to compute. Hence, the blocking probability can be readily obtained. Thus a cruder approximation can be obtained with less effort.

2.4 NUMERICAL EXAMPLES

It has been shown that B_0, B_1, and $F_{i,i+1}$ are relatively simple to compute. At this point, it is of interest to investigate how well the approximate formula is in agreement with results obtained by means of direct computation or simulation. When the switching network is relatively complex and all link occupancies are different, the blocking probability can be approximated quite easily with the method presented here. However, the exact blocking probability of such a network may not be easy to calculate. Therefore, it is only possible to compare the results by different methods for some special cases in which the blocking probability can be directly calculated or has been obtained by simulation.

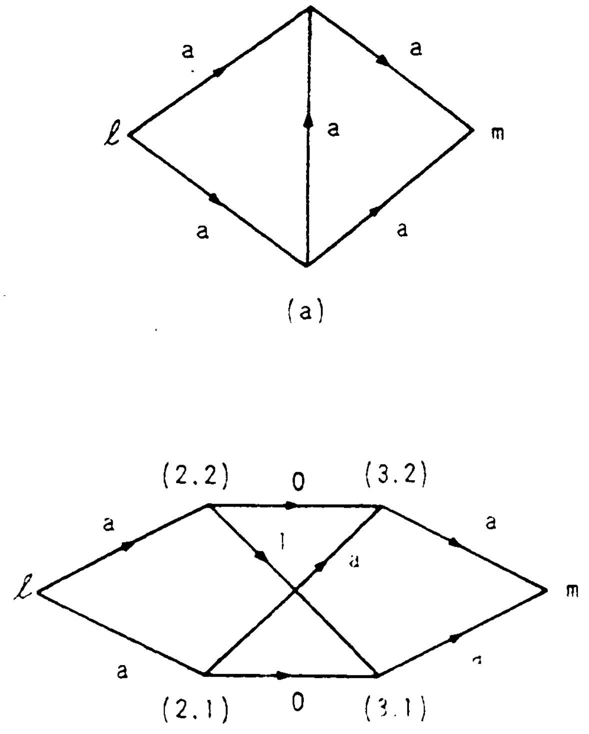
Case 1

Consider the network having a linear graph, as shown in Fig. 19(a), in which the occupancy of every link equals a. In this case, it is not difficult to find the blocking probability between terminals ℓ and m by applying either (2.4) or (4.3). Thus:

$$B = 1 - 2(1 - a)^2 - (1 - a)^3 + 3(1 - a)^4 - (1 - a)^5.$$

If the approximate blocking probability is to be found by the method presented here, first Fig. 19(a) must be made equivalent to the model considered in 2.1 above, as shown in Fig. 19(b). Thus, it becomes a standard 4-stage switching network. Therefore, according to (4.9), (4.10), and (4.11), respectively, one obtains

$$B_0 = a^2(2 - a^2)$$



(a)

(b)

FIG. 19 - NETWORK FOR CASE 1

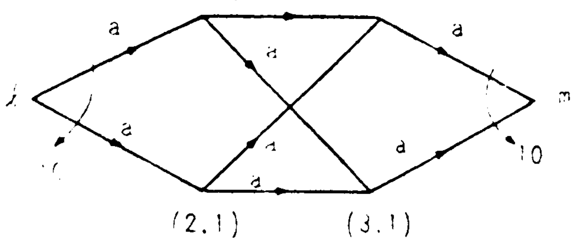


FIG. 20 - NETWORK FOR CASE 2

$$B_1 = a^2(2 - a)^2$$

and

$$F_{2,3} = \left(\frac{1 + a}{2}\right)^2.$$

Here $a_{(2,1)(3,1)} = 0 = a_{(2,2)(3,2)}$ have been retained and $a_{(2,1)(3,2)} = a$ and $a_{(2,2)(3,1)} = 1$ are equated to unity for obtaining B_1. Then the approximate blocking probability is given by (4.5). The results of B and B_a are plotted for various values of a in Fig. 24.

Case 2

Consider the network having a linear graph of Fig. 20, where the occupancy of every link equals a. If (2.4) or (4.3) is used to find its blocking probability between terminals ℓ and m, the process is very laborious. Since the occupancies of the links are identical, one can use the combinatorial analysis approach [14] to find its blocking probability. Thus,

$$B = \sum_{i=0}^{10} \binom{10}{i} a^{10-i} (1 - a)^i \left[a + (1 - a)a^i\right]^{10}.$$

The approximate blocking probability is given by (4.5), where

$$B_0 = 1-(1 - a^{10})^2 (1 - a^{100})$$

$$B_1 = 1 - (1 - a)^{3\ 10}$$

and

$$F_{2,3} = \left(\frac{10a}{9+a}\right)^{10}.$$

The results of both B and B_a are plotted in Fig. 25.

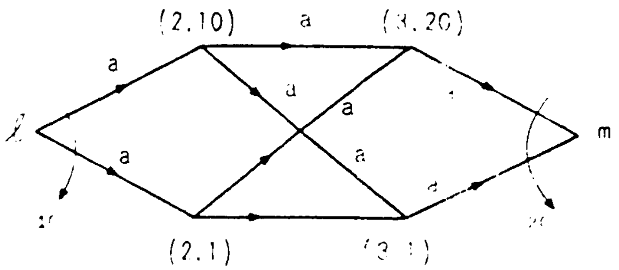


FIG. 21 - NETWORK FOR CASE 3.

Case 3

Consider a 4-stage switching network linear graph shown in Fig. 21, where the number of nodes in stages 2 and 3 are not equal. As the occupancies of all links are equal, one can still make use of the combinatorial analysis approach which gives:

$$B = \sum_{i=0}^{10} \binom{10}{i} a^{10-i}(1 - a)^i \left[a + (1 - a)a^i\right]^{20}.$$

The approximate blocking probability B_a is, of course, (4.5) with

$$B_0 = 1 - (1 - a^{10})(1 - a^{200})(1 - a^{20}),$$

$$B_1 = \left[1 - (1 - a)^3\right]^{10},$$

and

$$F_{2,3} = \left(\frac{20a}{19+a}\right)^{\sqrt{200}}$$

The results of both B and B_a are plotted in Fig. 26.

Case 4

Grantges and Sinowitz have studied a linear graph of moderate complexity, shown in Fig. 22. Their simulated results and the calculated blocking probabilities for $0.44 \leq a \leq 0.58$ are shown in Fig. 27. For comparison, in Fig. 27 the approximate blocking probability B_a by the method presented here are also plotted. Again B_a is (4.5) with

$$B_0 = 1 - (1 - a^8)^2 (1 - a^{32}),$$

$$B_1 = \left[1 - (1 - a)^3\right]^8,$$

and

$$F_{2,3} = \left(\frac{4(1 + a)}{7 + a}\right)^8 .$$

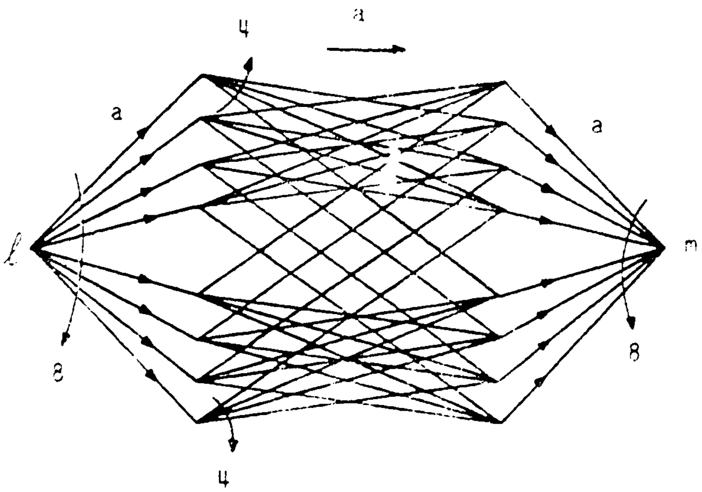


FIG. 22 - NETWORK FOR CASE 4.

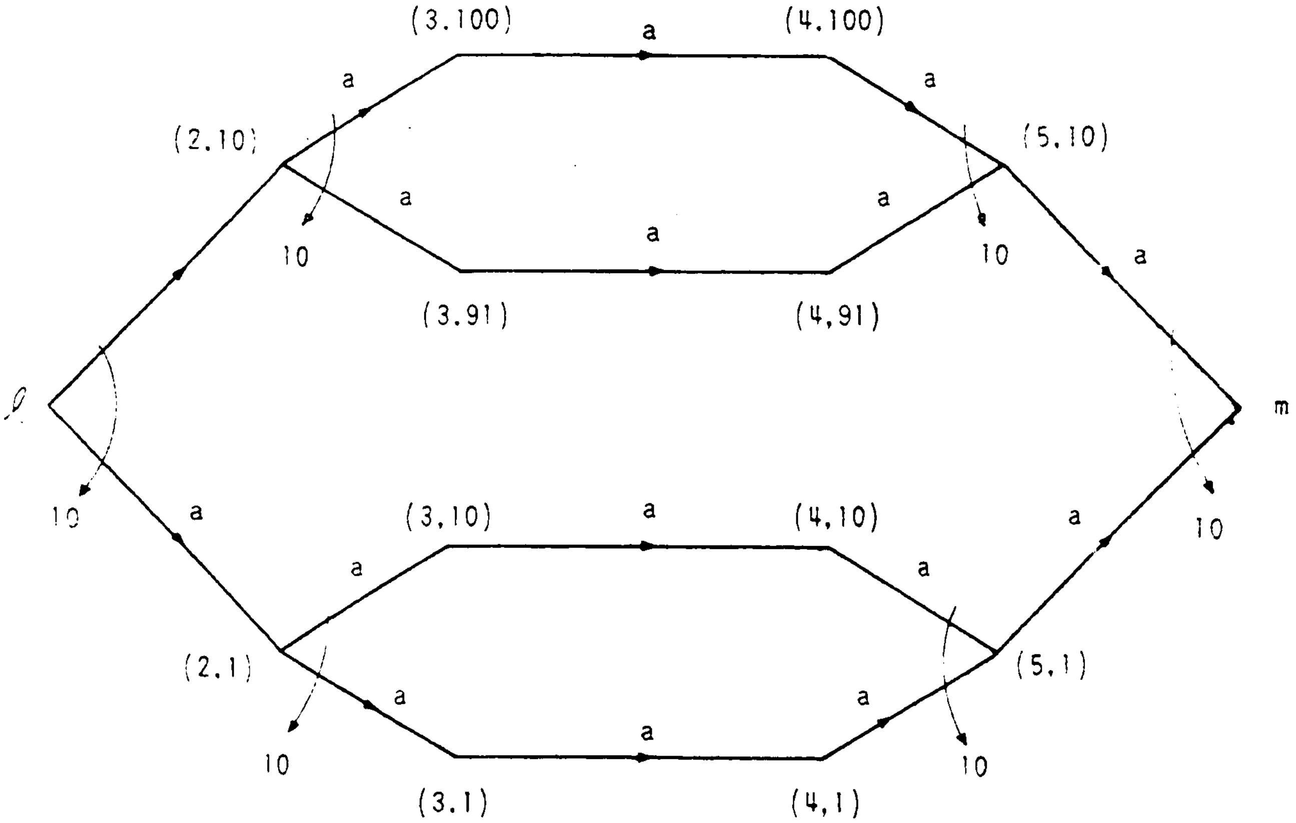


FIG. 23 - NETWORK FOR CASE 5.

Case 5

For a 6-stage network having a linear graph of Fig. 23, the blocking probability between terminals ℓ and m can be easily derived by means of the generating function (4.3). Thus,

$$B = \{1 - (1 - a)^2 \left[1 - (1 - (1 - a)^3)^{10}\right]\}^{10}.$$

Its approximate blocking probability can be found in either one of the two ways: namely, step-by-step method and the average F-factor method. Since the latter is much easier to use, the blocking probability is approximated by the average F-factor method only. Thus,

$$F_{2,3} = F_{4,5} = \left(\frac{90+10a}{99 + a}\right)^{\sqrt{1000}}$$

$$F_{3,4} = \left(\frac{990+10a}{999 + a}\right)^{100}$$

$$F = \frac{2F_{2,3}+F_{3,4}}{3}$$

$$B_0 = 1 - (1 - a^{10})^2 \ (1 - a^{100})^3$$

$$B_1 = \left[1 - (1 - a)^5\right]^{10}$$

and

$$B_a = B_0 + (B_1 - B_0)F.$$

Both B and B_a have been plotted, as shown in Fig. 28.

By examining the results (Figs. 24 - 28)of the above cases, it is seen that the characteristics of the approximate results are generally in agreement with those of the accurate results. A more complete evaluation of the approximation is made difficult by the fact that the exact answer is not easily available. At any rate the approximation appears sufficiently good to be used to obtain a preliminary evaluation of the network.

The approximate method proposed here for evaluating switching network blocking probability has been shown to produce reasonable results and has the advantage of being simple and inexpensive. In complex switching networks, it may serve as a guide to determine whether the effort of direct computation or simulation is worthwhile for a given network, and also provides a means of checking the results obtained from computation and simulation.

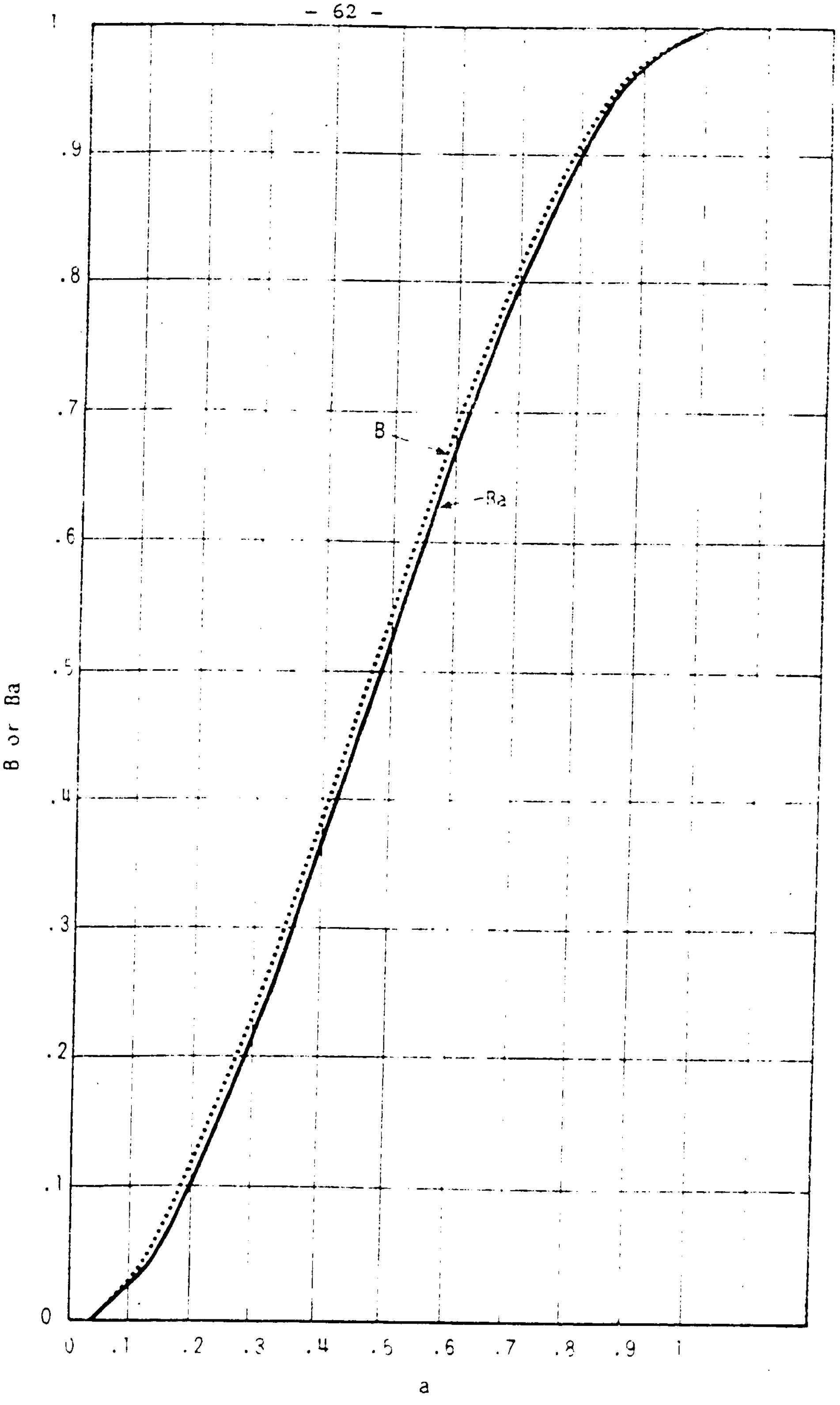


FIG. 24 - RESULTS FOR CASE 1.

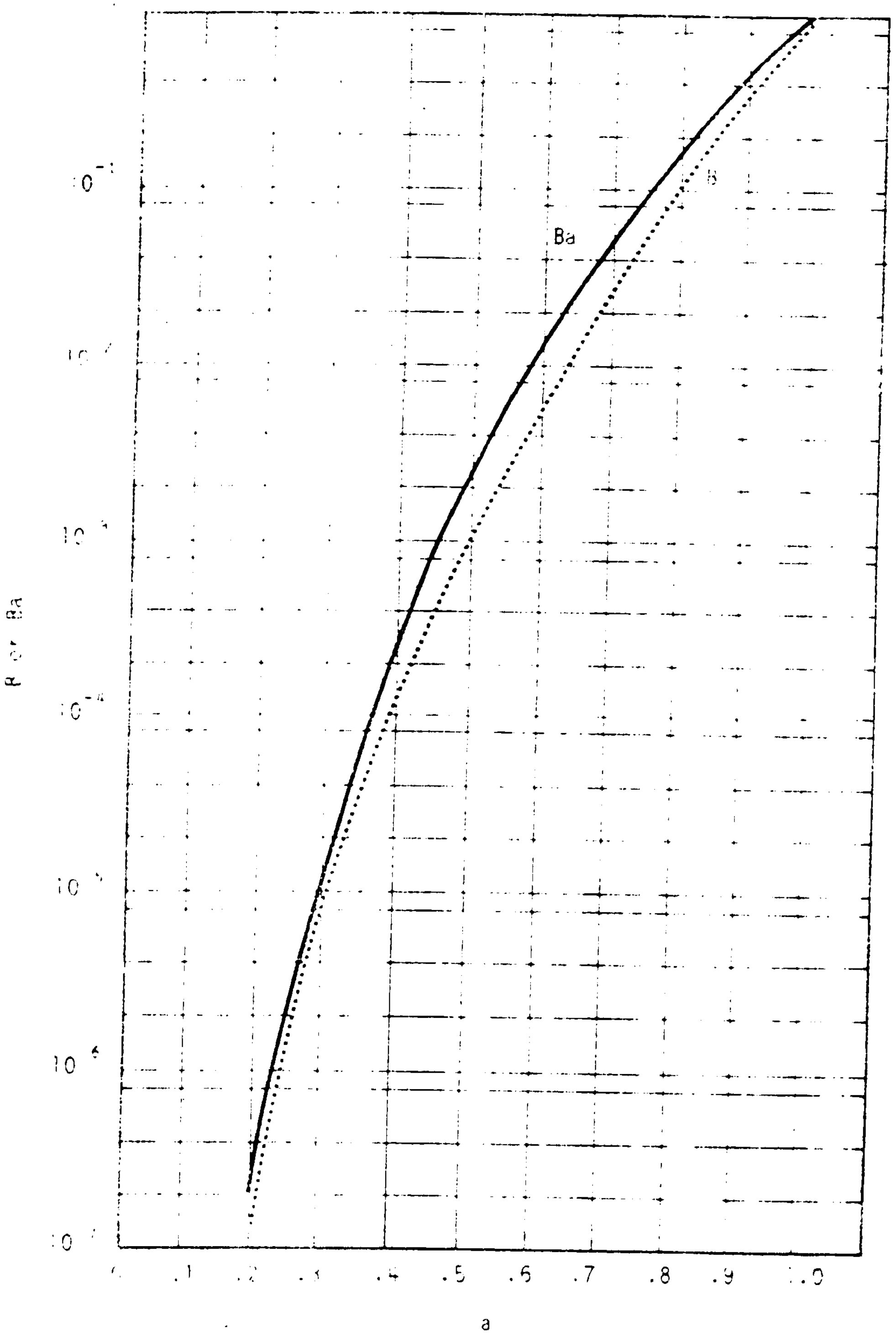


FIG. 25 - RESULTS FOR CASE 2.

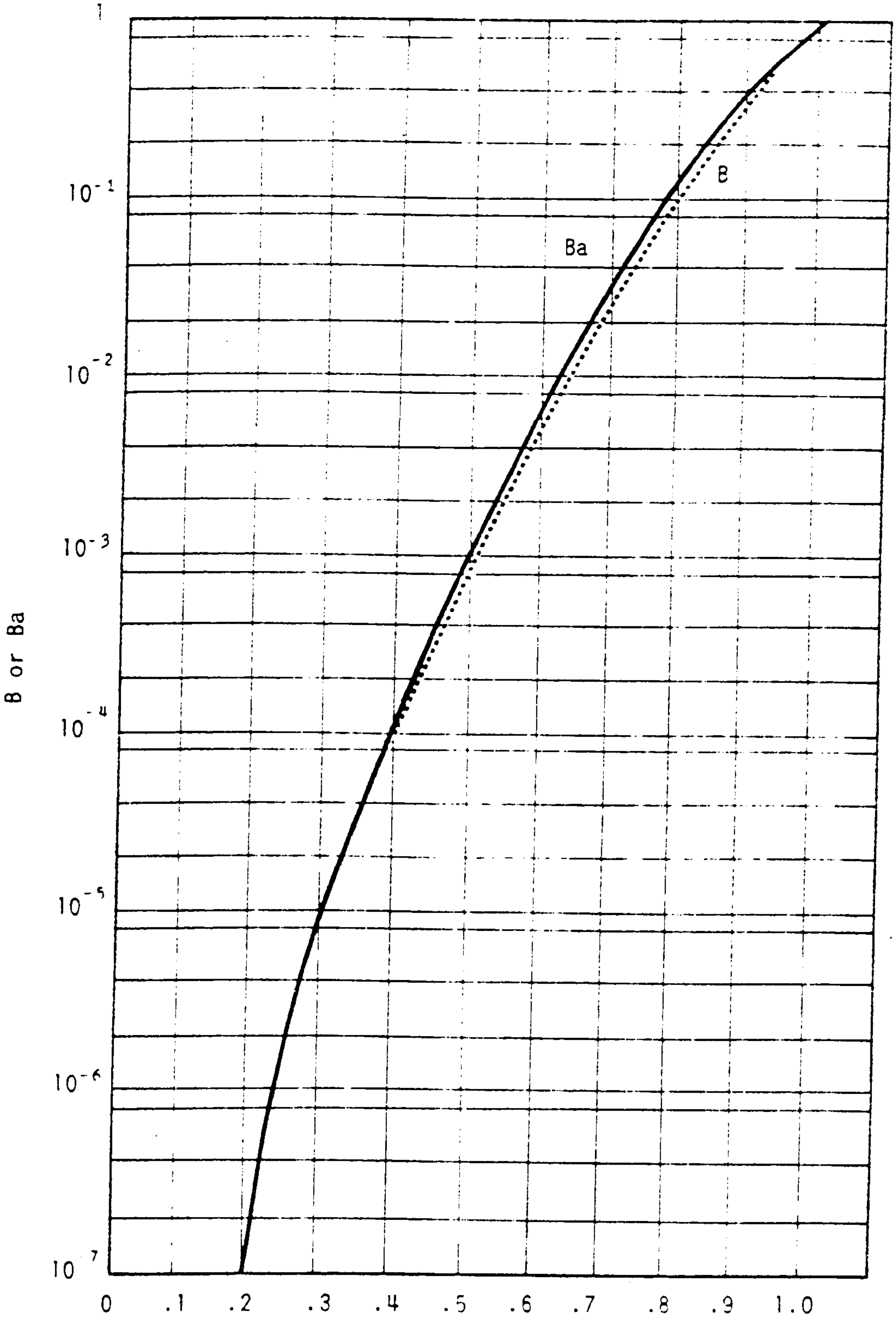


FIG. 26 - RESULTS FOR CASE 3.

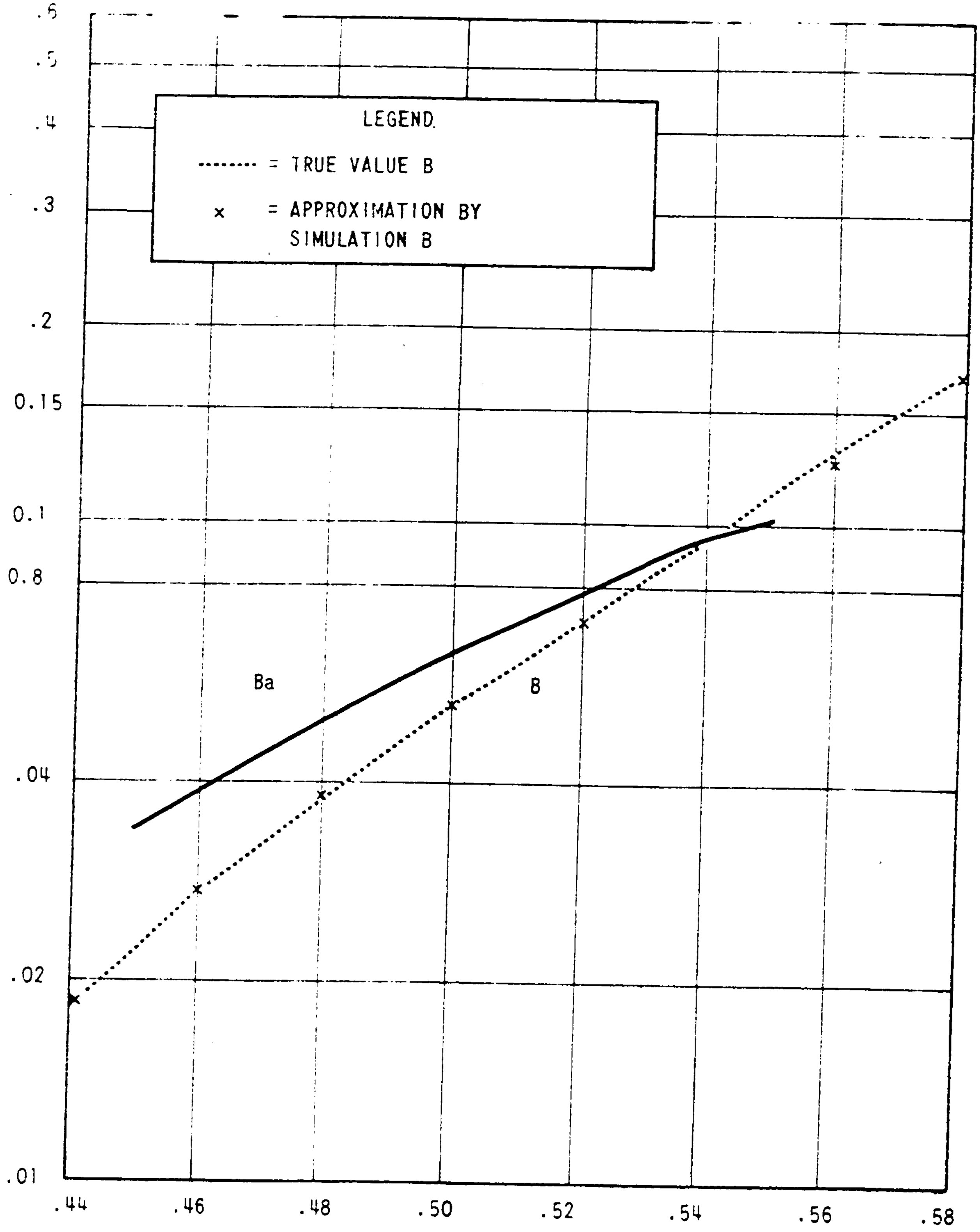


FIG. 27 - RESULTS FOR CASE 4.

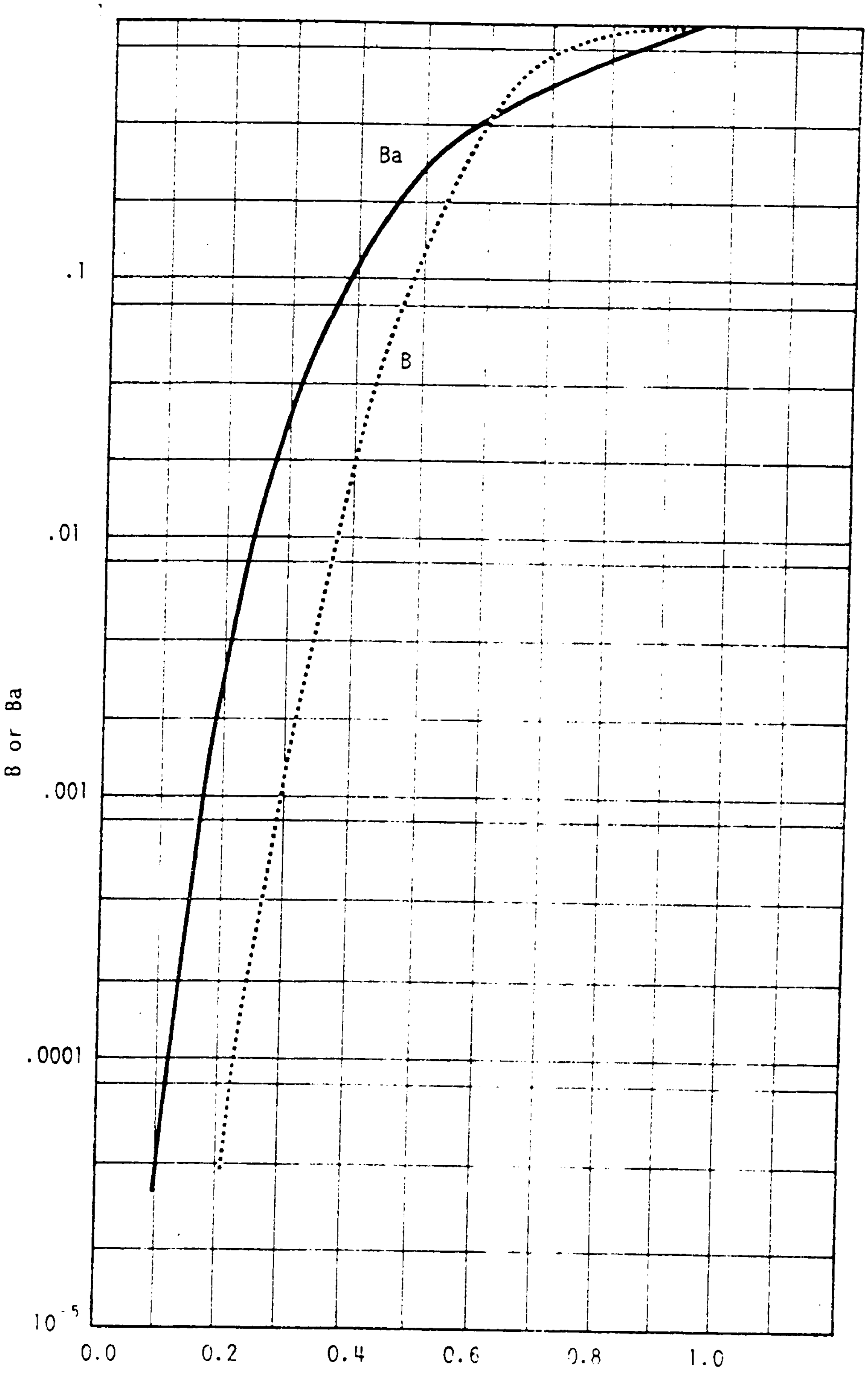


FIG. 28 - RESULTS FOR CASE 5.

3. Single Linking-Stage Network Loaded with Multitype Traffic

It should be noted that the Lee - LeGall model assumes that the state of a link (busy or idle) is independent of the states of the other links. In practice, this assumption is not strictly correct. As an example, in the case when an intra-office trunk is busy, two paths in the network (one of which connects a link to one of the two ports of the trunk, and the other line to the other port) must be busy simultaneously. Thus, when a link in a linking-stage is busy, it is certain that another link in the same stage must be busy. Fortunately, in a large switching network, the probability that the two input terminals involved belonging to the same switch is very small; and furthermore, the two output terminals connecting to the two ports of the intra-office trunk can be located in two different switches. Thus, the invalidity of the assumption of independent links due to this factor is insignificant for a large switching network.

However, for a small network where z=2 and W_1 is small (See Fig. 7), the fact that if a link is busy it causes other links to be busy simultaneously will make the assumption of independent links practically invalid. In order to have full access from the outlets of the switches in switching stage 1 to the trunks or service circuits, a network consisting of two switching stages may be connected as shown in Fig. 29, such that all output terminals connecting to the ports of a busy trunk appear busy at the same time. Assume that there are two types of traffic: one type causes one link busy per busy trunk (eg. incoming traffic) and another two per busy trunk (eg. intra-office traffic).

Then, the conditional blocking probability given that a trunk is available is different for the two types of traffic.

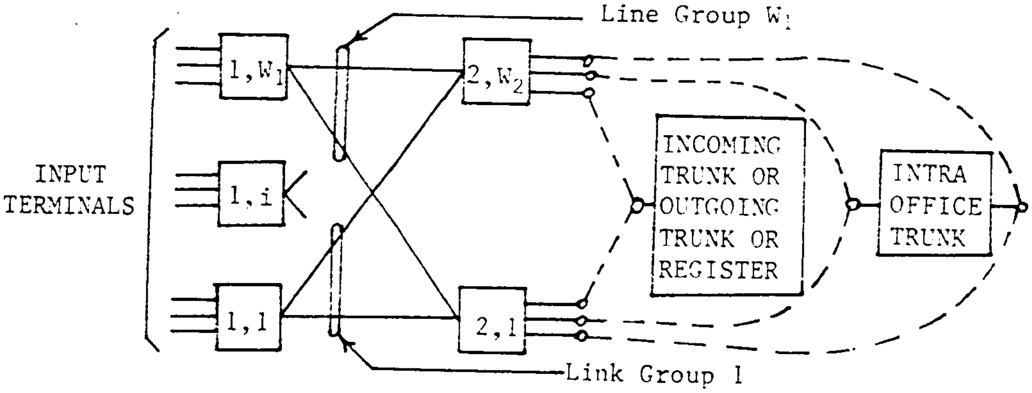


FIG. 29 - A NETWORK CONSISTING OF TWO SWITCHING STAGES.

For the type of traffic causing one link busy per busy trunk blocking occurs only if all outlets of the corresponding switch in switching stage 1 are busy. But, for the other type, blocking may occur even if there is still an idle outlet in the corresponding switch. This is possible if the two lines involved, say, in an intra-office call, are in the same switch. Thus, the application of the Lee - LeGall model for the blocking probability of such a network will result in an extremely poor approximation.

In order to give a more reasonable prediction for the blocking probability of such a network, a link group loaded with multitype traffic is investigated in the following.

3.1 Facts and Assumptions

Consider a link group consisting of $C(\geq 0)$ parallel links in a single linking stage network loaded with $m(\geq 1)$ types of traffic. A link can either be busy or idle (not busy) but not both at a time. When it is busy, it can be occupied by only one of the m types of traffic. When it is occupied by traffic type i, $T_i - 1$ other links are also occupied by traffic type i for the same interval of time forming a distinct group of T_i busy links. Of course, at a given time, there may be any number of such distinct groups of busy links in the link group. However, a busy link can only be in one of such distinct groups at a time.

Let these C links in the link group be numbered from 1 to C and the m types of traffic from 1 to m. Then, these two sets of numbers are $N = \{1, 2, \ldots, C\}$ and $M = \{1, 2, \ldots, m\}$ respectively. From these two sets, one can form a set $S = \{S_{i,j};\ i \in N \text{ and } j \in M\}$ such that each $S_{i,j} \in S$ itself is a set of subsets: $S_{i,j} = \{S_{i,j,k};\ k = 1, 2, \ldots, \binom{C-1}{T_j - 1}\}$ where each $S_{i,j,k} = \{n_\ell \in N;\ \ell = 1, 2, \ldots, T_j$ and i = one of these distinct numbers $n_\ell\}$. Thus, $S_{i,j}$ contains all the distinct groups each of which contains the number of T_j busy links in which i is occupied by type j traffic.

For example, in the case where there are $C = 3$ links in a link group loaded with $m = 2$ types of traffic such that $T_1 = 1$ and $T_2 = 2$, then $N = \{1, 2, 3\}$ and $M = \{1, 2\}$. Thus, $S = \{S_{1,1}, S_{1,2}, S_{2,1}, S_{2,2}, S_{3,1},$

$S_{3,2}\}$ where $S_{1,1} = \{1\}$, $S_{1,2} = \{\{1, 2\}, \{1, 3\}\}$, $S_{2,1} = \{2\}$, $S_{2,2} = \{\{2, 1\}, \{2, 3\}\}$, $S_{3,1} = \{3\}$ and $S_{3,2} = \{\{3, 1\}, \{3, 2\}\}$.

Also let each link i have associated with it a set of $\sum_{j=1}^{m} \binom{C-1}{T_j-1}$ random variables: $\{X_{i,j,k}(t);\ j\varepsilon M$, and $k = 1, 2, \ldots, \binom{C-1}{T_j-1}\}$ such that $\{X_{i,j,k}(t) = 1\} = E_{i,j,k}(t)$ represents that the ith link together with links $n_\lambda \varepsilon S_{i,j,k}$ are occupied by traffic type j at time t, and $\{X_{i,j,k}(t) = 0\} = \overline{E}_{i,j,k}(t)$ otherwise. Then, from the previous statements, the following are facts:

$$E_{i,j,k}(t) \cap \overline{E}_{i,j,k}(t) = \emptyset \tag{4.17}$$

where $\emptyset$ denotes a null event.

$$E_{i,j,k}(t) = \emptyset \qquad \text{if } T_j > C \tag{4.18}$$

$$E_{i,j,k}(t) = E_{i',j,k'}(t) \tag{4.19}$$

if $S_{i',j,k'} = S_{i,j,k}$.

$$E_{i,j,k}(t) \cap E_{i',j',k'}(t) = \emptyset \tag{4.20}$$

if (a) $i' = i$ but $j' \neq j$

(b) $i' = i$ and $j' = j$ but $k' \neq k$

(c) $i' \neq i$, and $i' \in S_{i,j,k}$ but $S_{i',j',k'} \neq S_{i,j,k}$ or

(d) $i' \neq i$, and $S_{i',j',k'} = S_{i,j,k}$ but $j' \neq j$.

In the following, $P\{\cdot\}$ denotes the probability of the occurrence of event $\{\cdot\}$. In order to derive useful results, it is assumed that

(a) at any given time t, for each traffic type $j \in M$,

$$P\{E_{i,j,k}(t)\} = P\{E_{i,j,k'}(t)\} \tag{4.21}$$

(b) there exists steady-state probabilities such that each

$$\lim_{t \to \infty} P\{E_{i,j,k}(t)\} = \pi_j . \tag{4.22}$$

The link group i is said to be blocked for traffic type j if the group can not handle any more traffic of the j type. Traffic type j can not be handled by the group only if there are less than T_j links idle at the service request time. Therefore, it is appropriate to define the blocking probability of the link group for traffic type j as the probability that there are $C - T_j + 1$ or more links busy in the group.

Based upon these assumptions and the above facts, it is required to find the link group blocking probabilities for various types of traffic.

3.2 The Link Occupancy

The link group blocking probabilities for various types of traffic are usually expressed in terms of the link occupancies. A link occupancy a is defined as the ratio of the total time that the link is busy to the specified measuring period of time. Thus, a link occupancy a_j due to traffic type j may be defined as the ratio of the total time that the link is busy due to traffic type j to the specified measuring period of time. For type j, under steady-state condition, is

$$a_j = \lim_{t\to\infty} P\{ \bigcup_{S_{i,j,k} \in S_{i,j}} E_{i,j,k}(t) \}$$

$$= \lim_{t\to\infty} \sum_{S_{i,j,k} \in S_{i,j}} P\{E_{i,j,k}(t)\}$$

$$= \lim_{t\to\infty} \binom{C-1}{T_j-1} P\{E_{i,j,k}(t)\}$$

$$= \binom{C-1}{T_j-1} \alpha_j \qquad (4.23)$$

because of (4.20) and assumptions. For the same reason, its overall occupancy is

$$a = \lim_{t \to \infty} P \{ \bigcup_{j \in M} \bigcup_{S_{i,j,k} \in S_{i,j}} E_{i,j,k}(t) \}$$

$$= \sum_{j=1}^{m} \binom{C-1}{T_j - 1} \alpha_j$$

$$= \sum_{j=1}^{m} a_j \qquad (4.24)$$

Because of the assumptions, each link in the group has the same overall occupancy as well as the same occupancies due to various types of traffic.

It is interesting to note the limiting values of the link occupancies due to various types of traffic, a_j's. Of course, the minimum value of each a_j is zero since there is no physical realization of a negative value of a_j. The maximum value of a_j is governed by (4.24), the overall link occupancy a. Although a link can be busy, at most, all the time in a given measuring period of time, it may be impossible sometimes for the maximum value of a to be unity. This may become obvious by considering the following simple case.

Suppose that a link group consisting of 3 links is loaded with a single type 1 traffic such that $T_1 = 2$. Fig. 30 illustrates this case in a particular way. The

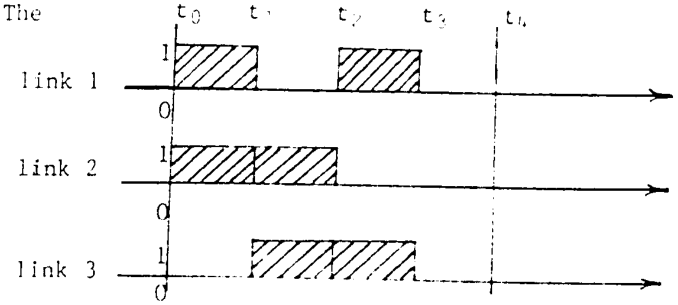


Fig. 30. Link occupancy for the case where C=3, and T_1=2.

busy condition of a link is represented by 1 and idle by 0. In this case, there are a total of 3 busy periods for the link group in the given time interval (t_0, t_4): links 1 and 2 are busy together during (t_0, t_1); links 2 and 3 during (t_1, t_2); and links 1 and 3 during (t_2, t_3). Here, $t_1 - t_0 = t_2 - t_1 = t_3 - t_2 = \alpha_1(t_4 - t_0)$ so that assumptions are satisfied. Since $T_1 = 2$ (i.e. the links are busy in pairs), the third link cannot become busy if the two other links are already busy. Therefore, a link can be busy, at most, $^2/3$ of the given time interval. This occurs when $t_4 - t_3 = 0$. In other words, the maximum value of $a = a_1$ is $^2/3$ which is less than unity.

In general, the maximum value of the link occupancy for a link group consisting of C links loaded with a single type j traffic may be determined as follows: Let max $(x|y)$ be the maximum value of x for a given condition of y. Then, the maximum value of the link occupancy required is the same as max $(a_j | \sum_{\substack{r=1 \\ r \neq j}}^{m} a_r = 0)$. Let t be the time of the measuring period. Then, $a_j t$ is the total amount of time that a link is busy. For a link group consisting of C links, the cumulative amount of busy time of all links is $Ca_j t$ because of the assumptions. If W denotes the maximum possible integer of W, the maximum possible number of links busy due to traffic type j at a time is $\left\lfloor \frac{C}{T_j} \right\rfloor T_j$. Therefore, the cumulative time $Ca_j t$ must spread at least over an actual interval of $\frac{Ca_j t}{\left\lfloor \frac{C}{T_j} \right\rfloor T_j}$. Thus, $\frac{Ca_j t}{\frac{C}{T_j} T_j}$ is the minimum interval of time of the measuring period. Therefore, the maximum link occupancy required is

$$\max\left(a_j \;\middle|\; \sum_{\substack{r=1\\ r\neq j}}^{m} a_r = 0\right) = \frac{a_j t}{C a_j t / \frac{C}{T_j}\, T_j}$$

$$= \frac{T_j}{C} \left| \frac{C}{T_j} \right. . \qquad (4.25)$$

The following obvious relations are of great interest in practice:

$$\max\left(a_j \;\middle|\; T_j = 1;\; \sum_{\substack{r=1\\ r\neq j}}^{m} a_r = x\right) = 1 - x \qquad (4.26)$$

$$\max\left(a_j \;\middle|\; \sum_{\substack{r=1\\ r\neq j}}^{m} a_r = x;\; k \prod_{r=1}^{m} T_r = C \text{ for}\right.$$

$$\left. k = 1, 2, \ldots\right) = 1 - x \qquad (4.27)$$

and

$$\max\left(a_j \;\middle|\; a_s = x \text{ for } T_s = 1;\; kT_j \neq C \text{ for } k = 1, 2, \ldots; \text{ and}\right.$$

$$\sum_{\substack{r=1 \\ r \neq j \neq s}}^{m} a_r = 0) = \min \left(1 - x, \left[\frac{C}{T_j}\right]^{\frac{T_j}{C}}\right) \tag{4.28}$$

where min $(y_1, y_2, \ldots)$ equals the smallest value of y_i in the set $\{y_i; i = 1, 2, \ldots\}$. Since the maximum values of a_j for other given conditions of a_r's where $r \neq j$ depend greatly on a_r's, these values may be derived directly as the cases occur.

3.3 The Blocking Probability

Let $Y(t)$ be the number of links idle at time t in a link group consisting of C links. Then, by definition, the link group blocking probability $B(a_j)$ for traffic type j under steady-state conditions is

$$B(a_j) = \lim_{t \to \infty} P\left\{ \bigcup_{y=0}^{T_j - 1} \{Y(t) = y\} \right\}$$

$$= \lim_{t \to \infty} \sum_{y=0}^{T_j - 1} P\{Y(t) = y\} \qquad (4.29)$$

But $Y(t) = y$ implies that $C-y$ other links must be busy at the same time. Furthermore, these idle and busy links may be in any order. If a specific order such that the links numbered from 1 to y are idle and $y+1$ to C busy at time t in a link group consisting of C links has a probability $Q_t(y; C-y)$, then, under steady-state conditions and assumptions,

$$\lim_{t \to \infty} P\{Y(t) = y\} = \binom{C}{y} \lim_{t \to \infty} Q_t(y; C-y)$$

$$= \binom{C}{y} Q(y; C-y) \qquad (4.30)$$

where $Q(y; C-y)$ is the corresponding steady-state probability that y links are idle and $C-y$ links busy.

At time t, a link is busy if event $\bigcup_{j \in M} \left\{ \bigcup_{S_{i,j,k} \in S_{i,j}} E_{i,j,k}(t) \right\}$

occurs and idle if $\bigcap_{j \in M} \{ \bigcap_{S_{i,j,k} \in S_{i,j}} \overline{E}_{i,j,k}(t) \}$ occurs. Note that no link idle implies all links busy. Therefore

$$
\begin{aligned}
Q_t(y;\, C-y) &= P\{ \bigcap_{i=1}^{C} \{ \bigcup_{j \in M} \{ \bigcup_{S_{i,j,k} \in S_{i,j}} E_{i,j,k}(t) \}\}\} && \text{if } y = 0 \\
&= P\{ \bigcap_{i=1}^{y} \{ \bigcap_{j \in M} \bigcap_{S_{i,j,k} \in S_{i,j}} \overline{E}_{i,j,k}(t) \}\} \bigcap_{i=y+1}^{C} \\
&\qquad \{ \bigcup_{j \in M} \{ \bigcup_{S_{i,j,k} \in S_{i,j}} E_{i,j,k}(t) \}\}\} && \text{if } 0<y<C \\
&= P\{ \bigcap_{i=1}^{C} \{ \bigcap_{j \in M} \{ \bigcap_{S_{i,j,k} \in S_{i,j}} \overline{E}_{i,j,k}(t) \}\}\} && \text{if } y = C
\end{aligned}
\tag{4.31}
$$

Thus, combining (4.29) to (4.31), the facts and assumptions outlined in Section 3.1, and the relations of link occupancies in Section 3.2, one can obtain the required link group blocking probability for a given type j traffic in terms of link occupancies.

However, when C>>1 and m>1, it is rather difficult, in general, to express the conditonal probabilities in terms of link occupancies. For practical purposes, when such difficulty is encountered, it is suggested to make approximation by using the *following additional assumption*:

The random variables $X_{i,j,k}(t)$'s are mutually independent except the facts outlined in Section 3.1

When $a_j = a_{j\max} = \max(a_j \mid \sum_{\substack{k=1 \\ k\neq j}}^{m} a_k)$, $B(a_j) = B(a_{j\max})$ should be unity since the link group can handle at most C $a_{j\max}$ traffic type j. But, with this additional assumption, it is possible that $B(a_{j\max}) \neq 1$. For this reason, it is suggested to normalize the link group blocking probability for a given traffic type j whenever this additional assumption is required such that the approximate link group blocking probability B_j for traffic type j is

$$B_j = \frac{B(a_j)}{B(a_{j\max})} \tag{4.32}$$

Thus, if $B(a_{j\max}) = 1$, the approximate and the exact link group blocking probability is identical.

To express B_j in terms of link occupancies, it is necessary to express Q(y; C-y) in terms of link occupancies. First, consider the case when all links are busy, ie. y=0. From (4.31),

$$Q_t(0;C) = P\{ \bigcap_{i=1}^{C-1} \{ \bigcup_{j\varepsilon M} \{ \bigcup_{S_{i,j,k}\varepsilon S_{i,j}} E_{i,j,k}(t)\}\} |$$

$$\bigcup_{j_C\varepsilon M} \{ \bigcup_{S_{C,j_C,k}\varepsilon S_{C,j_C}} E_{C,j_C,k_C}(t)\}\}$$

$$P\{ \bigcup_{j_C\varepsilon M} \{ \bigcup_{S_{C,j_C,k_C}\varepsilon S_{C,j_C}} E_{C,j_C,k_C}(t)\}\}$$

$$= \sum_{\substack{j_C=1\\ T_{j_C}<C}}^{M} \sum_{k_C=1}^{\binom{C-1}{T_{j_C}-1}} P\{ \bigcap_{i=1}^{C-1} \{ \bigcup_{j\varepsilon M} \{ \bigcup_{S_{i,j,k}\varepsilon S_{i,j}} E_{i,j,k}(t)\}\}$$

$$|E_{C,j_C,k_C}(t)\}\ P\{E_{C,j_C,k_C}(t)\}$$

$$= \sum_{\substack{j_C=1\\ T_{j_C}<C}}^{m} \sum_{k_C=1}^{\binom{C-1}{T_{j_C}-1}} P\{E_{C,j_C,k_C}(t) \bigcap_{i=1}^{C-1} \{ \bigcup_{j\varepsilon M}$$

$$\{ \bigcup_{S_{i,j,k}\varepsilon S_{i,j}} E_{i,j,k}(t)\}\}\}$$

$$+ \sum_{\substack{j_C=1 \\ T_{j_C}=C}}^{m} P\{E_{C,j_C,1}(t)\} \tag{4.33}$$

Continuing this approach, (4.33) can be expanded without using the additional assumption. However, it can be greatly simplified if this assumption is employed. Since

$$P\{ \bigcap_{i=1}^{C-T_{jC}} \{ \bigcup_{j \varepsilon M} \{ \bigcup_{\substack{S_{i,j,k} \varepsilon S_{i,j} \\ n_\lambda \varepsilon S_{C,jC,1} \not\subset S_{i,j,k}}} E_{i,j,k}(t)\}\}\}$$

is essentially the probability that all links in the link group consisting of C-T_{jC} links are busy at time t, ie. $Q_t(0;C-T_{jC})$, it then follows that, from (4.33) and using the additional assumptions,

$$Q_t(0;C) = \sum_{\substack{j=1 \\ T_j<C}}^{m} \binom{C-1}{T_j-1} P\{E_{C,j,1}(t)\} \, Q_t(0;C-T_j)$$

$$+ \sum_{\substack{j=1 \\ T_j=C}}^{m} P\{E_{C,j,1}(t)\}. \tag{4.34}$$

Another way of writing (4.34) is

$$Q_t(0;C) = \sum_{j=1}^{m} \binom{C-1}{T_j-1} \, P\{E_{C,j,1}(t)\} \, Q_t(0;C-T_j) \tag{4.35}$$

If one defined

$$\begin{aligned} Q_t(0;C-T_j) &= 0 \quad \text{for } C<T_j \\ &= 1 \quad \text{for } C=T_j \end{aligned} \tag{4.36}$$

Thus, to evaluate $Q_t(0;C)$, one only needs to carry out the recurrent relation of (4.35) until either one of the conditions (but not both) in (4.36) is reached. Under steady-state conditions and combining these recurrent relations into one expression, one obtains that,

$$\begin{aligned} Q(0;C) &= \lim_{t\to\infty} Q_t(0;C) \\ &= 0 \qquad\qquad\qquad\qquad\qquad\qquad \text{if } C<0 \\ &= 1 \qquad\qquad\qquad\qquad\qquad\qquad \text{if } C=0 \\ &= \sum_{j=1}^{m} \binom{C-1}{T_j-1} \alpha_j \, Q(0;C-T_j) \quad \text{if } C>0 \end{aligned} \tag{4.37}$$

Next, consider the case when all links are idle, ie. $y = C$. From (4.31), by simply apply in De Morgan's law [1].

$$Q_t(C;0) = P\{ \bigcap_{i=1}^{C} \{ \bigcap_{j\varepsilon M} \{ \bigcap_{S_{i,j,k}\varepsilon S_{i,j}} \bar{E}_{i,j,k}(t)\}\}\}$$

$$= 1 - P\{ \bigcup_{i=1}^{C} \{ \bigcup_{j\varepsilon M} \{ \bigcup_{S_{i,j,k}\varepsilon S_{i,j}} E_{i,j,k}(t)\}\}\}$$

$$= 1 - \sum_{i=1}^{C} P\{ \bigcup_{j\varepsilon M} \{ \bigcup_{S_{i,j,k}\varepsilon S_{i,j}} E_{i,j,k}(t)\}\}$$

$$+ \sum_{\substack{i_1,i_2=1 \\ i_1<i_2}}^{C} P\{ \bigcap_{x=1}^{2} \{ \bigcup_{j_x\varepsilon M} \{ \bigcup_{S_{i_x,j_x,k_x}\varepsilon S_{i_x,j_x}} E_{i_x,j_xk_x}(t)\}\}\}$$

$$- \sum_{\substack{i_1,i_2,i_3=1 \\ i_1<i_2<i_3}}^{C} P\{ \bigcap_{x=1}^{3} \{ \bigcup_{j_x\varepsilon M} \{ \bigcup_{S_{i_x,j_x,k_x}\varepsilon S_{i_x,j_x}} E_{i_x,j_x,k_x}(t)\}\}\}$$

$$+ \ . \ . \ . \qquad (4.38)$$

It is rather difficult to express (4.38) in terms of link occupancies explicitly. However, with more specific informations about the parameters, this difficulty may be eliminated. The following are some special cases:

(a) when C=1, then

$$Q_t(1;0) = 1 - \sum_{\substack{j=1 \\ T_j=1}}^{m} P\{E_{1,j,1}(t)\}$$

so that

$$Q(1;0) = 1 - \sum_{\substack{j=1 \\ T_j=1}}^{m} \alpha_j$$

$$= 1 - \sum_{\substack{j=1 \\ T_j=1}}^{m} a_j \tag{4.38a}$$

(b) When C=2, then

$$Q_t(2;0) = 1 - 2 \sum_{\substack{j=1 \\ T_j \leq 2}}^{m} P\{E_{1,j,1}(t)\}$$

$$+ \sum_{\substack{j_1=1 \\ T_{j_1} \leq 2}}^{m} P\{E_{1,j_1,1}(t) \cap \{ \bigcup_{j_2 \varepsilon M} \{ \bigcup_{S_{2,j_2,k_2} \varepsilon S_{2,j_2}} E_{2,j_2}(t)\}\}\}$$

$$= 1 - 2 \sum_{\substack{j=1 \\ T_j \leq 2}}^{m} P\{E_{1,j,1}(t)\}$$

$$+ \sum_{\substack{j_1=1 \\ T_{j_1}=1}}^{m} P\{E_{1,j_1,1}(t)\} \sum_{\substack{j_2=1 \\ T_{j_2}=1}}^{m} P\{E_{2,j_2,1}(t)\}$$

$$+ \sum_{\substack{j_1=1 \\ T_{j_1}=2}}^{m} P\{E_{1,j_1,1}(t)\}$$

so that

$$Q(2;0) = 1 - 2\,\Big(\sum_{\substack{j=1\\T_j=1}}^{m} \alpha_j + \sum_{\substack{j=1\\T_j=2}}^{m} \alpha_j\Big)$$

$$+ \sum_{\substack{j_1=1\\T_{j_1}=1}}^{m} \alpha_{j_1} \Big(\sum_{\substack{j_2=1\\T_{j_2}=1}}^{m} \alpha_{j_2}\Big) + \sum_{\substack{j=1\\T_j=2}}^{m} \alpha_j$$

$$= 1 - 2 \sum_{\substack{j=1\\T_j=1}}^{m} \alpha_j - \sum_{\substack{j=1\\T_j=2}}^{m} \alpha_j + \Big(\sum_{\substack{j=1\\T_j=1}}^{m} \alpha_j\Big)^2$$

$$= 1 - 2 \sum_{\substack{j=1\\T_{j=1}}}^{m} a_j - \sum_{\substack{j=1\\T_{j=2}}}^{m} a_j + \Big(\sum_{\substack{j=1\\T_j=1}}^{m} a_j\Big)^2 \qquad (4.38b)$$

(c) When $T_j=1$ for all $j \varepsilon M$, then

$$Q_t(C;0) = 1 - C \sum_{j=1}^{m} P\{E_{1,j,1}(t)\}$$

$$+ \binom{C}{2} \sum_{j_1=1}^{m} P\{E_{1,j_1,1}(t)\} \Big(\sum_{j_2=1}^{m} P\{E_{2,j_2,1}(t)\}\Big)$$

$$- \binom{C}{3} \sum_{j_1=1}^{m} P\{E_{1,j_1,1}(t)\} \Big(\sum_{j_2=1}^{m} P\{E_{2,j_2,1}(t)\}$$

$$\Big(\sum_{j_3=1}^{m} P\{E_{3,j_3,1}(t)\}\Big)\Big)$$

$$+ \ldots$$

so that

$$Q(C;0) = 1 - C \sum_{j=1}^{m} \alpha_j$$

$$+ \binom{C}{2} \left(\sum_{j=1}^{m} \alpha_j\right)^2$$

$$- \binom{C}{3} \left(\sum_{j=1}^{m} \alpha_j\right)^3 + \ldots$$

$$= \sum_{k=0}^{C} \binom{C}{k} (-a)^k$$

$$= (1-a)^C$$

etc. (4.38c)

Finally, consider the case when some links are idle and some busy, i.e. $0<y<C$. From (4.31),

$$Q_t(y,C-y) = P\{ \bigcap_{i=1}^{y} \{ \bigcap_{j \varepsilon M} \{ \bigcap_{\substack{S_{i,j,k} \varepsilon S_{i,j} \\ n_\lambda \varepsilon S_{i',j',k'} \not\subset S_{i,j,k}}} \overline{E}_{i,j,k}(t)\}\}$$

$$\bigcap_{i'=y+1}^{C} \{ \bigcup_{j'\varepsilon M} \{ \bigcup_{\substack{S_{i',j',k'}\varepsilon S_{i',j'} \\ n_{\lambda}\varepsilon S_{i,j,k} \not\subset S_{i',j',k'}}} E_{i',j',k'}(t)\}\}\}$$

$$= Q_t(y;0) \cdot Q_t(0;C-y) \qquad (4.39)$$

because of the facts and the assumptions. Under steady-state conditions, (4.39) becomes

$$Q(y;C-y) = Q(y;0) \cdot Q(0;C-y) \qquad (4.40)$$

In summary, combining (4.29) and (4.30), the link group blocking probability for traffic type j is

$$B(a_j) = \sum_{y=0}^{T_j-1} \binom{C}{y} Q(y;C-y) \qquad (4.41)$$

where Q(y;C-y) may be obtained by taking the limit of (4.31). For approximation, (4.32) may be used. In this case, $B(a_j)$ and $B(a_{jmax})$ are obtained from (4.41) with Q(y;C-y) obtained from (4.37), (4.38) and (4.40).

3.4 Sample Calculations

In this section, simple examples are given below to demonstrate the usefuleness of the formulae outlined in section 3.3 of this chapter.

Example 1:

A link group consisting of C links is loaded with a single type of traffic A_1, which requires the service of one link per request; ie. $T_1 = 1$. If the traffic is evenly loaded over the links, what is the probability that exactly y $(0 \leq y \leq C)$ links are busy?

Solution: Since the traffic A_1 is evenly loaded over the links, each link has an occupancy of

$$a_1 = \frac{A_1}{C}$$

The probability of exactly y links busy is (4.30). By (4.23), (4.37), (4.38c) and (4.40),

$$\begin{aligned} Q(0;C) &= {a_1}^{C} \\ Q(C;0) &= (1 - a_1)^{C} \\ \text{and} \quad Q(C-y;y) &= Q(C-y;0)\ Q(0;y) \\ &= (1 - a_1)^{C-y}\ {a_1}^{y} \qquad \text{if } 0<y<C. \end{aligned}$$

Therefore, the probability of exactly y links busy is,

$$\lim_{t \to \infty} P\{Y(t) = C-y\} = \binom{C}{C-y} (1-a_1)^{C-y}\ {a_1}^{y}$$

which is the Benroulli distribution [1,17] as expected.

Example 2:

A link group conssiting of C links is loaded with m types of traffic A_1, A_2, ..., A_m; all of which require the service of one link per request; i.e. $T_1 = T_2 = \ldots = T_m = 1$. If each traffic type is equally loaded over the links such that each link has an occupancy of

$$a_i = \frac{A_i}{C} \qquad i = 1, 2, \ldots, m;$$

what is the link group blocking probability for each type of traffic?

Solution: Because $T_1 = T_2 = \ldots = T_m = 1$, it follows from (4.23)

$$a_j = \alpha_j \qquad \text{for } j = 1, 2, \ldots, m.$$

By (4.26),

$$a_{j\,\max} = 1 - \sum_{\substack{i=1 \\ i \neq j}}^{m} a_i .$$

From (4.24), (4.37) and (4.41),

$$B(a_j) = Q(0{:}C) = \Big(\sum_{i=1}^{m} a_i\Big)^C$$

$$= a^C$$

But, when $a_j = a_{j\max}$, $\sum_{i=1}^{m} a_i = 1$. Therefore

$$B(a_{j\max}) = 1$$

By (4.32),

$$B_j = B(a_j) = a^C$$

Thus, the approximate link group blocking probability is the same as the link group blocking probability for each given traffic type j. This agrees with Lee - LeGall model when the occupancy of every link in the group is the same.

Example 3:

A link group consisting of C links is loaded with a single type of traffic A_1 which requires the service of x links per request; ie. $T_1 = x$. Assuming that the traffic is evenly loaded over the links such that each link occupancy is

$$a_1 = \frac{xA_1}{C}$$

What is the probability that all links in the group are busy?

Solution: The probability that all links in the group are busy is, by (4.30) and (4.37),

$$\begin{aligned}\lim_{t\to\infty} P\{Y(t) = 0\} &= Q(0;C) \\ &= 0 & \text{if } C<0 \\ &= 1 & \text{if } C=0 \\ &= a_1 Q(0;C-x) & \text{if } C>0\end{aligned}$$

Clearly, then, for C>0,

$$\begin{aligned}\lim_{t\to\infty} P\{Y(t) = 0\} &= \frac{a_1{}^{C/x}}{\{\binom{C-1}{x-1}\}^{C/x}} \prod_{k=0}^{\frac{C}{x}-1} \binom{C-kx-1}{x-1} & \text{if C is the multiple of x} \\ &= 0 & \text{if C is not the multiple of x}\end{aligned}$$

This is reasonable because of the following: It is expected that it is impossible for all the links to be busy in the group at any time if C is not a multiple of x since there are at least $(C-x\left[\frac{C}{x}\right])$ links idle all the time. Furthermore, for $^C/_x$ = an integer, it follows from (4.27), $a_{1\,\max} = 1$. By (4.32), (4.37), (4.40), (4.41) and the expressions above, the link group blocking probability is

$$B_j = (a)^{C/x}$$

as expected if $C>0$ and C is the multiple of x.

Example 4:

A link group consisting of 3 links is loaded with two types of traffic A_1 and A_2 such that $T_1 = 1$ and $T_2 = 2$. If each of these traffic is evenly loaded over the links such that the occupancies are

$$a_1 = \frac{A_1}{3}$$

$$\text{and} \qquad a_2 = \frac{2A_2}{3},$$

what is the link group blocking probability for each type of traffic:

Solution: First, consider type 1 traffic. By (4.23), (4.37) and (4.41),

$$B(a_1) = Q(0;3)$$

$$= \sum_{j=1}^{2} \binom{2}{T_i-1} \alpha_i \, Q(0;3-T_i)$$

$$= a_1 \sum_{i=1}^{2} \alpha_i \, Q(0;2-T_i) + a_2 \sum_{i=1}^{2} \alpha_i \, Q(0;1-T_i)$$

$$= a_1 \, (\alpha_1{}^2+\alpha_2) + a_2\alpha_1$$

$$= a_1{}^3 + \frac{3}{2} a_1 a_2$$

By (4.26)

$$a_{1\ max} = 1-a_2$$

Therefore,

$$B(a_{1\ max}) = (1-a_2)^3 + \frac{3}{2}\,a_2(1-a_2)$$

Thus, the approximate link group blocking probability for traffic type 1 is

$$B_1 = \frac{a_1{}^3 + \frac{3}{2}\,a_1a_2}{(1-a_2)^3 + \frac{3}{2}\,a_2(1-a_2)}$$

Next, consider type 2 traffic. Similarly,

$$B(a_2) = \sum_{y=0}^{1} \binom{3}{y}\ Q(y;3-y)$$

$$= Q(0;3) + 3Q(1;2)$$

But, by (4.38a), (4.37) and (4.40),

$$Q(1;2) = Q(1;0) \cdot Q(0;2)$$

$$= (1-a_1) \sum_{i=1}^{2} \alpha_i\ Q(0;2-T_i)$$

$$= (1 - a_1)\ (\alpha_1{}^2 + \alpha_2)$$

$$= (1 - a_1)\ (a_1{}^2 + \frac{a_2}{2})$$

Therefore

$$B(a_2) = a_1{}^3 + \frac{3}{2}\, a_1 a_2 + 3(1 - a_1)\ (a_1{}^2 + \frac{a_2}{2})$$

By (4.28),

$$a_{2\ \max} = \min\ (1 - a_1,\ \frac{2}{3})$$

Therefore,

$$\begin{aligned} B(a_{2\ \max}) = {} & a_1{}^3 + \frac{3}{2}\, a_1 \min\ (1 - a_1,\ \frac{2}{3}) \\ & + 3(1 - a_1) \cdot (a_1{}^2 \\ & + \tfrac{1}{2} \min(1 - a_1,\ \frac{2}{3})) \end{aligned}$$

Thus, the approximate link group blocking probability for traffic type 2 is

$$B_2 = \frac{B\ (a_2)}{B\ (a_{2\ \max})}$$

It is interesting to note that when $a_1 = \frac{1}{3}$ and $a_2 = \frac{2}{3}$, according to (4.25), it is expected that the link group is busy all the time. But in this case, both $B(a_1)$ and $B(a_2)$ do not equal unity. However, the normalized link group blocking probabilities are $B_1 = B_2 = 1$ as it should be.

Chapter V

Waiting Time Distributions

1. Introduction

To estimate the overall grade of service for delayed calls by means of the traffic flow graph, it is important to estimate the link delay distributions as exact as possible. Each link in the traffic flow graph represents a delay system of a call delayed for service in the exchange. Thus, the traffic flow graph represents essentially the network of delay systems through which the call is processed.

Usually, a delay system is defined as a full availability waiting system, that is a service system in which a request is allowed to wait if it cannot be served immediately upon its arrival. It has an input with M sources S_1, S_2, ..., S_M where $1 \leq M \leq \infty$, and an output which may be regarded as a source for the same and/or other delay systems. This is illustrated in Figure 31. A source may itself be regarded as the output of a delay system. If a call (or request) appears in source S_i of the delay system D_j, at time T_0, its service will begin at time $T_1 \geq T_0$ in accordance with some rule of selection in D_j, called the queue-discipline of D_j, and it will appear at the output of D_j when its service is completed at time $T_2 \geq T_1$. The time intervals $(T_1 - T_0)$, $(T_2 - T_1)$ and $(T_2 - T_0)$ are called respectively the queuing time, the service time and the waiting time of the

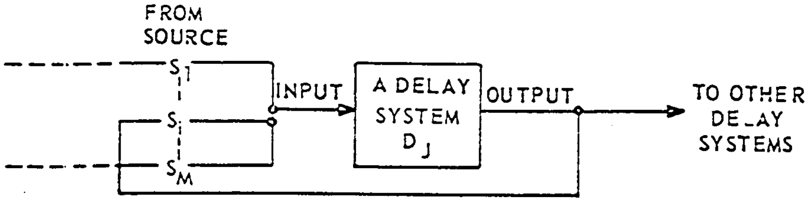


FIG. 31 - A DELAY SYSTEM

request from S_i with respect to D_j as shown in Figure 32. A delay system may be described by specifying the following three characteristics: the input process, the queue discipline and the service mechanism [8].

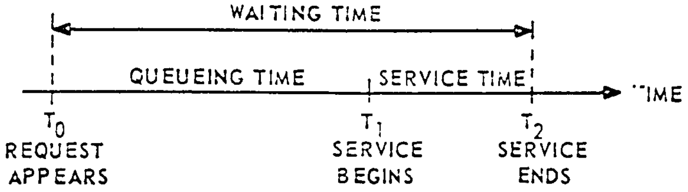


FIG. 32 - RELATIONS BETWEEN QUEUEING, SERVICE AND WAITING TIMES.

The <u>input process</u> of a delay system D_j can be described by a sum of M random variables,

$$X_j(t) = \sum_{i=1}^{M} x_i(t),$$

where each $x_i(t)$ represents the number of requests from source S_i of the system in the time interval (0,t). The <u>queue discipline</u> is the method by which a request is selected for service from all the requests waiting for service. It may be characterized by a set of conditional probabilities described in the next section. The service mechanism of a delay system can be described by the probability distribution of its service time T, $P\{T \leq t\}$, and the number of servers.

2. Methods of Determining Waiting Time Distribution

There are at least three methods which can be used for determining the queuing time distribution of a delay system, and hence its waiting time distribution if its service time distribution is known. The first method is the "particular case" approach. This is based on the characteristics of the delay system, namely: the input process, the queue discipline and the service mechanism. Practically all the waiting time distributions of the classical delay systems are determined by this method [1,19-21]. Therefore, this method will not be further discussed here. The second and third methods of determining the waiting time distribution of a delay system are the "general case" approach [22], and the traffic flow graph approach [5]. The "general case" approach is not any simpler than the first method but aids one to the understanding of the congestion processes producing delays. However, the third method, the traffic flow graph approach, may sometimes be simpler than the first two methods to be used in determining the waiting time distribution of a delay system. Both the second and third methods will be discussed in the following.

2.1 General Case Approach

In this approach, a single server (service channel) delay system is considered. This approach may be extended to the multi-service channels case. Let the single server delay system be subjected to a general stream of calls (demands for service) and operate in such a way that a call arriving, when the server is idle, obtains service immediately; otherwise, it is inserted in a queue where it remains until service is obtained. Moreover, calls are served one at a time, and while there are calls waiting, the server operates continuously.

At a particular moment t_0, a test call arrives and waits for a period of time D before gaining access to the service channel. Service times are initiated at times t_1, t_2, t_3, ... following the arrival of the test call. It is desired to find the queuing time distribution of this test call.

Now the determination of the probability that the time D is greater than t, $P\{D > t\}$, hinges on a very obvious fact: if $t > 0$, and the test call is still waiting at time $t_0 + t$, then there is at least one call in the system at time t_0, and whatever the number of service times initiated on the time interval $(t_0, t_0 + t]$, there is always at least one call waiting for service at those times t_1, t_2, t_3, ... such that

$$t_0 \leq t_1 \leq t_2 \leq t_3 \leq \ldots \leq t_0 + t.$$

Moreover, when the server is continuously busy between the times t_0 and $t_0 + t$, then exactly k service times are initiated in that interval of time provided the sum of k consecutive service times is at most t, whereas the sum of k + 1 consecutive service times exceeds t.

Observe that the validity of the previous statement is contingent neither on the queue discipline nor on any particular independence assumption about the delay system.

The relevant symbols are defined in the following.

Events

$\{D+\}$ the event that the test call is delayed, or that the server is busy at time t_0

$\{B_t\}$ the event that the server is continuously busy throughout the time interval $(t_0, t_0 + t]$

$\{D>t\}$ the event that the test call waits for a period of time exceeding t before obtaining service.

Random Variables

T_j the length of the service period beginning at t_{j-1}, $j = 2, 3, \ldots;$

T_1 refers to the remaining duration of the service in progress at time t_0, if any

S_j	the sum $\sum_{i=1}^{j} T_i$, for $j = 1, 2, \ldots$: but $S_0 = 0$
Z_1	the number of calls arriving between times t_0 and t_1, not including the test call
Z_j	the number of calls arriving during the service period beginning at t_{j-1}, $j = 2, 3, \ldots$
X_t	the number of service times initiated in the time interval $(t_0, t_0 + t]$
N_j	the number of calls waiting for service at time t_j, $j = 1, 2, \ldots$
N_0	the number of calls in the system at time t_0 excluding the test call, but including the call being served at time t_0, if any

The Queuing Time Distribution

It is convenient to summarize in compact forms the statements made above. To begin with, the delay of the test call cannot exceed t unless the server is busy at time t_0 and continues to remain busy throughout the interval $(t_0, t_0 + t]$, hence

$$\{D > t\} \subset \{D+\} \cap \{B_t\}. \tag{5.1}$$

If exactly k service times are initiated during the interval $(t_0, t_0 + t]$, then the occurrence of the event $\{D+\} \cap \{B_t\}$ implies that the variables N_0, N_1, ..., N_{k-1} and N_k cannot take the value 0. This leads to the following statement:

$$\{D+\} \cap \{B_t\} \subset \bigcup_{k=0}^{\infty} \{\{X_t = k\} \cap \{\bigcap_{i=0}^{k}\} \{N_i \geq 1\}\}. \qquad (5.2)$$

Finally, note that if both the events $\{X_t = k\}$ and $\{D+\} \cap \{B_t\}$ occur, then the variables N_0, N_1, N_2, ..., N_k must satisfy the relations

$$N_1 = N_0 + Z_1$$

$$N_j = N_{j-1} - 1 + Z_j, \; j = 2, 3, \dots, k. \qquad (5.3)$$

Now, the conditions expressed in the relations (5.2) and (5.3) completely specify the event $\{D+\} \cap \{B_t\}$; indeed this event is itself the aggregate of all mutually exclusive events, or congestion states

$$\{n_0, n_1, \dots, n_k, k\} = \{X_t = k\} \cap \{\bigcap_{i=0}^{k} \{N_i = n_i\}\}$$

such that

$$k = 0, 1, 2, \dots$$

$$n_0 \geq 1, \text{ for } k = 0, 1, 2, \dots$$

$$n_1 \geq n_0, \text{ and} \qquad (5.4)$$

$$n_i \geq \max(1, n_{i-1} - 1), \text{ for } k = 1, 2, 3, \dots$$

If R represents this aggregate, then the relations (5.1) and (5.4) imply that

$$P\{D > t\} = \sum_{\{n_0,n_1,n_2,\ldots n_k, k\}\varepsilon R} P\{D > t/n_0,\ n_1,\ \ldots,\ n_k,\ k\}p\{n_0,\ n_1,\ \ldots,\ n_k, k\} \qquad (5.5)$$

where the summation operator

$$\sum_{\{n_0,\ n_1,\ \ldots,\ n_k\ \ k\}\varepsilon R}$$

is to be read

$$\sum_{n_0=1}^{\infty}\ \sum_{k=0}^{\infty}\ \sum_{n_1=n_0}^{\infty}\ \cdots\ \sum_{n_i=\max(1,n_i-1)}^{\infty}\ \cdots\ \sum_{n_k=\max(1,\ n_{k-1}\ -\ 1)}^{\infty}.$$

and $P\{\cdot\}$ is the probability of the occurrence of event $\{\cdot\}$.

The equality (5.5) is, therefore, a general representation of the probability $P\{D > t\}$ based on elementary probability properties.

The Queue Discipline

Having characterized all the congestion states relevant to the analysis of the event $\{D > t\}$, the expression $P\{D > t/n_0,\ n_1,\ \ldots,\ n_k,\ k\}$, which denotes the conditional probability of this event given the particular congestion state $\{n_0,\ n_1,\ \ldots,\ n_k,\ k\}$, must naturally represent the law governing the selection of calls for service at times $t_1,\ t_2,\ \ldots,\ t_k$.

Needless to say, a very complex law of choice may result in a very complicated form for the expression $P\{D > t/n_0, n_1, \ldots, n_k, k\}$. Nevertheless, provided this law of choice can be expressed mathematically, the relation (5.5) determines the queueing time distribution completely. As examples, consider the three classical queue disciplines.

Strict Queuing

The test call arriving at t_0, and finding $n_0 - 1$ calls waiting, takes up the n_0th position in the queue and enters the service channel at time t_{n_0}. Therefore

$$\begin{aligned} P\{D > t/n_0, n_1, \ldots, n_j, j\} &= 0; \text{ if } n_0 \leq j,\ j > 0 \\ &= 1; \text{ if } n_0 > j,\ j \geqslant 0. \qquad (5.6) \end{aligned}$$

Random Order of Service

At each of the moments $t_1, t_2, \ldots, t_{x_t}$, the next call to enter the service channel is selected at random from among those waiting; hence

$$\begin{aligned} P\{D > t/n_0, n_1, \ldots, n_j, j\} &= 1, \text{ if } j = 0,\ n_0 \geqslant 1 \\ &= \prod_{i=1}^{j} \left(1 - \frac{1}{n_i}\right), \quad \text{if } j > 0. \qquad (5.7) \end{aligned}$$

Reverse Order of Service

The test call arriving at t_0 remains in the queue as long as the server is busy with calls that arrive after time t_0; therefore

$$
\begin{aligned}
P(D > t \mid n_0, n_1, \dots n_j, j) &= 1; \text{ if } j = 0,\ n_0 \geq 1 \\
&= 1; \text{ if } j > 0,\ n_i - n_0 > 0 \\
&\qquad i = 1, 2, \dots, j \\
&= 0; \text{ otherwise.}
\end{aligned} \tag{5.8}
$$

To see that this is so note that at time t_0 the test call becomes the first waiting in a queue of n_0 calls. If at any time after t_0, and before the test call enters the service channel, there are n calls waiting altogether, then $n - n_0$ calls must be served before this test call.

The Congestion States Probability

The general representation of the waiting time distribution being determined, the fact remains that, in many practical situations it would be of little help were it not for certain assumptions regarding the stream and disposal of calls. It is suggested, however, that rather than setting down assumptions beforehand for any given situation, it is desirable to study the general representation of the distribution, since this study may, of itself, suggest what assumptions ought to be made, and what approximations are suitable. In the final analysis, it is the very structure of this distribution relevant to a particular situation which dictates to what extent assumptions can be made in order to attain numerical results that are reasonable.

For the simplest types of queue disciplines, as in the examples discussed above, the following assumptions are usually made:

1) calls arriving in disjoint time intervals are independent;

2) the arrivals do not depend on the state of the system:

3) service times are independent of the number of calls waiting:

4) the queue, if any, may be unbounded, but the mean inter-arrival time exceeds the mean service time; and finally,

5) the call intensity is also time independent.

With these assumptions made, let $v_s\{k\}$ denote the probability that exactly k calls arrive during an arbitrary time interval of length s. In addition, define

$$P\{\tau_j \leq s\} = H(s),\ j = 2, 3, \ldots,$$

Then

$$P\{T_1 \leq s\} = H^*(s) = \frac{\int_0^s 1 - H(u)\,du}{\int_0^\infty 1 - H(u)\,du} = \sum_{n_0=1}^{\infty} P\{n_0\}P\{T_1 \leq s|n_0\}$$

$$= \sum_{n_0=1}^{\infty} P\{n_0\}\, H^{*}n_0(s)$$

Under these conditions and because of relation (5.3) and the fact that $s_k \leq t < s_{k+1}$,

$$p\{n_0, n_1, \ldots, n_k, k\} = p\{n_0\}p\{n_1, n_2, \ldots, n_k, k/n_0\} =$$

$$p\{n_0\} \int\int_{\{x_1+x_2+\ldots+x_k \leq t\}} \ldots \int v_{x_1}\{n_1 - n_0\} \prod_{i=2}^{k} v_{x_i}\{n_i - n_{i-1} + 1\}$$

$$(1 - H(t - x_1 - x_2 \ldots - x_k))dH^{*}_{n_0}(x_1)dH(x_2)$$

$$\ldots dH(x_k) \qquad (5.9)$$

where $p\{n_0\}$ is the stationary probability that there be n_0 calls in the system at t_0.

Thus, having the knowledge of the congestion states probability and the queue discipline of the delay system, the queuing time distribution can be determined from equation (5.5).

2.2 Traffic Flow graph Approach

To use the traffic flow graph approach for determining the waiting time distribution of a delay system one must subdivide the system into sub-systems forming a network of delay systems. In order that a delay system can be regarded as a network of delay systems, each delay system in the network may be defined as a time delay system which contributes a certain amount of time $t \geq 0$ to the total time required to process the request through the network. It has the same structure as Fig. 31. The time required to process a request from a source S_i to the output of the delay system D_j is called the delay time of the request from S_i with respect to D_j. Once a delay system is represented by such a network that the probabilities of a request (or call) processed through the sub-delay systems and the distributions of the corresponding delay times are known, the delay system is said to be represented by a traffic flow graph. Thus, the overall delay time distribution of the corresponding traffic flow graph is the waiting time distribution of the delay system.

However, to use the general formula (2.8) to determine the overall delay time distribution for a given traffic flow graph, it is necessary to know the directed links involved in each path between D_1 and D_n. For a relatively simple traffic flow graph, it is not a difficult task to enumerate the directed links in each path. As an example, consider the traffic flow graph shown in Figure 33.

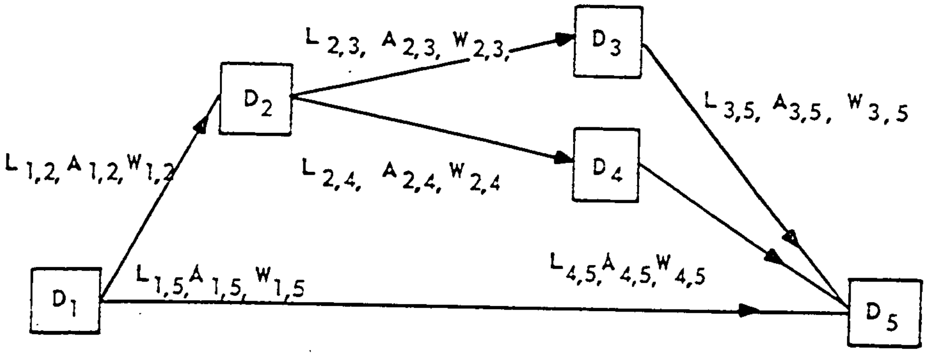


FIG. 33 - A SIMPLE TRAFFIC FLOW GRAPH

There are 5 nodes. D_1 is the source and D_5 the sink. There is one single step path between nodes D_1 and D_5, i.e., $Z^{(1)} = 1$, namely: $L_{1,5}$ with associated link delay $W_{1,5}$ and link selection variable $A_{1,5}$. Also there are two 3-step paths, i.e., $Z^{(3)} = 2$, namely: one path consisting of $L_{1,2}$, $L_{2,3}$ and $L_{3,5}$ in series, and another path consisting of $L_{1,2}$, $L_{2,4}$ and $L_{4,5}$ in series. Since $Z^{(k)} = 0$ for $k = 2$ and $k \geq 4$, the overall delay time distribution in accordance with (2.8) is therefore,

$$P\{W \leq t\} = P\{W_1^{(1)} \leq t\} + \sum_{r=1}^{2} P\{W_r^{(3)} \leq t\}$$

$$= P\{\{W_{1,5} \leq t\} \cap \{A_{1,5} = 1\}\}$$

$$+ P\{\{(W_{1,2} + W_{2,3} + W_{3,5}) \leq t\} \cap \{A_{1,2} = 1\} \cap \{A_{2,3} = 1\}$$

$$\cap \{A_{3,5} = 1\}\}$$

$$+ P\{\{(W_{1,2} + W_{2,4} + W_{4,5}) \leq t\} \cap \{A_{1,2} = 1\} \cap \{A_{2,4} = 1\}$$

$$\cap \{A_{4,5} = 1\}\}$$

It follows, by (2.6) and (2.7) that:

$$P\{W \leq t\} = \int_{0}^{t} P_{1,5} \cdot dP\{W_{1,5} \leq t\}$$

$$+ \int_{t_1=0}^{t} \int_{t_2=0}^{t-t_1} \int_{t_3=0}^{t-t_1-t_2} P_{1,2} \cdot dP\{W_{1,2} \leq t_3\}$$

$$\cdot P_{2,3} \cdot dP\{W_{2,3} \leq t_2\} \cdot P_{3,5} \cdot dP\{W_{3,5} \leq t_1\}$$

$$+ \int_{t_1=0}^{t} \int_{t_2=0}^{t-t_1} \int_{t_3=0}^{t-t_1-t_2} P_{1,2} \cdot dP\{W_{1,2} \leq t_3\}$$

$$\cdot P_{2,4} \cdot dP\{W_{2,4} \leq t_2\} \cdot P_{4,5} \cdot dP\{W_{4,5} \leq t_1\}$$

If a traffic flow graph is relatively complex, it may be decomposed into sub-flowgraphs in order to make the computation of the overall delay time distribution easier. The only condition required is that each sub-flow graph, when fitted into the final traffic flow graph, can be replaced by a directed link such that:

a) The associated link selection probability equals the sum of the link selection probabilities associated with the outgoing links from the source in the sub-flow graph.

b) The associated link delay distribution equals the overall delay time distribution of the sub-flow graph.

However, when the traffic flow graph is relatively complex, it may be a difficult task not only to enumerate the paths, but also to decompose it into sub-flow graphs. For this reason, a matrix technique for path enumerations is introduced.

Matrix Technique for Path Enumerations

To enumerate paths from D_1 to D_n in a given traffic flow graph by the matrix technique, the following conventions are adopted. The product of directed links indicates that the directed links are connected in series, and the sum, in parallel. It is assumed that there exists at most one directed link $L_{i,j}$ connecting any node D_i to any adjacent node D_j which may be D_i itself. If any directed link $L_{i,j}$ does not exist, $L_{i,j}$ is equated to zero.

In a given traffic flow graph, there may be a number of paths, consisting of k directed links, connecting D_i to D_j. Let this number be $Z_{i,j}^{(k)}$. In particular, $Z_{1,n}^{(k)} = Z^{(k)}$. If $L_{i,j}^{(k)}(r)$ is one of these $Z_{i,j}^{(k)}$ paths, say, the r^{th} path, then,

$$L_{i,j}^{(k)}(r) = L_{i,r_1} \cdot L_{r_1,r_2} \cdots L_{r_{k-1},j} \qquad \text{for } k > 1$$

$$= L_{i,j} \qquad \text{for } k = 1$$

where $1 \le r_u \le n$ and L_{i,r_1}, L_{r_1,r_2}, $\cdots$ and $L_{r_{k-1},j}$ are the directed links in the path. Note that the general formula (2.8) will produce the same result even if the positions of the directed links within the expression $L^{(k)}_{i,j}(r)$ are put into a different order. When $L^{(k)}_{i,j}(r) = 0$ it signifies that such a path does not exist. If the $L^{(k)}_{i,j}$ represents all the $Z^{(k)}_{i,j}$ paths, then

$$L^{(k)}_{i,j} = \sum_{r=1}^{Z^{(k)}_{i,j}} L^{(k)}_{i,j}(r) \tag{5.10}$$

For $k > 1$,

$$L^{(k)}_{i,j} = \sum_{r=1}^{Z^{(k)}_{i,j}} L^{(k-1)}_{i,r_{k-1}}(r) \cdot L_{r_{k-1},j}(r)$$

$$= \sum_{u=1}^{n} L_{u,j} \sum_{r=1}^{Z^{(k-1)}_{i,u}} L^{(k-1)}_{i,u}(r)$$

$$= \sum_{u=1}^{n} L^{(k-1)}_{i,u} \cdot L_{u,j}$$

For $k = 1$, obviously,

$$L^{(1)}_{i,j} = L_{i,j}$$

Suppose that an n x n matrix L is constructed such that the element of the i^{th} row and j^{th} column is $L_{i,j}$. If the product of L itself k times is L^k, then, the i^{th} row and j^{th} column element of L^k is exactly $L_{i,j}^{(k)}$ [23]. In particular, $L_{1,n}^{(k)}$ is the first row and n^{th} column element of L^k and represents all the $Z^{(k)}$ k-step paths in the given traffic flow graph.

As an example, the matrix representation for the traffic flow graph shown in Figure 33 is

$$L = \begin{array}{c} \\ D_1 \\ D_2 \\ D_3 \\ D_4 \\ D_5 \end{array} \begin{array}{c} \begin{array}{ccccc} D_1 & D_2 & D_3 & D_4 & D_5 \end{array} \\ \left[\begin{array}{ccccc} 0 & L_{1,2} & 0 & 0 & L_{1,5} \\ 0 & 0 & L_{2,3} & L_{2,4} & 0 \\ 0 & 0 & 0 & 0 & L_{3,5} \\ 0 & 0 & 0 & 0 & L_{4,5} \\ 0 & 0 & 0 & 0 & 0 \end{array}\right] \end{array}$$

Therefore, as expected,

$$L_{1,5}^{(1)} = L_{1,5}$$

$$L_{1,5}^{(2)} = 0$$

$$L_{1,5}^{(3)} = L_{1,2} \cdot L_{2,3} \cdot L_{3,5} + L_{1,2} \cdot L_{2,4} \cdot L_{4,5}$$

and

$$L_{1,5}^{(k)} = 0 \qquad \text{for } k \geq 4.$$

Once the paths for a given traffic flow graph have been enumerated in the form of (5.10) the procedure of obtaining $P\{W_r^{(k)} \geq t\}$ is very simple since $\{W_r^{(k)} \leq t\}$ is the joint event that the sum of the link delays associated with the directed links in the paths $L_{1,n}^{(k)}(r)$ does not exceed t, and that all link selection variables associated with the directed links in the path $L_{1,n}^{(k)}(r)$ are equal to one. Admittedly, when $Z^{(k)}$ increases with k, it is also laborious to obtain the overall delay time distribution $P\{W \leq t\}$. Fortunately, in many practical cases, if $Z^{(k)}$ increases with k, $P\{W_r^{(k)} \leq t\}$ usually converges fairly rapidly as k increases, so that k need not be too large in order to obtain a reasonably accurate result.

Traffic Flow Graph with Serial Dependence

Sometimes, the delay systems in the network of delay systems are interdependent with each other. If such a network is considered as a traffic flow graph, the resulting overall delay time distribution is only an approximation. The approximation is improved if some of these interdependencies are taken into consideration. In the following, a model called the traffic flow graph with serial dependence is studied.

Similar to a traffic flow graph, a traffic flow graph with serial dependence is defined as a graph consisting a set of n, where $1 < n \leq \infty$, nodes $\{D_1, D_2, \ldots, D_n\}$ interconnected by directed links in some fashion such that there is at most one link from D_i to D_j and such that

(1) Node D_1, called the source, has only outgoing links; and node D_n, called the sink, has only incoming links;

(2) Each directed link $L_{i,j}$ connecting node D_i to its adjacent node D_j, where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, n$ has the link selection binary variable $A_{i,j}$ taking values from $\{0,1\}$ with $\{A_{i,j} = 1\}$ representing the event that the link $L_{i,j}$ is selected ; and $\{A_{i,j} = 0\}$, otherwise;

(3) The link selection variables and the link delays in a given path from D_1 to D_n are interdependent, and furthermore, the events $\{A_{i,j} = 1\}$'s for a given i are mutually exclusive; otherwise, the link selection variables and the link delays are independent with each other.

Let $Z^{(k)}$ be the number of paths consisting of k directed links from D_1 to D_n:

$\{W_r^{(k)} \leq t\}$ be the event that the total delay time does not exceed t for the rth path with k links;

$\{W \leq t\}$ be the event that the overall delay time W by selecting a path between D_1 and D_n does not exceed t; and

$P\{.\}$ be the probability of the occurrence of event $\{.\}$.

Then, the overall delay time distribution for a given traffic flow graph with serial dependence is

$$P\{W \leq t\} = P\{\bigcup_{k=1}^{\infty} \bigcup_{r=1}^{Z^{(k)}} \{W_r^{(k)} \leq t\}\}$$

$$= \sum_{k=1}^{\infty} \sum_{r=1}^{Z^{(k)}} P\{W_r^{(k)} \leq t\} \tag{5.11}$$

which is the same as that for a given traffic flow graph because of the exclusive links . However, the probability that the total delay time does not exceed t for the rth one of the $Z^{(k)}$ path is

$$P\{W_r^{(k)} \leq t\} = P\{\{(W_{1,r_1} + W_{r_1,r_2} + \dots + W_{r_{k-1},n}) \leq t\} \cap$$

$$\{\Lambda_{1,r_1} = 1\} \cap \{\Lambda_{r_1,r_2} = 1\} \cap \dots$$

$$\cap \{\Lambda_{r_{k-1},n} = 1\}\}$$

$$= \int_{t_n=0}^{t} \int_{t_{r_{k-1}}=0}^{t-t_n} \cdots \int_{t_{r_1}=0}^{t-t_n-\sum_{m=1}^{k-2} t_{r_{k-m}}}$$

$$\alpha \, . \, dP\{\{W_{1,r_1} \leq t_{r_1}\} \Big| \{W_{r_1,r_2} = t_{r_2}\} \cap \dots \cap \{W_{r_{k-2},r_{k-1}} = t_{r_{k-1}}\}$$

$$\cap \, \{W_{r_{k-1},n} = t_n\}\} \dots dP\{\{W_{r_{k-2},r_{k-1}} \leq t_{r_k}\} \Big| \{W_{r_{k-1},n} = t_n\}\}$$

$$dP\{W_{r_{k-1},n} \leq t_n\} \tag{5.12}$$

where each subscript r_u takes a value from $\{2, 3, \dots, n-1\}$, the associated links L_{1,r_1}, L_{r_1,r_2}, ..., $L_{r_{k-1},n}$ form the rth one of the $Z^{(k)}$ path, and

$$\alpha = P\{\{A_{1,r_1} = 1\} \Big| \{A_{r_1,r_2} = 1\} \cap \dots \cap \{A_{r_{k-1},n} = 1\}$$

$$\cap\{(W_{1,r_1} + W_{r_1,r_2} + \dots + W_{r_{k-1},n}) \leq t\}\}$$

$$\cdots P\{\{A_{r_{k-1},n} = 1\} \Big| \{(W_{1,r_1} + W_{r_1,r_2} + \ldots + W_{r_{k-1},n}) \leq t\}\Big\}$$

In particular, when the link selection variables and the link delays in the path are mutually independent, (5.12) reduces to that of a traffic flow graph as it should.

3. Typical Example

To demonstrate the power of the general-case approach and the traffic flow graph approach for determining the waiting time distribution of a delay system, the Erlang model [24] of a delay system with a Poisson input, first-come, first-served discipline, and negative exponential service time is considered here. Note that only the queuing time distribution for this model is required to be determined since the waiting time is the sum of queuing time and service time which has been given. The specification of this model is as follows:

The input process is

$$v_s(k) = \frac{(as)^k}{k!} e^{-as} \qquad (5.13)$$

where $v_s(k)$ is the probability that exactly k calls arrive during an arbitrary time interval of length s, and a is the arrival rate. The queue discipline is specified by the conditional probability (5.6). And the service mechanism is specified by the service time distribution:

$$H(s) = 1 - e^{-bs} \qquad (5.14)$$

and c servers such that the interdeparture time distribution is

$$H'(s) = 1 - e^{-bcs} \qquad (5.14a)$$

where b^{-1} is the average service time.

General Case Approach

Since the general formula (5.5) was developed for a single server delay system, only the case where c=1 is considered here. Combining (5.5) and (5.6), the queuing time distribution for the single server delay system is 1-P{D>t} such that

$$P\{D>t\} = \sum_{n_0=1}^{\infty} \sum_{k=0}^{\infty} P\{n_0,k\} \qquad n_0>k\geq 0$$

$$= \sum_{n_0=k+1}^{\infty} \sum_{k=0}^{\infty} P\{n_0,k\} \tag{5.15}$$

since the calls arriving later than the test call will not affect the delay of the test call to be served. By (5.9),

$$P\{n_0,k\} = P\{n_0\} \int\int \dots \int_{\{x_1+x_2+\dots+x_k \leq t\}} (1-H(t-x_1-x_2-\dots-x_k))$$

$$dH^*_{n_0}(x_1)\ dH(x_2)\ \dots\ dH(x_k) \tag{5.16}$$

But [25]

$$P\{n_0\} = \int_0^{\infty} U_{n_0+1}(x)\ dH(x) \tag{5.17}$$

and

$$H^*_{n_0}(\tau) = \frac{a}{P\{n_0\}} \int_{x=0}^{\tau} \int_{t=x}^{\infty} U_{n_0}(t-x) \, . \, dH(t) \, . \, dx \qquad (5.18)$$

where

$$U_k(s) = \sum_{n_0=0}^{k} P\{n_0\} \, v_s(k-n_0). \qquad (5.19)$$

Let σ = t-x then (5.18) becomes

$$H^*_{n_0}(\tau) = \frac{a}{P\{n_0\}} \int_{x=0}^{\tau} \int_{\sigma=0}^{\infty} U_{n_0}(\sigma) \, . \, dH(\sigma+x) \, . \, dx \qquad (5.20)$$

Again, let $w=\sigma+x$, then

$$H^*_{n_0}(\tau) = \frac{a}{P\{n_0\}} \int_{\sigma=0}^{\infty} U_{n_0}(\sigma) \int_{w=\sigma}^{\sigma+\tau} dH(w) \, dx$$

$$= \frac{a}{P\{n_0\}} \int_{\sigma=0}^{\infty} U_{n_0}(\sigma) \, (H(\sigma+\tau)-H(\sigma)) \, d\sigma \qquad (5.21)$$

But, by (5.14)

$$H(\sigma+\tau) - H(\sigma) = (1-e^{-b\tau})e^{-b\sigma}$$

$$= H(\tau)e^{-b\sigma} \qquad (5.22)$$

Therefore, (5.21) becomes

$$H^*_{n_0}(\tau) = \frac{a}{P\{n_0\}} \int_{\sigma=0}^{\infty} U_{n_0}(\sigma)\, H(\tau) e^{-b\sigma}\, d\sigma$$

$$= \frac{a\ H(\tau)}{b\ P\{n_0\}} \int_{\sigma=0}^{\infty} U_{n_0}(\sigma)\, dH(\sigma) \qquad (5.23)$$

But [19]

$$P\{n_0\} = (1-\frac{a}{b})\, (\frac{a}{b})^{n_0} \qquad (5.24)$$

Thus, combining (5.17), (5.23) and (5.24) results in

$$H^*_{n_0}(\tau) = H(\tau) \qquad (5.25)$$

That is, the conditional residual service time distribution equals the service time distribution for the exponential service time distribution case. Substituting (5.14) and (5.25) into (5.16), one obtains

$$P\{n_0,k\} = P\{n_0\} \int\int \dots \int_{\{x_1+x_2+\dots+\ x_k\leq t\}} b^k\ e^{-bt}\ dx_1 dx_2 \dots dx_k \qquad (5.26)$$

Let $\frac{x_i}{t} = w_i$ for $i = 1, 2, \dots, k$. Then (5.26) becomes

$$P\{n_0,k\} = P\{n_0\} (bt)^k e^{-bt} \int\int\cdots\int_{\{w_1+w_2+\dots+w_k \leq 1\}} dw_1\, dw_2 \dots dw_k \tag{5.27}$$

According to Dirichlet integral [1], (5.27) can be written as

$$\begin{aligned} P\{n_0,k\} &= P\{n_0\} (bt)^k e^{-bt} \int\int\cdots\int_{\{w_1+w_2+\dots+w_k \leq 1\}} \prod_{i=1}^{k} w_i^{1-1}\, dw_i \\ &= P\{n_0\} (bt)^k e^{-bt} \frac{\prod_{i=1}^{k} \Gamma(1)}{\Gamma(k)} \int_{w=0}^{1} w^{k-1}\, dw \end{aligned}$$

$$= \frac{P\{n_0\}\ (bt)^k}{k!} e^{-bt} \tag{5.28}$$

Therefore, from (5.15), (5.24) and (5.28), the required queuing time distrbution is

$$\begin{aligned} P\{D \leq t\} &= 1 - P\{D > t\} \\ &= 1 - \sum_{n_0=k+1}^{\infty} \sum_{k=0}^{\infty} \frac{(1-\frac{a}{b})\ (\frac{a}{b})^{n_0}\ (bt)^k}{k!} e^{-bt} \end{aligned}$$

$$= 1 - (1-\frac{a}{b})\, e^{-bt} \sum_{k=0}^{\infty} \frac{(bt)^k}{k!} (\frac{a}{b})^{k+1} \sum_{n_0=0}^{\infty} (\frac{a}{b})^{n_0}$$

$$= 1 - e^{-bt}\, \frac{a}{b} \sum_{k=0}^{\infty} \frac{(at)^k}{k!}$$

$$= 1 - \frac{a}{b}\, e^{-(b-a)t} \qquad (5.29)$$

which is the same as obtained by the presently known method [19].

Traffic Flow Graph Approach

The processing of a service request in the Erlang Model of a delay system with c servers may be displayed graphically as shown in Figure 34. In this traffic flow graph, node D_1 represents the source of the request and node D_n the completion of its service. When the request appears in D_1 at time t_0, the system must be in one of the following states:

a) There are c-1 or less requests in the system at t_0. That is, there is at least one of the c servers idle at t_0.

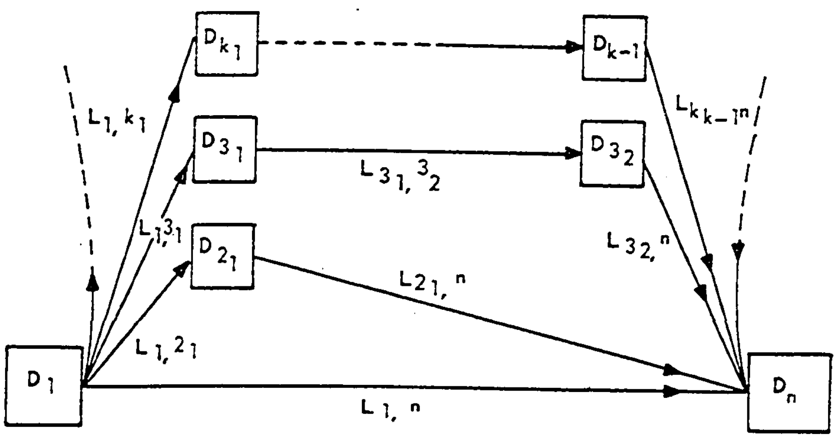


FIG. 34 - THE TRAFFIC FLOW GRAPH FOR THE ERLANG MODEL OF A DELAY SYSTEM.

This is represented by event $\{A_{1,n}=1\}$. In this case, the service of the request begins also at t_0. The request will appear at D_n when its service is completed. Thus, it is delayed to reach D_n from D_1 by its service time. This delay time, and hence its service time, is represented by $W_{1,n}$. In other words, the request is processed through a single step path $L_{1,n}$ with a probability $P_{1,n}$ and a delay time $W_{1,n}$.

b) There are exactly $(c + k - 2)$ requests in the system at t_0, where $k \geq 2$. That is, all c servers are busy and there are (k-2) other requests waiting at t_0. This is represented by $\{A_{1,k_1}=1\}$. In this case, because of the first-come, first-served queue discipline, the service of the request begins only after all the other waiting requests have been served and one of the servers becomes idle. If $(k-2)=0$, its service begins at the end of the residual interdeparture time (ie., at the time when one of the servers becomes idle). It appears at D_n when its own service time is completed. Thus, if k=2, the delay time of the request equals the residual interdeparture time plus its own service time. This residual time is represented by $W_{1,2_1}$ and its own service time by $W_{2_1,n}$ with the associated link selection variable $A_{2_1,n}$ always being 1. In other words, if k=2, the request is first processed through a delay system D_{2_1} with a probability $P_{1,2_1}$ and a delay time $W_{1,2_1}$, then through another delay system D_n with a probability $P_{2_1,n}=1$ and a delay time $W_{2_1,n}$. Similarly, if (k-2)>0, the first delay for the request to reach D_n from D_1 is the residual interdeparture time, the last delay its own service time, and the other delays the interdeparture times for the other requests. In other words, the request is processed through a path consisting of k delay systems, the first being D_{k_1} with a probability P_{1,k_1} and a delay time W_{1,k_1} which equals the residual interdeparture time, the last D_n with a probability $P_{k_{k-1},n}=1$ and a delay time $W_{k_{k-1},n}$ which is its own service time, and each of the others D_{k_i}, with a probability $P_{k_{i-1},k}=1$ and a delay time W_{k_{i-1},k_i} which is one interdeparture time where i = 2, 3, ..., k-1.

Because the queueing rather than the waiting time distribution is required, the service time of the request must be equated to zero in order that (2.8) can be applied. That is

$$P\{W_{1,n} \leq t\} = P\{W_{k_{k-1},n} \leq t\} = 1 \quad \text{for } t \geq 0 \qquad (5.30)$$

Since the interdeparture time distribution is negative exponential, the residual interdeparture time distribution is also negative exponential.

If P_x is the stationary probability that there are x requests in the system at the arrival time t_0 of the request, then [19],

$$P_x = P_0 \frac{(a/b)^x}{x!} \qquad x \leq c$$

$$P_{c+m} = P_c \frac{(a/b)^m}{c^m} \qquad m > 0$$

and

$$P_0^{-1} = 1 + \frac{a}{b} + \ldots + \frac{(a/b)^{c-1}}{(c-1)!} + \frac{c(a/b)^c}{c!(c-a/b)}$$

Therefore, it follows,

$$P_{1,n} = \sum_{x=0}^{c-1} P_x = P_0 \sum_{x=0}^{c-1} \frac{(a/b)^x}{x!} \tag{5.31}$$

and

$$P_{1,k_1} = P_{c+k-2} = P_c \frac{(a/b)^{k-2}}{c^{k-2}} \quad \text{for } k \geq 2 \tag{5.32}$$

Hence, from (2.6), (5.30) and (5.31)

$$P\{W_1^{(1)} \leq t\} = P_0 \sum_{x=0}^{c-1} \frac{(a/b)^x}{x!} \qquad (5.33)$$

and from (2.6), (5.14a) (5.30) and (5.32), for $k \geq 2$,

$$P\{W_1^{(k)} \leq t\} = P_c \frac{(a/b)^{k-2}}{c^{k-2}} \left(1 - e^{-bct} \sum_{x=0}^{k-2} \frac{(bct)^x}{x!}\right) \qquad (5.34)$$

since $p_{k_{k-1},n}=1$, and $p_{k_{i-1},k_i}=1$, where $i=2,3,\ldots,k-1$. Finally, since $Z^{(k)}=1$ for $k \geq 1$, substituting (5.33) and (5.34) into (2.8) results in the required queuing time distribution as follows:

$$\begin{aligned}
P\{W \leq t\} &= P_0 \sum_{x=0}^{c-1} \frac{(a/b)^x}{x!} \\
&\quad + \sum_{k=2}^{\infty} P_c \frac{(a/b)^{k-2}}{c^{k-2}} \left(1 - e^{-bct} \sum_{x=0}^{k-2} \frac{(bct)^x}{x!}\right) \\
&= P_0 \sum_{x=0}^{c-1} \frac{(a/b)^x}{x!} + P_c \sum_{m=0}^{\infty} \frac{(a/b)^m}{c^m} \\
&\quad - P_c e^{-bct} \sum_{m=0}^{\infty} \left(\frac{a}{bc}\right)^m \sum_{x=0}^{m} \frac{(bct)^x}{x!} \\
&= 1 - P_c e^{-bct} \sum_{x=0}^{\infty} \frac{(bct)^x}{x!} \sum_{m=x}^{\infty} \left(\frac{a}{bc}\right)^m
\end{aligned}$$

$$= 1-P_c e^{-bct} \sum_{x=0}^{\infty} \frac{(at)^x}{x!} \cdot \frac{1}{1-a/bc}$$

$$= 1-P_0 \frac{(a/b)^c}{c!} \cdot \frac{c}{c-a/b} \cdot e^{-(bc-a)t} \qquad (5.35)$$

which is again the same as obtained by the presently known methods [10]. Note that the matrix technique is not used in this case for two reasons:

1. the traffic flow graph has infinite number of nodes, and
2. the paths are easily seen.

Chapter VI

A Simple Telephone Exchange

1. Introduction

In this chapter, the methods outlined in this research for estimating the overall grade of service of a telephone exchange are illustrated by considering a simple hypothetical telephone switching system. This system has the basic structure as shown in Fig. 1. Its switching network has two switching stages as shown in Fig. 29. There is only one common control unit in the system. The types of trunks connecting to the switching networks are:

(a) Dial Tone Service Circuit (DTSC) - which provides dial tone to a subscriber request and stores the digits subsequently dialed (or keyed) by the subscriber;

(b) Outgoing trunk (OT) - which connects a subscriber-line through the network to the distant exchange;

(c) Intra-office trunk (IAO) - which connects a pair of subscriber lines through the network;

(d) Incoming trunk (IT) - which connects a subscriber line from the distant exchange to a subscriber line through the network;

In short, the complete switching system considered has the essential physical structure as shown in Fig. 35. It should be noted that there is a group of senders (SDR) and a group of incoming registers (IR) which are not connected to the switching network but to a non-blocking switch.

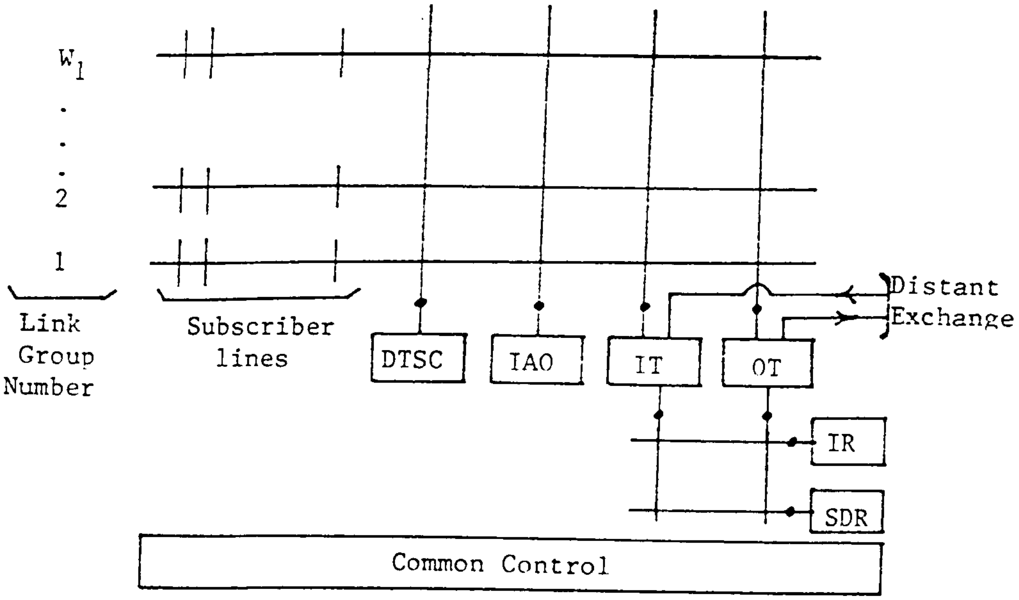


FIG. 35 - A SIMPLE SWITCHING SYSTEM

The SDR is used to outpulse the called number through the outgoing trunk to the distant exchange: whereas the IR is used to receive the digits through the incoming trunk from the distant exchange.

Let the objective grade of service for this exchange in a given period of time be as follows:

(a) The probability, $P\{W > t\}$, that a dial tone request waits longer than time t is no more than δ: i.e.

$$P\{W > t\} \leq \delta \qquad (6.1)$$

and (b) The probability, B, that a call encounters overflow conditions is no more than β; ie.

$$B \leq \beta \tag{6.2}$$

The problem is to provide enough sub-units in the system so that the objective grade of service is met.

2. System Mode of Operation

In order to represent the processing of various types of calls in the switching system by the probability linear graph or the traffic flow graph, it is necessary to have the knowledge of the system mode of operation. It should be noted that a different probability linear graph or traffic flow graph will result for the switching system of the same physical structure, if the system mode of operation is different. As an example, let the dial tone call be processed so that an overflow tone is returned to the subscriber if it encounters congestion. Then, the dial tone call is considered as a lost call since no dial tone will return to the subscriber regardless of the length of time that the subscriber will wait. In this case, the dial tone may return to the subscriber only if the subscriber makes another dial tone call. Clearly, this mode of operation for dial tone calls is different from the mode of operation shown in

Fig. 2 described in Chapter II. To estimate the overall grade of service, as pointed out in Chapter II, the probability linear graph should be used for the former, and the traffic flow graph for the latter.

In the switching system considered here, the modes of operation for various types calls are assumed as follows:

(a) Dial tone call - same as shown in Fig. 2.

(b) Intra-office call - same as shown in Fig. 3.

(c) Outgoing call - After the subscriber lifts his receiver (off-hook), obtains dial tone, and dials (or keys) the number of the called subscriber in the distant exchange, this number is stored in the DTSC until the common control analyses it. When the common control recognizes that the number in the distant exchange, it searches for a free OT in the trunk group connecting this exchange to the distant exchange. If all OT's in the trunk group are busy, an overflow tone is sent to the calling subscriber through the DTSC. If there is a free OT, the common control uses the same link in the network connecting the calling subscriber line to the DTSC to connect the OT to the calling subscriber line. Then, the DTSC is released. Because of this, this call can not be blocked by the network. After the calling subscriber line is connected to the OT, the common control searches for a free SDR. If all SDR are busy, an overflow tone is returned to the calling subscriber through the OT. If there is a free SDR, the common control will pass the called number in the distant exchange to the SDR and connects the SDR to the OT through the nonblocking switch. Then, the common control releases itself and the SDR will send the called number to the distant exchange when the distant exchange is ready to receive digits. After the called number is sent, the SDR releases itself. Thus, the calling subscriber line is connected to the distant exchange.

(d) Incoming call - An IT is seized by an OT in the distant exchange. The IT signals the common control to connect it to an idle IR. If all IR's are busy, an overflow tone is returned to the distant exchange. If there is an idle IR, the common control connects the IT to the IR through the nonblocking switch and then releases itself. The IR will send a "ready" singal to the distant exchange through the IT so that the distant exchange will subsequently send the called number of a subscriber in this exchange to the IR. Upon receiving all the digits, the IR signals the common control for service. The common control determines the location of the called line from the digits received by the IR. Then, releasing the IR, the common control searches for a free path in the network between the IT and the called line. If the network is blocked, an overflow tone is returned to the distant exchange. If there is a free path, the called line and the IT are connected. A ringing tone is sent to the subscribers and the common control releases itself. Thus, the called line is connected to the calling line in the distant exchange.

These modes of operations for outgoing and incoming calls are displayed graphically as shown in Figs. 36 and 37 respectively.

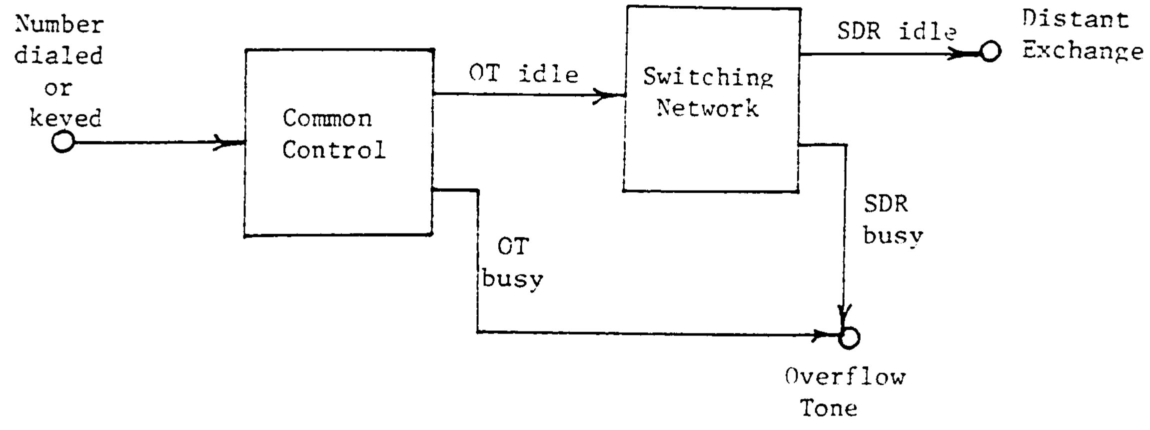


FIG. 36 - MODE OF OPERATIONS FOR OUTGOING CALL.

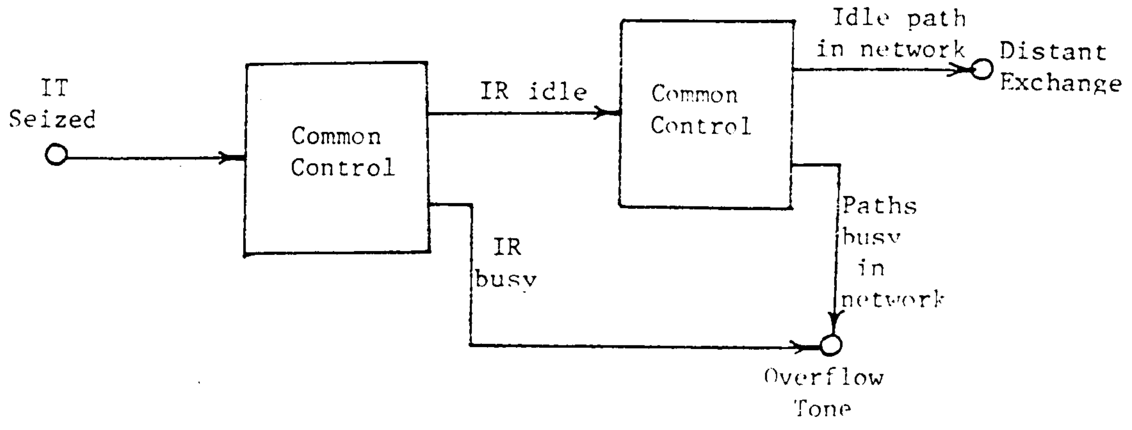


FIG. 37 - MODE OF OPERATIONS FOR INCOMING CALL.

3. Estimation of the overall grade of service

According to the system mode of operation, the dial tone call is on delay basis, whereas the other types of calls are on lost call basis. Therefore, the traffic flow graph may be used to estimate the dial tone delay, and the probability linear graph to estimate the probability of lost calls.

3.1 Traffic flow graph for dial tone calls

From the system mode of operation, the process of a dial tone call (request) may be displayed by a traffic flow graph as shown in Fig. 38.

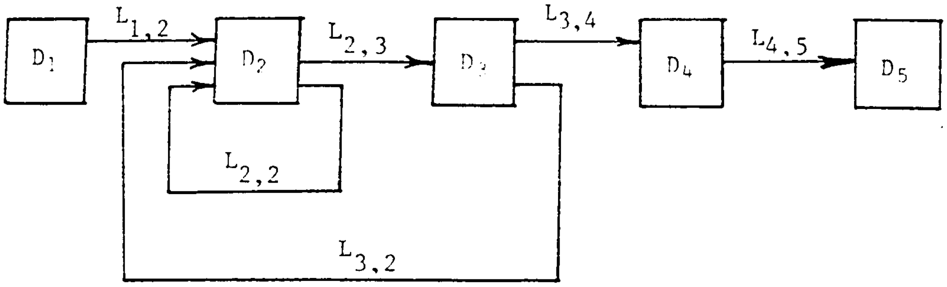


FIG. 38 - THE TRAFFIC FLOW GRAPH FOR A DIAL TONE CALL.

The physical representations of the nodes are as follows: node D_1 represents the telephone subscriber off-hook signal; node D_2, the output of the common control; node D_3, the output of the DTSC; node D_4, the output of the switching network; node D_5, the return of dial tone to the subscriber. The physical representations of the directed links are:

(1) $L_{1,2}$: the associated link delay $W_{1,2}$ represents the waiting time of the dial tone request for the common control; and the associated link selection variable $A_{1,2}$ has a probability $P_{1,2} = 1$, i.e. the request will wait as long as necessary until it is served by the common control.

(2) $L_{2,2}$: the associated link delay $W_{2,2}$ represents the waiting time of the request for the common control given that the request has been served at least once by the common control and encounters all DTSC's busy. The associated link selection variable $A_{2,2}$ has a probability $P_{2,2}$ which is the congestion function of the DTSC group.

(3) $L_{2,3}$: the associated link delay $W_{2,3}$ represents the service (operating) time of the DTSC for the request. The associated link selection variable $A_{2,3}$ has a probability of $1 - P_{2,2}$ which is the probability that there is at least one DTSC idle at the service requesting time.

(4) $L_{3,2}$: After the DTSC has served the request, the request must wait for the service of the common control again if

all the available network paths are busy. This waiting time is represented by the associated link delay $W_{3,2}$ and the probability that all paths in the network are busy is $P_{3,2}$.

(5) $L_{3,4}$: If the network is not blocked which has a probability $(1 - P_{3,2})$, the request must wait for the network to complete its operating (service) time for the request. This is represented by the link delay $W_{3,4}$.

(6) $L_{4,5}$: After the network has served the request, the request must wait for additional operating time of the DTSC in order to receive dial tone. This time is represented by $W_{4,5}$ and the associated link selection probability is $P_{4,5} = 1$.

The paths from D_1 to D_5 in the traffic flow graph may be enumerated by means of the matrix technique. Thus

$$L = \begin{array}{c} \\ D_1 \\ D_2 \\ D_3 \\ D_4 \\ D_5 \end{array} \begin{array}{c} \begin{array}{ccccc} D_1 & D_2 & D_3 & D_4 & D_5 \end{array} \\ \left(\begin{array}{ccccc} 0 & L_{1,2} & 0 & 0 & 0 \\ 0 & L_{2,2} & L_{2,3} & 0 & 0 \\ 0 & L_{3,2} & 0 & L_{3,4} & 0 \\ 0 & 0 & 0 & 0 & L_{4,5} \\ 0 & 0 & 0 & 0 & 0 \end{array} \right) \end{array} \qquad (6.3)$$

Clearly,

$$
\begin{aligned}
L_{1,5}^{(k)} &= 0 && \text{for } k \leq 3 \\
&= L_{1,2}\, L_{2,3}\, L_{3,4}\, L_{4,5} && \text{for } k = 4 \\
&= L_{1,2}\, L_{2,2}\, L_{2,3}\, L_{3,4}\, L_{4,5} && \text{for } k = 5 \\
&= L_{1,2}\, L_{2,2}^2\, L_{2,3}\, L_{3,4}\, L_{4,5} && \\
&\quad + L_{1,2}\, L_{2,3}^2\, L_{3,2}\, L_{3,4}\, L_{4,5} && \text{for } k = 6 \\
& && \text{etc.}
\end{aligned}
$$

Thus, the overall delay distribution is

$$
\begin{aligned}
P\{W \leq t\} = {} & (1-P_{2,2})(1-P_{3,2})P\{(W_{1,2} + W_{2,3} + W_{4,5}) \leq t\} \\
& + P_{2,2}(1-P_{2,2})(1-P_{3,2})P\{(W_{1,2} + W_{2,2} + W_{2,3} + W_{3,4} \\
& \qquad + W_{4,5}) \leq t\} \\
& + P_{2,2}^2(1-P_{2,2})(1-P_{3,2})P\{(W_{1,2} + W_{2,2} + W_{2,3} + W_{2,2} \\
& \qquad + W_{3,4} + W_{4,5}) \leq t\} \\
& + (1-P_{2,2})^2 P_{3,2}(1-P_{3,2})P\{(W_{1,2} + W_{2,3} + W_{2,3} \\
& \qquad + W_{3,2} + W_{3,4} + W_{4,5}) \leq t\} \\
& + \ldots \qquad\qquad (6.4)
\end{aligned}
$$

3.2 Probability linear graph for overflow calls

Whenever overflow tone is returned to the subscriber, the call is considered as a lost call. For intra-office call, from Fig. 3, the overflow tone is returned if

(a) all IAO trunks are busy, or

(b) the switching network is blocked, or

(c) both (a) and (b).

Let P_{IAO} be the probability that all IAO trunks are busy, and P_{SN1} be the switching network blocking probability for an intra-office call. Then, the probability linear graph for an intra-office call to become lost is as shown in Fig. 39.

FIG. 39 - PROBABILITY LINEAR GRAPH FOR AN INTRA-OFFICE CALL TO BECOME LOST

Let B_{IAO} be the probability that an intra-office call is lost. Then, by (2.2) and (2.4)

$$B_{IAO} = P_{IAO} + (1 - P_{IAO})\, P_{SN1} \qquad (6.5)$$

Similarly, the probability B_{OG} that an outgoing call is lost in this exchange is, from Fig. 36,

$$B_{OG} = P_{OT} + (1 - P_{OT}) P_{SDR} \tag{6.6}$$

where P_{OT} is the probability that all outgoing trunks in the trunk group are busy and P_{SDR} is the probability that all senders are busy.

From Fig. 37, the probability B_{INC} that an incoming call is lost in this exchange is

$$B_{INC} = P_{IR} + (1 - P_{IR}) P_{SN2} \tag{6.7}$$

where P_{IR} is the probability that all incoming registers are busy, and P_{SN2} is the network blocking probability for an incoming call.

An inter-exchange call may receive overflow tone in this exchange or from the distant exchange. If the distant exchange has the same switching system as this exchange, then, the probability linear graph representation of such a call is as shown in Fig. 40. Therefore, if B_{IEX} is the

FIG. 40 - PROBABILITY LINEAR GRAPH FOR AN INTER-EXCHANGE CALL TO BECOME LOST

probability that an inter-exchange call receives overflow tone, then,

$$
\begin{aligned}
B_{IEX} &= B_{OG} + (1 - B_{OG})\, B_{INC} \\
&= P_{OT} + (1 - P_{OT})\, P_{SDR} + (1 - P_{OT})(1 - P_{SDR})\, P_{IR} \\
&\quad + (1 - P_{OT})(1 - P_{SDR})(1 - P_{IR}) P_{SN2} \qquad (6.8)
\end{aligned}
$$

3.3 Grade of service estimation

In order to satisfy the objective grade of service (6.1) and (6.2), it is essential that

$$1 - P\{W \leq t\} \leq S \qquad (6.9)$$

$$B_{IAO} \leq \beta \qquad (6.10)$$

and

$$B_{IEX} \leq \beta \qquad (6.11)$$

To verify whether (6.4) satisfies (6.9), it is necessary to have the knowledge of the congestion functions of $P_{2,2}$, $P_{3,2}$, $P\{W_{1,2} \leq t\}$, $P\{W_{2,2} \leq t\}$, $P\{W_{2,3} \leq t\}$, $P\{W_{3,2} \leq t\}$, $P\{W_{3,4} \leq t\}$ and $P\{W_{4,5} \leq t\}$. According to the mode of operations, $P_{2,2}$ may be approximated by the Erlang B formula, $B(C, A, \infty)$, where A is the traffic offered to the DTSC group and C is the number of

DTSC in the group. Since A can be estimated and C is known, $P_{2,2}$ can be computed. $P_{3,2}$ is the blocking probability for a single linking-stage network loaded with multitypes of traffic. Hence, it can be approximated in accordance with the method outlined in section 3 of Chapter IV. In general, the distribution functions for $W_{1,2}$, $W_{2,2}$, $W_{2,3}$, $W_{3,2}$, $W_{3,4}$ and $W_{4,5}$ are rather complex. For practical results, one may assume that the operating times of DTSC and of the network are constant and that $P\{W_{1,2} \leq t\} = P\{W_{2,2} \leq t\} = P\{W_{3,2} \leq t\}$ with Poisson arrivals. Since there is only one common control in the switching system, $P\{W_{1,2} \leq t\}$ can be expressed as a function of λ, μ and t by any one of the methods outlined in section 2 of Chapter V where λ is the arrival rate and μ the service rate of the common control. As λ and μ can be estimated, $P\{W \leq t\}$ of (6.4) can therefore be evaluated for the specified value of t. Thus, one can verify whether the objective grade of service for a dial tone request is met.

To verify whether (6.5) and (6.8) satisfy (6.10) and (6.11) respectively, one must also know the congestion function of P_{IAO}, P_{SN1}, P_{OT}, P_{SDR}, P_{IR}, and P_{SN2}. The network blocking probabilities P_{SN1} and P_{SN2} are, in general, different. These probabilities may be approximated in accordance with section 3 of Chapter IV in this case. The other congestion functions depend also on the trunking arrangements. Since the trunking arrangements are known in this case, these congestion functions can be approximated in accordance with the methods outlined in Chapter III. Thus, one can verify whether the objective grade of service for calls encountered overflow conditions is met.

Chapter VII

Conclusion

From the foregoing discussions, the following conclusions are drawn:

1. For administration purposes, the overall grade of service provided in a given telephone exchange must be estimated. In order to do this analytically, two models are required: one for lost calls and another for delayed calls. For practical results, the probability linear graph can be employed as the model for the former, and the traffic flow graph for the latter. Both of these models are expressed in terms of the congestion functions of the sub-units of the switching system in the exchange. Thus, it is a simple procedure to determine the quantity of equipment for the switching system in the exchange to satisfy the objective grade of service.

2. The degree of accuracy, in using these two models to evaluate the overall grade of service of an exchange, depends greatly on the accuracy of the congestion functions describing the congestion conditions of the sub-units of the switching system in the exchange. Usually, the trunking arrangements and the modes of operation of the sub-units are very complex, and their congestion functions can only be approximated. Furthermore,

these sub-units are usually not independent of each other. Therefore, the use of these two models to evaluate the overall grade of service of an exchange yields only approximate results.

3. In order to improve the accuracy of estimating the overall grade of service of an exchange, much more research work must be done to improve the accuracy of the congestion functions describing the congestion conditions of the sub-units. The following lists some of the areas requiring further attention :

A. Trunk group congestion conditions

Up to the present, only idealized trunking arrangements and modes of operation have been investigated. In practice, for economical reasons, besides the complexity of interconnections, a trunk group may have to serve more than one kind of traffic. As an example, a full availability trunk group may have to serve calls on lost-calls basis as well as on delayed calls basis [26]. Clearly, a large number of problems remain to be solved in order that the accuracy of the congestion functions describing the congestion conditions of the trunk groups can be improved. As for example:

(a) More than one kind of traffic offered to a grading trunk group;

(b) Overflow traffic and alternate routing [27] with finite number of traffic sources;

(c) Complex interconnected trunk group with constant service time.

B. Switching networks

As pointed out in Chapter IV, the simple Lee - LeGall model is not applicable for single linking stage network loaded with multi-type of traffic. Although the new model presented in this thesis appears to give reasonable results, traffic measurements should be made to justify its practical application.

The accuracy of the network blocking probability depends greatly on the accuracy of the prediction of the link occupancies, which, in turn, depends on the amount of traffic generated from the telephone subscribers. The subscribers are usually arranged in groups in the switching systems. Thus, some groups of subscribers may generate more traffic than the others. This results unbalance traffic in the exchange. Hence, the link occupancies are affected. Although some effects of unbalance traffic have been studied from telephone administration point of view [28,29], the effect of unbalance traffic on the prediction of link occupancies still remains as a problem.

C. Waiting system

Even among the simple special cases of the investigated waiting systems, some of the congestion functions are practically intractable for numerical computations. For such cases and more complex waiting systems, the numerical results may be approximated by means of the traffic flow graph discussed in this research. However, on planning to design a switching system, it is useful to have numerical tables or graphs of congestion functions for various waiting systems available for consultations. Therefore, it is worth to make numerical computations for some frequently used waiting systems. Some of the waiting systems whose congestion functions are not available in numerical form and frequently encountered in practice are

(a) Single or multi-servers delay system with Poissonian input, discrete constant service times (i.e. constant service time for each type of traffic), all on first-come first-served basis.

(b) Same as (a), but all on random order of service basis.

(c) Same as (a), but the order of service is on priority basis. Within each priority, the order of service is first-come first-served.

(d) Same as (c), but the order of service within each priority is random.

4. With the two models discussed in this research for estimating the overall grade of service given to the telephone subscribers in an exchange, it is possible to dimension the switching system in the exchange in more than one way and yet satisfy the objective grade of service. Thus, taking the costs of the sub-units of the switching system in the exchange into account, there should be an optimum way of dimensioning such that the objective grade of service is satisfied and the cost of the system is minimum. This optimization has not been investigated in this research, but any progress in this direction is highly desirable.

References

1. R. Syski,

   "Introduction to congestion theory in telephone systems". London: Oliver and Boyd, 1960.

2. B.V.M. Ahlstedt,

   "Duration of the congestion state in local telephone traffic", the 4th International Teletraffic Congress, London, July 15-21, 1964.

3. Y. Rapp,

   "Extension of telephone plant with regard to the value of subscribers' time", Ericsson Technics, No. 1, 1961.

4. C.Y. Lee,

   "Analysis of Switching Network", Bell Sys. Tech. J., Vol. 34, pp. 1287 - 1315, 1955.

5. L. Lee,

   "Waiting-time distributions for networks of delay systems", the 5th International Teletraffic Congress, New York, June 14-20, 1967.

6. G.S. Berkeley,

"Traffic and trunking principles in automatic telephone", 2nd revised edition, 1949, Ernest Benn, Ltd., London.

7. E. Brockmeyer, H.L. Halstrom, and A. Jensen,

"The life and works of A.K. Erlang", Trans. Danish Acac. Tech. Sci. No. 2, 1948. Copenhagen.

8. C. Clos, and R.I. Wilkinson,

"Dialing habits of telephone customers", Bell Sys. Tech. J., Vol. 31, pp. 32 - 67, 1952.

9. G.F. O'Dell,

"An outline of the trunking aspect of automatic telephony", Journal of the IEE 65(1927), p. 185 - 214.

10. A. Lotz,

"History and development of grading theory", the 5th International Teletraffic Congress, New York, June 14 - 20, 1967.

11. C. Palm,

"Table of the Erlang Loss Formula", 2nd edition, Telefonaktiebolaget L.M. Ericsson, Stockholm, 1954.

12. C. Clos,
"A study of non-blocking switching networks", Bell Sys. Tech. J. Vol. 32, pp. 406 - 424, March, 1953.

13. R.F. Grantges, and N.R. Sinowitz,
"NEASIM", Bell Sys. Tech. J., Vol. 3, pp. 965 - 1004, May 1964.

14. C. Jacobaeus,
"A study on congestion in link systems", Ericsson Tech. (Sweden) No. 48, 1950.

15. K. Kummerle,
"An analysis of loss approximations for link systems", The 5th International Teletraffic Congress, New York, June 14 - 20, 1967.

16. L. Lee, and J.A. Brzozowski,
"An approximate method for computing blocking probability in switching network", IEEE on Communication Technology Vol. Com - 14, No. 2, 1966, pp. 85 - 93.

17. A. Papoulis,
"Probability, Random Variables, and Stochastic Processes", New York, McGraw-Hill Book Company, 1965.

18. D.G. Kendall,

"Some problems in the theory of queues", J.R. Statist. Soc. Ser. B, 13, pp. 151 - 185, 1951.

19. J. Riordan,

"Stochastic service systems", John Wiley and Sons, Inc., New York, 1962.

20. T.L. Saaty,

"Elements of queuing theory with applications", McGraw-Hill Book Co., Inc., New York, 1961.

21. D.R. Cox and W.L. Smith,

"Queues", John Wiley & Sons, Inc., New York, 1961.

22. L.E.N. Delbrouck and L. Lee,

"The anatomy of waiting time distributions in single server systems", IEEE Transactions on Communication Technology, Vol. Com - 15, No. 1, Feb. 1967, pp. 11 - 17.

23. R.A. Frazer, W.J. Duncan, and A.R. Collar,

"Elementary matrices", Cambridge: at the University Press, 1952.

24. A.K. Erlang,

"Solutions of some problems in the theory of probabilities of significance in automatic telephone exchanges", Post Off. Elect. Engrgs. J. (GB), 10, 4, pp. 189 - 197, 1917 - 1918.

25. A.Y. Khintchine,

"Mathematical methods in the theory of queuing", Hafner Publishing Company, New York, 1960.

26. J.W. Cohen,

"Certain delay problems for a full availability trunk group loaded by two traffic sources", Comm. News, Vol. 26, No. 3, April 1956.

27. R.I. Wilkinson,

"Theories for toll traffic engineering in the U.S.A.", Bell Sys. Tech. J. Vol. 35, pp. 421 - 514, March, 1956.

28. A. Rodriguez and J.P. Dartois,

"Small groups of subscribers traffic unbalance", The 5th International Teletraffic Congress, New York, June 14 - 20, 1967.

29. W. Helly,

"The balancing of loads among inputs to loss and delay networks", The 5th International Teletraffic Congress, New York, June 14 - 20, 1957.

VITA

Name: Lansun Lee

Born: Aug. 15, 1931, China

Educated:

Primary: China

Secondary: China and Regina, Saskatchewan,

University: (a) University of Saskatchewan, Saskatoon, Sask.
1957 Bachelor of Science in Engineering
Physics (B.E.).

(b) McGill University, Montreal, P.Q.
1960 Master of Engineering (M. Eng.)
Electrical Engineering.