A STUDY ON MESSAGE ROUTING AND CAPACITY ASSIGNMENT FOR STORE-AND-FORWARD COMPUTER COMMUNICATION NETWORKS

MASAHARU KOMATSU

DECEMBER 1977

A STUDY ON MESSAGE ROUTING AND CAPACITY ASSIGNMENT FOR STORE-AND-FORWARD COMPUTER COMMUNICATION NETWORKS

MASAHARU KOMATSU

DECEMBER 1977

ACKNOWLEDGEMENTS

The author would like to acknowledge the continuing guidance and encouragement of Professor Yoshikazu Tezuka througout this investigation.

The author would like to express his appreciation to Professor Kiyoyasu Itakura, Professor Toshihiko Namekawa, Professor Nobuaki Kumagai, and Professor Yoshiro Nakanishi.

Dozens of past and present members of Tezuka Laboratory have contributed in completing this thesis, and special appreciation goes to Associate Professor Hidehiko Sanada, Dr. Hikaru Nakanishi, Mr. Seiichi Uchinami, Dr. Yukuo Hayashida, Mr. Jun Ishii, Mr. Yasuhiro Ouchi, Mr. Taizo Nanbu, Mr.Nobuo Teraji, Mr.Takeshi Inoue, Mr. Yuji Shibata, Mr. Ichiro Akiyoshi, Mr. Tsutomu Iwahashi, and Mr. Yoji Komatsu, for their helpful suggestions and discussions.

To Mr. Tsuyoshi Nakatani, Mr. Kenji Kawamura, Mr.Satoshi Kageyama, Mr. Jun Taniguchi, Mr. Yoji Isota, Mr. Seiji Iwasaki, and Mr. Kazuhiko Chiba, I give my gratitude for proofreading.

ABSTRACT

This thesis considers message routing and channel capacity assignment problems for store-and-forward computer communication networks.

In Chaper 1, fundamental aspects of a computer communication network are presented. Furthermore, a review of the previous researches and the problems studied in this thesis are summarized.

In Chapter 2, a store-and-forward computer communication network is mathematically modeled as a simple queueing network. Using this model, many results are given later on.

Next, the optimum route assignment problem is formulated as a problem finding the optimum route assignment with the minimum total average message delay. Its solution is derived as the optimum route assignment theorem which gives the necessary and sufficient conditions to minimize the total average message delay.

Finally, detouring behavior of the optimum route assignment is compared with that of the equal-delay-principle route assignment, and the difference between them is clarified

In Chapter 3, a new adaptive routing procedure based on the optimum route assignment theorem is proposed. And, from simulation results, it is verified that the new procedure is able to select the route on which total message is transmitted by smaller delay than the ARPA procedure.

In Chapter 4, the optimum channel capacity assignment problem is formulated, and its solution is obtained as the optimum channel capacity assignment theorem, which gives the necessary and sufficient conditions to minimize the total average message delay in the case

of general message length.

From numerical results, the difference between the characteristics of the optimum channel capacity assignment and that of the most plausible assignment, i.e. the proportional channel capacity assignment, is clarified.

In Chapter 5, extended optimum channel capacity assignment problems are considered. The optimum channel capacity assignment problem as first given by Kleinrock is to minimize the total average message delay. From the Little's formula, this problem may be interpreted as a problem finding the channel capacity assignment to minimize the total number of messages within the network. Therefore, an extended optimum channel capacity assignment problem to reduce variation among queue lengths may be formulated. The solution to this problem is derived, and some interesting properties are clarified. Another extended problem is given by Meister et al., which is a problem to reduce variation among channel delays. A dual relation between these two extended problems is shown.

In Chapter 6, the overall conclusions obtained in this dissertaion are summarized.

CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Both computer technology and communication technology have been playing an extremely important role in human society. The close connection of these technologies produced a new information processing system which is called "computer network".

A computer network is generally defined as a set of autonomous, independent computer systems, interconnected so as to permit interactive resource sharing between any pair of systems [1]. As mentioned in the above definition, the main purpose of a computer network is to share the resources, i.e. data base, hardware, and software. Furthermore, it has another purposes such as high reliability and load sharing. At present, there are several computer networks in the world. However, they are still in the laboratory stage.

This dissertation mainly studies design problems for a computer communication network interconnecting large computer systems.

1.1 Computer Communication Network

The general configuration of a computer network is shown in Fig.1.1. A computer network consists of two subnetworks. One of them is a local network, which is also called a low level network, with large computers and terminals connected to them by low-speed channels. The large computers carry out the useful processing and storage tasks. The other is a computer communication network, mutually connected by high-speed data communication channels. The node computers carry out the communication oriented

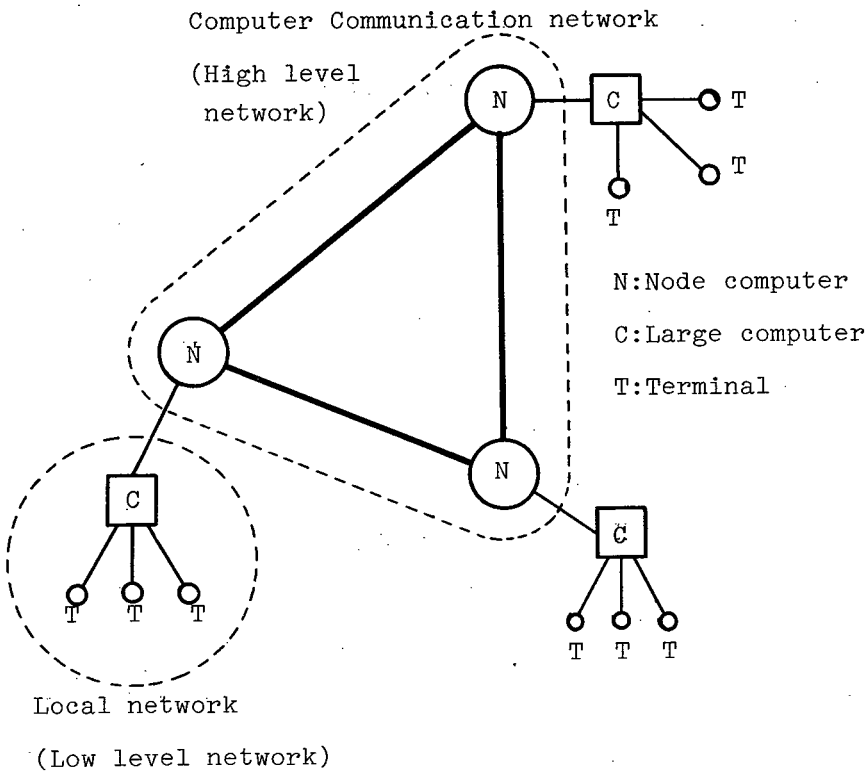

Fig.1.1 Computer network

task.

The topological configuration of communication networks may be divided into four types:

(1) Centralized network (Star network)

A number of nodes are connected to a central node with control function such as switching. The centralized network is superior in simplicity and cost, but is not very reliable.

The examples of this type are the COINS network, the NETWORK/440 [2], the OCTOPUS network of Lawrence Radiation Laboratory [3], and the TUCC network in North Carolina.

(2) Loop network [4,5]

In this network, nodes form a ring or a loop. This network has an advantage in cost and line length.

The example of this type is the DCS of University of California.

(3) Distributed network

Communication control functions are distributed to each node which is connected to many neighbouring nodes. The reliability of this network is high because there are several alternating routes between source node and destination node. In this network, routing procedure is needed.

The ARPA network [1,6,7,8,9] is the most representative network of this type. Another examples are the NPL network [9,10,11], the CYBERNET, and the MERIT computer network.

For computer communication networks, switching methods are classified into three main methods as follows:

(1) Circuit (Line)-switching

In this method, a complete path of connected lines is established from source node to destination node by a call before messages are transmitted.

(2) Message-switching

Message is transmitted from its source node to its destination node in store-and-forward fashion.

(3) Packet-switching

The packet-switching is basically the same as the message-switching, except that the message is decomposed into packets.

Store-and-forward switching is general term for message-switching and packet-switching.

In computer-computer communication, most of messages are interactive messages with short length. Therefore, the store-and-forward switching, especially packet-switching, is more suitable to the computer communication networks [12,13].

On the other hand, circuit-switching is more suitable to transmitting a long message such as a file message. Therefore, the hybrid-switching with both advantages of circuit- and packet-switching is also proposed [14,15].

In the future, as a large computer communication network, the distributed store-and-forward network is expected to develop.

In evaluating the computer communication network, several performance measures are considered:

(1) Message delay
(2) Throughput
(3) Cost
(4) Reliability

These must be considered in the following design problems of computer communication networks.

(1) Topological design
(2) Channel capacity assignment problem
(3) Route assignment problem
(4) Message routing procedure
(5) Flow control

Concerning the design of the optimum network, it is impossible to deal with these design problems simultaneously. Therefore, in general, these problems are considered independently.

1.2 Review of The Previous Researches

In this section, the brief review of the previous researches on store-and-forward computer communication networks is shown.

The earliest mathematically modeling and analysis of the store-and-forward network were given by Kleinrock [16] based on the results of Berk [17] and Jackson [18]. Moreover, the message length independence assumption was derived, and its validity was verified from simulation results [16]. As the result, each queueing unit within the network may be considered as an independent unit. Most of the analytical considerations of store-and-forward networks, especially message-switching networks are based on Kleinrock's modeling [See for example Miyahara [19]]. On the other hand, the analysis of a packet-switching network is extremely difficult. Approximate analysis of it is given by Fultz [20], Rubin[21,22,23], Okada [24,25], and Hashida [26].

The optimum design problem of store-and-forward network was formulated as a problem to achieve minimum total average message delay as a fixed cost by appropriately choosing the network topology, the channel capacity assignment, the message routing, and flow control [16]. However, these variables are mutually related, and the optimum design of a network is imposible in this sense. Therefore, in general, that problem is divided into some individually independent problems. Concerning the topological design of the network, optimum solution is not still found, but good suboptimum procedures designing the network topology are given by Doll[27] and Frank[28,29]. The optimum channel capacity assignment problem for a message-switching network is formulated and solved by Kleinrock [16]. Meister discussed the channel

capacity assignment reducing variation among channel delays [30], and further considered the case of nodal cost and capacities [31]. Frank [29] devised an optimum procedure for selecting discrete channel capacity for tree network.

The message routing problem is divided into two problems. One of them is a problem finding the optimum routes of messages in steady state. Concerning this problem, the basic concept of the maximum flow between source node and destination node was discussed by Frank [32], Ford [33], Rothfarb [34], and Sanada [35]. On the other hand, several algorithms finding the optimum set of routes of messages in order to minimize total average message delay were given by Frank [29], Cantor [36], Frata [37], and Schwartz [38]. The other is a design problem of routing procedure which is one of the most important problems in operational network. Prosser investigated the random routing [39] and the directory routing [40]. Boehm [41], Furtz [42], McQuillan [43], Rubin [44], and Butrimenko [45] examined or proposed adaptive routing procedures. Pickholtz [46] discussed the effect of priority discipline in routing.

Flow control is a technique to prevent congestion which is a major hazard to store-and-forward network. Pennotti [47] and Sanada [48] analyzed congestion phenomena. Kahn [49] and Herrman [50] showed the flow control method in the ARPA network. Davies [51] proposed the Isarithmic method, and the behavior is analyzed by Price [52,53] and Okada [54].

1.3 Research Problems

In this thesis, we study some problems as mentioned in Sec. 1.2. We summarize them as follows:

(1) Route assignment problem

The optimum route assignment problem is formulated as a problem finding the set of routes of messages to minimize total average message delay. The solution is given as the optimum route assignment theorem. Furthermore, the difference between the behavior of the optimum route assignment and that of the equal-delay-principle route assignment is discussed.

(2) Adaptive routing procedure

A new adaptive routing procedure based on the optimum route assignment theorem is proposed. And, from the simulation results, the superiority of our new procedure to the ARPA one is verified.

(3) Optimum channel capacity assignment problem

The optimum channel capacity assignment problem as first given by Kleinrock is to minimize the total average message delay. He solved this problem in the case of exponential message length. In this thesis, this problem is solved in the case of general message length. The solution is given as the optimum channel capacity assignment theorem which gives the necessary and sufficient conditions to minimize the total average message delay. From numerical results, the difference between the optimum channel capacity assignment and the most plausible assignment, i.e. the proportional channel capacity assignment, is considered.

(4) Extended optimum channel capacity assignment problem

Kleinrock's channel capacity assignment problem may be interpreted as a problem minimizing the number of messages within

the network. We extend this problem to a problem reducing variation among queue lengths, and its solution is obtained. On the other hand, Meister et al. extended it to a problem reducing variation among channel delays. It is found that these two assignments have a dual relation to each other.

CHAPTER 2

OPTIMUM ROUTE ASSIGNMENT PROBLEM

2.1 Introduction

One of the important problems for computer communication networks is a message routing problem, which is referred to as the optimum route assignment problem. The optimum route assignment problem is to find the optimum set of routes on which messages have to be transmitted in order to minimize total average message delay. By this time, various algorithms [36,37,38] have been proposed for solving this nonlinear optimization problem. The conditions to minimize the total average message delay are also well known in the single commodity case in which all messages are transmitted from the same source node to the same destination node, and the distribution of message length is exponential [35].

Store-and-forward computer communication networks may be mathematically modeled by queueing networks. The queueing network consists of a number of queueing units mutually connected in series or parallel. By introducing the message length independence assumption and the general independence assumption, we may consider each queueing unit as an independent queueing unit M/G/1.

In this chapter, we consider the optimum route assignment problem in the multi-commodity case in which there are many pairs of source node and destination node, and the distribution of message length is general. Concerning this problem, we derive "optimum route assignment theorem" which gives the necessary and sufficient conditions to obtain the optimum route assignment minimizing the total average message delay.

Furthermore, we consider a simple multiple-channel model. By applying the optimum route assignment theorem, the optimum route assignment is derived for this model. And, from numerical results, the difference between the optimum route assignment and the equal-delay-principle route assignment is clarified.

2.2 Mathematical model of Store-and-Forward Computer Communication Networks

First, for mathematically modeling the store-and-forward switching communication networks such as message-switching networks and packet-switching networks, we introduce the elementary concepts associated with the network. The store-and-forward switching network consists of a number of nodes connected to each other by channels as shown in Fig.2.1. The node is a switching center. The message (or packet) is specified by its source node and destination node, length and priority class. In this network, a message (or packet) is transmitted from its source node to its destination node in store-and-forward fashion: As an example, suppose that a message originate at node 1 and is destined for node 8. This message originate at node 1, that is, enters the network at node 1 from the outside of the network. Upon origination of the message at node 1, the nodal processor receives it, and must make a decision as to whether to send the message to node 2 or node 5, that is, as to whether to send the message via channel 1 or channel 2. The decision rule is referred to as the routing procedure. After the routing decision, the message joins the queue coressponding to the assigned route or channel, say channel 2. If the channel 2 is busy, the message must wait in the queue before the channel

becomes available. And, if the channel is empty, the message is transmitted to node 5 immediately. When the message arrives at node 5 via channel 2, the same process as at node 1 is performed. That is, the store-and-forward process as above is repeated at each node belonging the route from the source node to destination node until the message arrives at the destination node. Eventually the message leaves the network at the destination node 8.

Therefore, the store-and-forward communication network may be mathematically modeled by the queueing network which consists of a number of queueing units mutually connected in series or parallel. Figure 2.2 shows a part of the queueing network corresponding to the store-and-forward communication network as shown in Fig.2.1. The queueing unit consists of a single server and a waiting room. The former is a channel, and the latter is a buffer in which after the routing decision, the messages (or packets) wait until the channel becomes available. It is extremely difficult to analyze this model because it leads to a rather complex mathematical model in which the permanent assignment of length to each message gives rise to a dependency between the interarrival time and length of adjacent messages as they travel through the network. However, by introducing the message length independence assumption [16] and the general independence assumption[20], we may consider the queueing unit as an independent unit. These two assumptions are as follows:

The message length independence assumption

Each time a message is received at a node within the network, a new length v is chosen for this message from the following probability function.

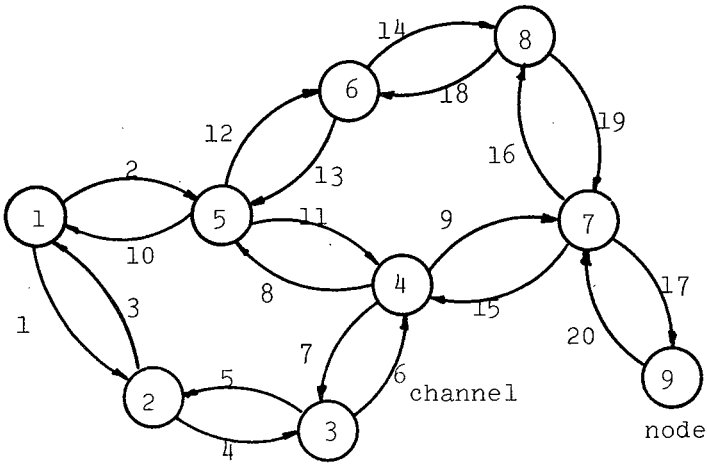

Fig.2.1 Store-and-forward Network

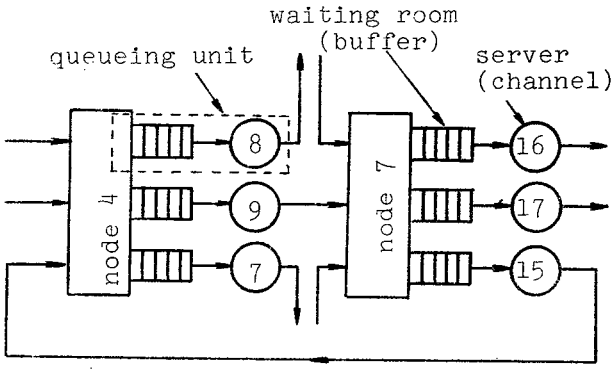

Fig.2.2 Mathematical model of store-and-forward network

$$f(v)=\mu e^{-\mu v}$$

Of course, in this assumption, we assume that the message length is exponential. Using this assumption, the queueing unit can be modeled as M/M/1.

On the other hand, the following assumption is useful in the case of general message length.

The general independence assumption†

Assume that the message interarrival times at each node within the network are Poisson.

Using this assumption, the delay for any channel can be computed from Pollaczek-Khinchin formula [55], assuming that the queueing unit is characterized as M/G/1.

Before proceeding, we define and list below some of the important quantities and symbols.

B_i	;the i-th channel within the network
N_i	;the i-th node within the network
$R_j(s,d)$	;the j-th route from source node N_s to destination node N_d
N	;number of channels within the network
M	;number of nodes within the network
$n(s,d)$	;number of routes from N_s to N_d
C_i	;channel capacity of B_i (bits/sec)
λ_i	;average traffic rate at B_i (messages/sec)
$1/\mu_i$	;average message length for B_i (bits)
ρ_i	;average channel utilization of B_i

† See APPENDIX A

T_i ;average channel delay on B_i (average delay for a message passing through B_i, which includes both time on queue and time in transmission) (sec)

L_i ;average queue length on B_i (includes number of messages on queue and number of messages in transmission) (messages)

γ_{sd} ;average arrival rate of messages with source node N_s and destination node N_d (messages/sec)

Z_{sd} ;average message delay of messages with source node N_s and destination node N_d (sec)

T ;total average message delay over a entire network (sec)

γ ;total message arrival rate from external sources to the network (messages/sec)

Having these definitions, several relations can be stated which will be used later on.

$$\gamma = \sum_{s,d} \gamma_{sd} \tag{2.1}$$

$$\rho_i = \frac{\lambda_i}{\mu_i C_i} \tag{2.2}$$

$$T = \sum_{s,d} \frac{\gamma_{sd}}{\gamma} Z_{sd} = \sum_{i} \frac{\lambda_i}{\gamma} T_i \tag{2.3}^{\dagger}$$

And, from well known Little's formula

$$L_i = \lambda_i T_i \tag{2.4}$$

---

† See APPENDIX B

T can be rewritten as

$$T = \sum_i \frac{L_i}{\gamma} \tag{2.5}$$

Furthermore, the average delay T_i is given by

$$T_i = \frac{1}{\mu_i C_i - \lambda_1} \qquad \text{for M/M/1} \tag{2.6}$$

$$T_i = \frac{1}{\lambda_i}\left[\rho_i + \frac{\rho_i^2 + \lambda_i^2 \sigma_i^2 / C_i^2}{2(1-\rho_i)}\right] \qquad \text{for M/G/1} \tag{2.7}$$

where σ_i^2 is the variance of message length in B_i.

For computing the average channel delay T_i, we use Eqs.(2.6) and (2.7) for exponential message length and for general message length respectively.

2.3 Optimum Route Assignment Theorem for Store-and-Forward Computer Communication Networks

In previous section, we mathematically modeled the store-and-forward communication network, defined the quantities and symbols, and stated some relations.

In this section, the optimum route assignment problem for store-and-forward communication network is formulated, and solved. The solution is referred to as the optimum route assignment theorem, which gives the necessary and sufficient conditions to find the optimum route assignment minimizing total average message delay.

2.3.1 Optimum Route Assignment Problem

The optimum route assignment problem for store-and-forward computer communication networks is to determine the optimum routing, i.e. the optimum set of routes on which messages have to be transmitted to optimize a well defined objective function. The objective functions are total average message delay, cost, or throughput, etc. In this section, we use the total average message delay as the objective function.

Furthermore, we assume that the network topology, traffic rate γ_{sd}, and channel capacity C_i are given.

Let define x_j^{sd} as route assignment probability which gives the probability of assignment of traffic γ_{sd} to route $R_j(s,d)$. x_j^{sd} must satisfy

$$\sum_{j=1}^{n(s,d)} x_j^{sd}=1 \ , \quad x_j^{sd}\geq 0 \tag{2.8}$$

And, λ_i is given by

$$\lambda_i = \sum_{j,s,d \mid B_i \epsilon R_j(s,d)} x_j^{sd}\gamma_{sd} \tag{2.9}$$

Concerning the optimum route assignment problem, it is important to derive the conditions which x_j^{sd} must satisfy to minimize the total average message delay.

Thus, the optimum route assignment problem may be formulated as follows:

Optimum Route Assignment Problem

Given: network topology

traffic rate γ_{sd}

channel capacity C_i

Minimize: total average message delay T

With respect to: route assignment probability $[x_j^{sd}]$

Under constraint: channel utilization $\rho_i < 1$

2.3.2 Optimum Route Assignment Theorem

The solution to the optimum route assignment problem formulated in the previous section is given by optimum route assignment theorem as stated in;

Optimum Route Assignment Theorem

The solution $[x_j^{sd}]$ to the optimum route assignment problem is optimum if and only if

$$\sum_{i|B_i \epsilon R_j(s,d)} \frac{\partial L_i}{\partial \lambda_i} \begin{cases} = D(s,d) & ; x_j^{sd} > 0 \\ \geq D(s,d) & ; x_j^{sd} = 0 \end{cases} \qquad (2.10)$$

for all source node N_s and destination node N_d, where D(s,d) is a constant value for a pair of N_s and N_d.

Proof. The optimum route assignment problem can be reformulated as the equivalent optimum route assignment problem given as follows:

Equivalent optimum route assignment problem

Objective function: $S(x) = -T = -\sum_i L_i/\gamma \rightarrow \max.$ (2.11)

With respect to: x

Under constraint:

$$K(x) = [g_1(x), \cdots, g_N(x), h^{1,2}(x), \cdots, h^{sd}(x), \cdots,$$
$$h^{M,M-1}(x), f_1^{1,2}(x), \cdots, f_j^{sd}(x), \cdots,$$
$$f_{n(M,M-1)}^{M,M-1}(x)] \leq 0 \qquad (2.12)$$

where

$$g_i(x)=\rho_i-1 \qquad i=1,2,\cdots,N \tag{2.13}$$

$$h^{sd}(x)=x_1^{sd}+x_2^{sd}+\cdots+x_{n(s,d)-1}^{sd}-1 \qquad s,d=1,2,\cdots,M \tag{2.14}$$

$$f_j^{sd}(x)=-x_j^{sd} \qquad j=1,2,\cdots,n(s,d)-1 \quad s,d=1,2,\cdots,M \tag{2.15}$$

$$x=(x_1,x_2,\cdots,x_N) \tag{2.16}$$

$$x_i=(x_1^{i,1},x_2^{i,1},\cdots,x_{n(i,1)-1}^{i,1},x_1^{i,2},x_2^{i,2},\cdots, x_{n(i,2)-1}^{i,2},\cdots,x_1^{i,i-1},x_2^{i,i-1},\cdots,x_{n(i,i-1)-1}^{i,i-1}, x_1^{i,i+1},x_2^{i,i+1},\cdots,x_{n(i,i+1)-1}^{i,i+1},\cdots, x_1^{i,M},x_2^{i,M},\cdots,x_{n(i,M)-1}^{i,M}) \tag{2.17}$$

For the above problem, we use Lagrange function

$$\phi(x,\eta)=S(x)-\eta\cdot K(x) \tag{2.18}$$

where η is Lagrange multiplier given as follows:

$$\eta=(\alpha_1,\alpha_2,\cdots,\alpha_N,\beta_{1,2},\beta_{1,3},\cdots,\beta_{M,M-1}, \delta_1^{1,2},\delta_2^{1,2},\cdots,\delta_{n(M,M-1)-1}^{M,M-1}) \tag{2.19}$$

From Kuhn-Tucker theorem [56], the necessary and sufficient conditions for (x^o,η^o) to maximize S under constraint Eq.(2.12) are as follows:

(i) Necessary conditions

$$\nabla_x \phi|_{x^o,\eta^o} \le 0, \quad (\nabla_x \phi, x)_{x^o,\eta^o} = 0, \quad x^o \ge 0 \tag{2.20}$$

$$\nabla_\eta \phi|_{x^o,\eta^o} \ge 0, \quad (\nabla_\eta \phi, \eta)_{x^o,\eta^o} = 0, \quad \eta^o \ge 0 \tag{2.21}$$

(ii) Sufficient conditions

$$\phi(x,\eta^o) \le \phi(x^o,\eta^o) + (\nabla_x \phi|_{x^o,\eta^o}, x-x^o) \tag{2.22}$$

$$\phi(x^o,\eta) \ge \phi(x^o,\eta^o) + (\nabla_\eta \phi|_{x^o,\eta^o}, \eta-\eta^o) \tag{2.23}$$

Later on, x^o and η^o will be omitted.

First, we consider the necessay conditions, i.e. Eqs.(2.20) and (2.21). Since T_i given by Eqs.(2.6) or (2.7) is differentiable with respect to λ_i, and the relation between T_i and L_i is given by Eq.(2.4), L_i is a differentiable function with respect to λ_i. Furthermore, from Eq.(2.9), it is easily recognized that λ_i is a linear function of the vector x with components $x_j^{sd} \ge 0$. Therefore, the partial derivative of L_i by x_j^{sd} is given by

$$\frac{\partial L_i}{\partial x_j^{sd}} = \frac{\partial L_i}{\partial \lambda_i} \frac{\partial \lambda_i}{\partial x_j^{sd}}$$

$$= \begin{cases} 0 & ; B_i \notin R_j(s,d),\ B_i \notin R_{n(s,d)}(s,d) \\ \gamma_{sd} \dfrac{\partial L_i}{\partial \lambda_i} & ; B_i \in R_j(s,d),\ B_i \notin R_{n(s,d)}(s,d) \\ 0 & ; B_i \in R_j(s,d),\ B_i \in R_{n(s,d)}(s,d) \\ -\gamma_{sd} \dfrac{\partial L_i}{\partial \lambda_i} & ; B_i \notin R_j(s,d),\ B_i \in R_{n(s,d)}(s,d) \end{cases} \tag{2.24}$$

And, ρ_i is also differentiable with respect to λ_i since ρ_i is a linear function of λ_i as given by Eq.(2.2). Thus, the partial derivative of g_i by x_j^{sd} is given as follows:

$$\frac{\partial g_i}{\partial x_j^{sd}} = \frac{\partial g_i}{\partial \lambda_i} \frac{\partial \lambda_i}{\partial x_j^{sd}}$$

$$= \begin{cases} 0 & ;B_i \notin R_j(s,d),\ B_i \notin R_{n(s,d)}(s,d) \\ \dfrac{1}{\mu_i C_i}\gamma_{sd} & ;B_i \in R_j(s,d),\ B_i \notin R_{n(s,d)}(s,d) \\ 0 & ;B_i \in R_j(s,d),\ B_i \in R_{n(s,d)}(s,d) \\ -\dfrac{1}{\mu_i C_i}\gamma_{sd} & ;B_i \notin R_j(s,d),\ B_i \in R_{n(s,d)}(s,d) \end{cases} \tag{2.25}$$

Furthermore, the partial derivatives of h^{kr} and f_k^{rm} by x_j^{sd} are easily obtained as follows:

$$\frac{\partial h^{kr}}{\partial x_j^{sd}} = \begin{cases} 1 & ;k=s,\ r=d \\ 0 & ;\text{elsewhere} \end{cases} \tag{2.26}$$

$$\frac{\partial f_k^{rm}}{\partial x_j^{sd}} = \begin{cases} -1 & ;k=j,\ r=s,\ m=d \\ 0 & ;\text{elsewhere} \end{cases} \tag{2.27}$$

Therefore, the first condition of the necessary conditions, i.e. Eq.(2.20), is the same as the following equations.

$$\frac{\partial \phi}{\partial x_j^{sd}} = -\frac{1}{\gamma} \sum_{i \left| \begin{array}{l} B_i \epsilon R_j(s,d) \\ B_i \notin R_{n(s,d)}(s,d) \end{array} \right.} \gamma_{sd} \frac{\partial L_i}{\partial \lambda_i}$$

$$+\frac{1}{\gamma} \sum_{i \left| \begin{array}{l} B_i \notin R_j(s,d) \\ B_i \epsilon R_{n(s,d)}(s,d) \end{array} \right.} \gamma_{sd} \frac{\partial L_i}{\partial \lambda_i}$$

$$- \sum_{i \left| \begin{array}{l} B_i \epsilon R_j(s,d) \\ B_i \notin R_{n(s,d)}(s,d) \end{array} \right.} \alpha_i \frac{\gamma_{sd}}{\mu_i C_i}$$

$$+ \sum_{i \left| \begin{array}{l} B_i \notin R_j(s,d) \\ B_i \epsilon R_{n(s,d)}(s,d) \end{array} \right.} \alpha_i \frac{\gamma_{sd}}{\mu_i C_i}$$

$$- \beta_{sd} + \delta_j^{sd} \leq 0 \qquad (2.28)$$

$$\nabla_x \phi = \sum_{j,s,d} x_j^{sd} \frac{\partial \phi}{\partial x_j^{sd}} = 0 \qquad (2.29)$$

$$x_j^{sd} \geq 0 \qquad (2.30)$$

The above three equations must be satisfied simultaneously. Therefore, if $x_j^{sd} > 0$, then

$$-\frac{\gamma_{sd}}{\gamma}\sum_{i\left|\begin{matrix}B_i\varepsilon R_j(s,d)\\ B_i\notin R_{n(s,d)}(s,d)\end{matrix}\right.}\frac{\partial L_i}{\partial \lambda_i}+\frac{\gamma_{sd}}{\gamma}\sum_{i\left|\begin{matrix}B_i\notin R_j(s,d)\\ B_i\varepsilon R_{n(s,d)}(s,d)\end{matrix}\right.}\frac{\partial L_i}{\partial \lambda_i}$$

$$-\gamma_{sd}\sum_{i\left|\begin{matrix}B_i\varepsilon R_j(s,d)\\ B_i\notin R_{n(s,d)}(s,d)\end{matrix}\right.}\frac{\alpha_i}{\mu_i C_i}+\gamma_{sd}\sum_{i\left|\begin{matrix}B_i\notin R_j(s,d)\\ B_i\varepsilon R_{n(s,d)}(s,d)\end{matrix}\right.}\frac{\alpha_i}{\mu_i C_i}$$

$$-\beta_{sd}+\delta_j^{sd}=0 \qquad (2.31)$$

and, if $x_j^{sd}=0$, then

$$-\frac{\gamma_{sd}}{\gamma}\sum_{i\left|\begin{matrix}B_i\varepsilon R_j(s,d)\\ B_i\notin R_{n(s,d)}(s,d)\end{matrix}\right.}\frac{\partial L_i}{\partial \lambda_i}+\frac{\gamma_{sd}}{\gamma}\sum_{i\left|\begin{matrix}B_i\notin R_j(s,d)\\ B_i\varepsilon R_{n(s,d)}(s,d)\end{matrix}\right.}\frac{\partial L_i}{\partial \lambda_i}$$

$$-\gamma_{sd}\sum_{i\left|\begin{matrix}B_i\varepsilon R_j(s,d)\\ B_i\notin R_{n(s,d)}(s,d)\end{matrix}\right.}\frac{\alpha_i}{\mu_i C_i}+\gamma_{sd}\sum_{i\left|\begin{matrix}B_i\notin R_j(s,d)\\ B_i\varepsilon R_{n(s,d)}(s,d)\end{matrix}\right.}\frac{\alpha_i}{\mu_i C_i}$$

$$-\beta_{sd}+\delta_j^{sd}\leq 0 \qquad (2.32)$$

The partial derivatives of Lagrange function ϕ by the Lagrange multipliers, i.e. α_i, β_{sd}, and δ_j^{sd}, are given as follows:

$$\frac{\partial \phi}{\partial \alpha_i}=-\left(\frac{\lambda_i}{\mu_i C_i}-1\right) \qquad (2.33)$$

$$\frac{\partial \phi}{\partial \beta_{sd}} = -(x_1^{sd} + x_2^{sd} + \cdots + x_{n(s,d)-1}^{sd} - 1) \tag{2.34}$$

$$\frac{\partial \phi}{\partial \delta_j^{sd}} = x_j^{sd} \tag{2.35}$$

Therefore, the second condition (Eq.(2.21)) of the necessary conditions implies that the optimum solution x must satisfy the following equations.

$$-\left(\frac{\lambda_i}{\mu_i C_i} - 1\right) \geq 0 \tag{2.36}$$

$$-(x_1^{sd} + x_2^{sd} + \cdots + x_{n(s,d)}^{sd} - 1) \geq 0 \tag{2.37}$$

$$x_j^{sd} \geq 0 \tag{2.38}$$

$$-\sum_i \alpha_i \left(\frac{\lambda_i}{\mu_i C_i} - 1\right) - \sum_{s,d} \beta_{sd}(x_1^{sd} + x_2^{sd} + \cdots + x_{n(s,d)-1}^{sd} - 1) + \sum_{j,s,d} \delta_j^{sd} x_j^{sd} = 0 \tag{2.39}$$

$$\alpha_i \geq 0, \quad \beta_{sd} \geq 0, \quad \delta_j^{sd} \geq 0 \tag{2.40}$$

From Eqs.(2.36)-(2.40), it is easily recognized that;

(a) If $\lambda_i/\mu_i C_i \to 1$, then $L_i \to \infty$, i.e. $S \to -\infty$. Thus, α_i must be equal to zero.

(b) If $x_1^{sd} + x_2^{sd} + \cdots + x_{n(s,d)-1}^{sd} - 1 = 0$, i.e. $x_{n(s,d)}^{sd} = 0$, then $\beta_{sd} \geq 0$, and if $x_1^{sd} + x_2^{sd} + \cdots + x_{n(s,d)}^{sd} - 1 < 0$, i.e. $x_{n(s,d)} > 0$, then $\beta_{sd} = 0$.

(c) If $x_j^{sd}=0$, then $\delta_j^{sd}\geq 0$, and if $x_j^{sd}>0$, then $\delta_j^{sd}=0$.

Thus, from the above conditions (a), (b) and (c), Eqs.(2.31) (2.32) may be divided into four cases as follows:

(1) If $x_j^{sd}>0$ and $x_{n(s,d)}^{sd}>0$, then

$$-\frac{\gamma_{sd}}{\gamma}\sum_{i\left|\begin{matrix}B_i\varepsilon R_j(s,d)\\ B_i\notin R_{n(s,d)}(s,d)\end{matrix}\right.}\frac{\partial L_i}{\partial \lambda_i}+\frac{\gamma_{sd}}{\gamma}\sum_{i\left|\begin{matrix}B_i\notin R_j(s,d)\\ B_i\varepsilon R_{n(s,d)}(s,d)\end{matrix}\right.}\frac{\partial L_i}{\partial \lambda_i}=0 \qquad (2.41)$$

(2) If $x_j^{sd}>0$ and $x_{n(s,d)}^{sd}=0$, then

$$-\frac{\gamma_{sd}}{\gamma}\sum_{i\left|\begin{matrix}B_i\varepsilon R_j(s,d)\\ B_i\notin R_{n(s,d)}(s,d)\end{matrix}\right.}\frac{\partial L_i}{\partial \lambda_i}+\frac{\gamma_{sd}}{\gamma}\sum_{i\left|\begin{matrix}B_i\notin R_j(s,d)\\ B_i\varepsilon R_{n(s,d)}(s,d)\end{matrix}\right.}\frac{\partial L_i}{\partial \lambda_i}=\beta_{sd} \qquad (2.42)$$

(3) If $x_j^{sd}=0$ and $x_{n(s,d)}^{sd}>0$, then

$$-\frac{\gamma_{sd}}{\gamma}\sum_{i\left|\begin{matrix}B_i\varepsilon R_j(s,d)\\ B_i\notin R_{n(s,d)}(s,d)\end{matrix}\right.}\frac{\partial L_i}{\partial \lambda_i}+\frac{\gamma_{sd}}{\gamma}\sum_{i\left|\begin{matrix}B_i\notin R_j(s,d)\\ B_i\varepsilon R_{n(s,d)}(s,d)\end{matrix}\right.}\frac{\partial L_i}{\partial \lambda_i}\leq -\delta_j^{sd} \qquad (2.43)$$

(4) If $x_j^{sd}=0$ and $x_{n(s,d)}^{sd}=0$, then

$$-\frac{\gamma_{sd}}{\gamma}\sum_{i\left|\begin{matrix}B_i\varepsilon R_j(s,d)\\ B_i\notin R_{n(s,d)}(s,d)\end{matrix}\right.}\frac{\partial L_i}{\partial \lambda_i}+\frac{\gamma_{sd}}{\gamma}\sum_{i\left|\begin{matrix}B_i\notin R_j(s,d)\\ B_i\varepsilon R_{n(s,d)}(s,d)\end{matrix}\right.}\frac{\partial L_i}{\partial \lambda_i}\leq \beta_{sd}-\delta_j^{sd} \qquad (2.44)$$

Even if we appropriately choose the n(s,d)-th route $R_{n(s,d)}(s,d)$

to satisfy that $x^{sd}_{n(s,d)}>0$, the generality of the proof is not lost.

And, concerning the summation for the channels B_i, the following relations exist.

$$\sum_{i\,|\,B_i\in R_j(s,d),\ B_i\notin R_{n(s,d)}(s,d)} + \sum_{i\,|\,B_i\in R_j(s,d),\ B_i\in R_{n(s,d)}(s,d)} = \sum_{i\,|\,B_i\in R_j(s,d)} \tag{2.45}$$

$$\sum_{i\,|\,B_i\notin R_j(s,d),\ B_i\in R_{n(s,d)}(s,d)} + \sum_{i\,|\,B_i\in R_j(s,d),\ B_i\in R_{n(s,d)}(s,d)} = \sum_{i\,|\,B_i\in R_{n(s,d)}(s,d)} \tag{2.46}$$

From Eqs.(2.41),(2.43),(2.45), and (2.46), it is recognized that

$$\sum_{i\,|\,B_i\in R_j(s,d)} \frac{\partial L_i}{\partial \lambda_i} = \sum_{i\,|\,B_i\in R_{n(s,d)}(s,d)} \frac{\partial L_i}{\partial \lambda_i} \quad ; x_j^{s,d}>0 \tag{2.47}$$

$$\sum_{i\,|\,B_i\in R_j(s,d)} \frac{\partial L_i}{\partial \lambda_i} = \sum_{i\,|\,B_i\in R_{n(s,d)}(s,d)} \frac{\partial L_i}{\partial \lambda_i} + \delta_j^{sd}$$

$$\geq \sum_{i\,|\,B_i\in R_{n(s,d)}(s,d)} \frac{\partial L_i}{\partial \lambda_i} \quad ; x_j^{sd}=0 \tag{2.48}$$

We define that

$$D(s,d)= \sum_{i\,|\,B_i\in R_{n(s,d)}(s,d)} \frac{\partial L_i}{\partial \lambda_i} \tag{2.49}$$

Finally, Eq.(2.10) is obtained from Eqs.(2.47),(2.48), and (2.49).

Next,we consider the sufficient condition, i.e. Eqs.(2.22) and (2.23). These equations imply that the objective function S must be a continuous, differentiable and concave function of vector x, and the constraint functions must be convex functions of x. The objective function Eq.(2.11)† and the constraint functions Eqs. (2.13) and (2.15) clearly satisfy these conditions. Q.E.D.

The optimum route assignment theorem gives the necessary and sufficient conditions to minimize the total average message delay by using the partial derivative of L_i with respect to λ_i. Since the relation between L_i and T_i is given by Little's formula Eq. (2.4), we can also write the necessary and sufficient conditions by useing T_i.

From Eq.(2.4), we have

$$\frac{\partial L_i}{\partial \lambda_i} = T_i + \lambda_i \frac{\partial T_i}{\partial \lambda_i} \tag{2.50}$$

Thus, the necessary and sufficient conditions given by Eq.(2.10) may be rewritten as follows:

$$\sum_{i|B_i \epsilon R_j(s,d)} (T_i + \lambda_i \frac{\partial T_i}{\partial \lambda_i}) \begin{cases} =D(s,d) & ; x_j^{sd}>0 \\ \geq D(s,d) & ; x_j^{sd}=0 \end{cases} \tag{2.51}$$

for all N_s and N_d

Assuming that the message length distribution is general, $\partial L_i/\partial \lambda_i$

---

† See APPENDIX C.

in Eq.(2.10) is given by

$$\frac{\partial L_i}{\partial \lambda_i} = \frac{1}{\mu_i C_i}\left[1 + \frac{\rho_i(2-\rho_i)(1+\mu_i^2\sigma_i^2)}{2(1-\rho_i)^2}\right] \tag{2.52}$$

Especially, when the message length is Erlangian with phase k, $\sigma_i^2 = 1/k\mu_i^2$. Thus, $\partial L_i/\partial \lambda_i$ is written by

$$\frac{\partial L_i}{\partial \lambda_i} = \frac{1}{\mu_i C_i}\left[1 + \frac{\rho_i(2-\rho_i)(1+1/k)}{2(1-\rho_i)^2}\right] \tag{2.53}$$

For k=1, the message length is exponential, and

$$\frac{\partial L_i}{\partial \lambda_i} = \frac{1}{\mu_i C_i (1-\rho_i)^2} \tag{2.54}$$

For k=∞, the message length is constant, and

$$\frac{\partial L_i}{\partial \lambda_i} = \frac{1}{\mu_i C_i}\left[1 + \frac{\rho_i(2-\rho_i)}{2(1-\rho_i)^2}\right] \tag{2.55}$$

Let us consider some other plausible route assignment. An intuitively reasonable assignment is the assignment based on the equal-delay-principle as follows:

Equal-delay-principle route assignment [57]

The route assignment probabilities for the equal-delay-principle route assignment are decided so that

$$\sum_{i|B_i \varepsilon R_j(s,d)} T_i \begin{cases} = E(s,d) & ; x_j^{sd} > 0 \\ \geq E(s,d) & ; x_j^{sd} = 0 \end{cases} \qquad (2.56)$$

for all N_s and N_d

where $E(s,d)$ is a constant value of the average delay for messages from the source node N_s to the destination node N_d.

2.3.3 Optimum Route Assignment for Multiple-Channel

In this section, we consider a simple multiple-channel model as shown in Fig.2.3, and compare the optimum route assignment to the equal-delay-principle route assignment.

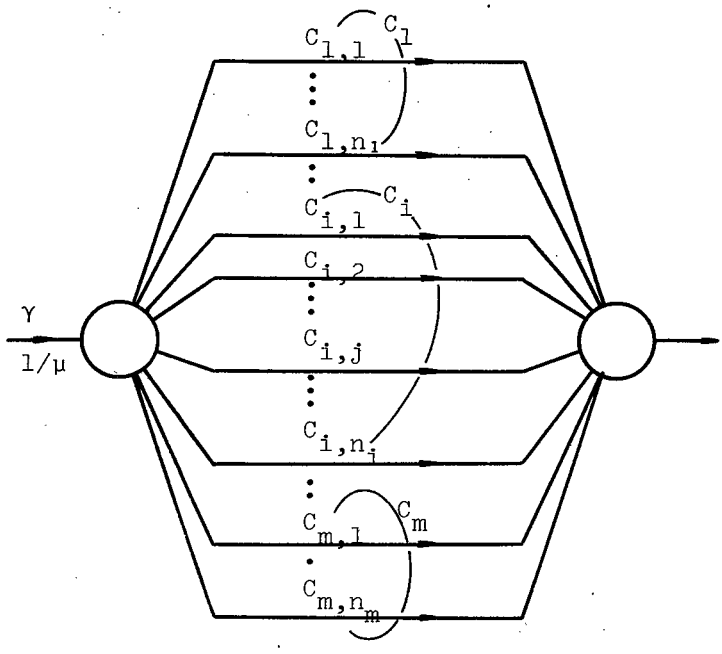

Fig.2.3 Multiple-channel model

In Fig.2.3, there are m classes of channels with capacity C_i where $C_1>C_2>\cdots>C_m$. The i-th class has n_i channels, and the j-th channel of class i has capacity C_{ij}, that is, $C_{ij}=C_i$ $(j=1,2,\cdots, n_i)$. Furthermore, we assume that the message length is exponential with mean $1/\mu$.

On the above assumption, the optimum route assignment and the equal-delay-principle route assignment are obtained as follows:

Optimum route assignment

For $\gamma_{k-1}^{o}\leq\gamma<\gamma_k^{o}$,

$$\lambda_{ij}=\begin{cases}\mu C_i-\dfrac{\sqrt{C_i}}{\sum_{r=1}^{k} n_r\sqrt{C_r}}\left(\sum_{u=1}^{k} n_u\mu C_u-\gamma\right); & i=1,2,\cdots,k \\ & j=1,2,\cdots,n_i \\ 0 & ;i=k+1,\cdots,m,\ j=1,2,\cdots,n_i\end{cases} \tag{2.57}$$

where

$$\gamma_k^{o}=\begin{cases}\mu\left(\sum_{i=1}^{k} n_iC_i-\sqrt{C_{k+1}}\sum_{i=1}^{k} n_i\sqrt{C_i}\right) & ;i=1,2,\cdots,m-1 \\ 0 & ;i=0 \\ \mu\sum_{i=1}^{m} n_iC_i & ;i=m\end{cases} \tag{2.58}$$

Proof. The channel can be mathematically modeled by queueing unit M/M/1. Thus, the average queue length L_{ij} on the channel C_{ij} is given by

$$L_{ij}=\frac{\lambda_{ij}}{\mu C_{ij}-\lambda_{ij}} \tag{2.59}$$

where λ_{ij} is the message arrival rate on the channel C_{ij}. From

Eqs.(2.10) and (2.59), the necessary and sufficient conditions to minimize the total average message delay are given by

$$\frac{\mu C_{ij}}{(\mu C_{ij}-\lambda_{ij})^2} = D \qquad ;\lambda_{ij}>0$$

$$\frac{1}{\mu C_{ij}} \geq D \qquad ;\lambda_{ij}=0 \tag{2.60}$$

From Eq.(2.60), it is easily recognized that if, for certain traffic rate γ, the channels of the classes $1,2,\cdots,k$ are used and the other classes are not used, then the following equations must be satisfied.

$$\frac{\mu C_{ij}}{(\mu C_{ij}-\lambda_{ij})^2} = D \qquad ;i=1,2,\cdots,k \quad j=1,2,\cdots,n_i \tag{2.61}$$

$$\frac{1}{\mu C_{ij}} \geq D \qquad ;i=k+1,\cdots,m \quad j=1,2,\cdots,n_i \tag{2.62}$$

From Eq.(2.61), we obtain

$$\mu C_{ij}-\lambda_{ij}=\sqrt{\mu C_{ij}/D}$$

$$\lambda_{ij}=\mu C_{ij}-\sqrt{\mu C_{ij}/D} = \mu C_i-\sqrt{\mu C_i/D} \tag{2.63}$$

Summing Eq.(2.63) on i and j, we find

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\lambda_{ij} = \gamma = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\mu C_i-\sqrt{\mu C_i/D}\right)$$

$$= \sum_{i=1}^{k} n_i\mu C_i - \frac{1}{\sqrt{D}}\sum_{i=1}^{k} n_i\sqrt{\mu C_i} \tag{2.64}$$

from which we obtain

$$D=\left[\frac{\sum_{i=1}^{k} n_i\sqrt{\mu C_i}}{\sum_{i=1}^{k} n_i\mu C_i-\gamma}\right]^2 \qquad (2.65)$$

Substituting Eq.(2.65) into Eq.(2.63), we arrive at

$$\lambda_{ij}=\mu C_i-\frac{\sqrt{C_i}}{\sum_{r=1}^{k} n_r\sqrt{C_r}}\left(\sum_{u=1}^{k} n_u\mu C_u-\gamma\right) \quad ; \begin{matrix} i=1,2,\cdots,k \\ j=1,2,\cdots,n_i \end{matrix}$$

From Eq.(2.65), D is a monotone increasing function of γ, and $\mu C_{ij}/(\mu C_{ij}-\lambda_{ij})^2$ is also a monotone increasing function of λ_{ij}. Therefore, from Eqs.(2.60) and (2.62), it is found that γ_k^o (hereafter referred to as detour traffic rate) is the traffic rate γ so that

$$\frac{1}{\mu C_{ij}}=D \qquad ;i=k+1 \qquad (2.66)$$

where D is given by Eq.(2.65). Thus,

$$\frac{1}{\mu C_{k+1}}=\left[\frac{\sum_{i=1}^{k} n_i\sqrt{\mu C_i}}{\sum_{i=1}^{k} n_i\mu C_i-\gamma_k^o}\right]^2 \qquad (2.67)$$

from which we arrive at

$$\gamma_k^o=\mu\left(\sum_{i=1}^{k} n_i C_i-\sqrt{C_{k+1}}\sum_{i=1}^{k} n_i\sqrt{C_i}\right)$$

Moreover, γ_m^o is the traffic rate γ so that

$$\rho = \frac{\gamma}{\mu \sum_{i=1}^{m} \sum_{j=1}^{n_i} C_{ij}} = 1 \tag{2.68}$$

from which we arrive at

$$\gamma_m^o = \mu \sum_{i=1}^{m} n_i C_i$$

Q.E.D.

Equal-delay-principle route assignment

For $\gamma_{k-1}^E \leq \gamma < \gamma_k^E$,

$$\lambda_{ij} = \begin{cases} \mu C_i - \dfrac{\sum_{i=1}^{k} n_i \mu C_i - \gamma}{\sum_{i=1}^{k} n_i} & i=1,2,\cdots,k \\ & j=1,2,\cdots,n_i \\ 0 & i=k+1,\cdots,m \quad j=1,2,\cdots,n_i \end{cases} \tag{2.69}$$

where

$$\gamma_k^E = \begin{cases} \mu \sum_{i=1}^{k} n_i (C_i - C_{k+1}) & ;k=1,2,\cdots,m-1 \\ 0 & ;k=0 \\ \mu \sum_{i=1}^{m} n_i C_i & ;k=m \end{cases} \tag{2.70}$$

Proof. The average channel delay T_{ij} in the channel C_{ij} is given by

$$T_{ij} = \frac{1}{\mu C_{ij} - \lambda_{ij}} \tag{2.71}$$

From Eqs.(2.56) and (2.71), we can easily obtain Eqs.(2.61) and (2.70) in the same manner in which we obtained Eqs.(2.57) and (2.58). Q.E.D.

It is interesting to note the relation between γ_k^o and γ_k^E. Since $C_1>C_2>\cdots>C_k>C_{k+1}$,

$$\gamma_k^o = \mu \sum_{i=1}^{k} n_i(C_i-\sqrt{C_{k+1}}\sqrt{C_i})$$

$$<\mu \sum_{i=1}^{k} n_i(C_i-\sqrt{C_{k+1}}\sqrt{C_{k+1}})=\gamma_k^E \qquad (2.72)$$

Therefore, we find that

$$\gamma_k^o < \gamma_k^E \qquad (2.73)$$

From Eq.(2.73), it is recognized that as the traffic rate increases, the commencement of detour for the optimum route assignment appears always earlier than that for the equal-delay-principle route assignment.

Numerical examples are shown in Figs.2.4 and 2.5 which give a reference to the comment made above. For this example, we assume that $m=3$, $n_i=1$ $(i=1,2,3)$, and $C_1:C_2:C_2=3:2:1$. In Fig.2.4, we show the behavior of the route assignment probability λ_i/γ, where λ_i is the traffic rate which is carried by the i-th channel, ρ is the network utilization given by $\rho=\gamma/\mu(C_1+C_2+C_3)$, and ρ_k^o and ρ_k^E are the network utilization corresponding to the traffic rate γ_k^o and γ_k^E respectively.

For the optimum route assignment, (i) for $\rho<\rho_1^o=0.092$, only the first channel with the greatest capacity is used, (ii) for $\rho_1^o\leq\rho<$

ρ_2^o=0.309, the second channel is also used, but the third channel with the minimum capacity is not used, and (iii) for $\rho \geq \rho_2^o$, all channels are used. Thus, the detouring phenomena are observed at $\rho=\rho_1^o$ and ρ_2^o.

For the equal-delay-principle route assignment, the simillar detouring phenomena are observed at $\rho=\rho_1^E$=0.167 and ρ_2^E=0.5. Furthermore, it is clear that $\rho_1^o<\rho_1^E$ and $\rho_2^o<\rho_2^E$.

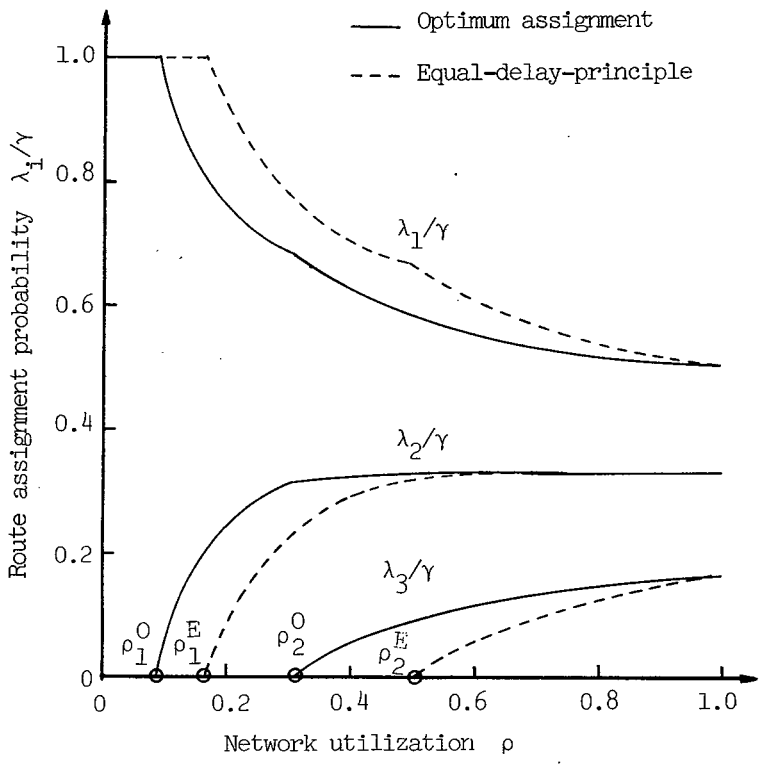

Fig.2.4 Route assignment probability versus network utilization

In Fig.2.5, we show the total average message delay. The curve plotting the total average message delay consists of three curves which satisfy some conditions as follows:

For the optimum route assignment,

(1) $\lambda_1=\gamma$, $\lambda_2=\lambda_3=0$; $\rho<\rho_1^o$

(2) $\partial L_1/\partial\lambda_1=\partial L_2/\partial\lambda_2$, $\lambda_3=0$; $\rho_1^o\leq\rho<\rho_2^o$

(3) $\partial L_1/\partial\lambda_1=\partial L_2/\partial\lambda_2=\partial L_3/\partial\lambda_3$; $\rho\geq\rho_2^o$

For the equal-delay-principle route assignment,

(1) $\lambda_1=\gamma$, $\lambda_2=\lambda_3=0$; $\rho<\rho_1^E$

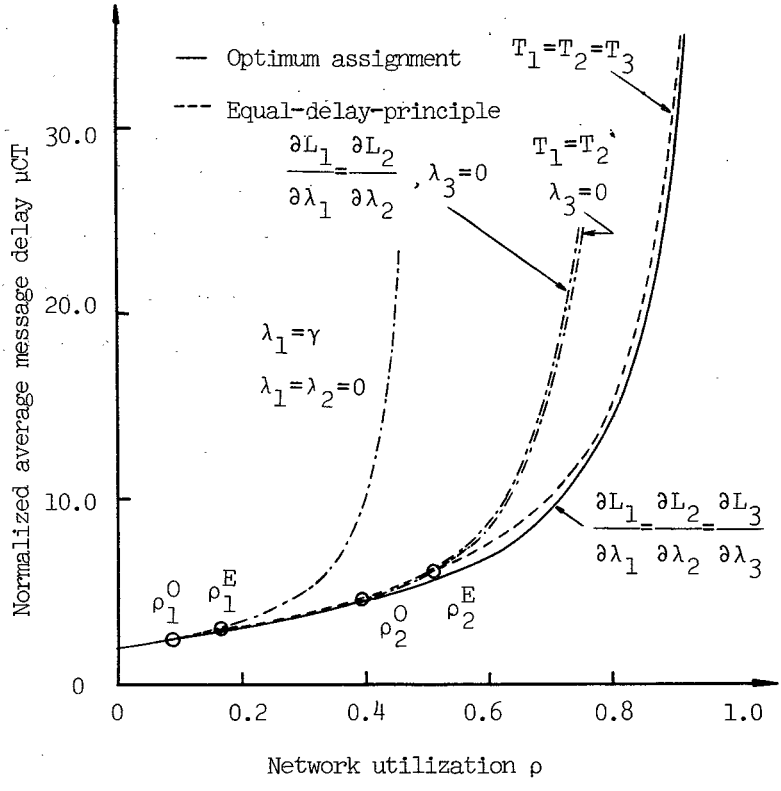

Fig.2.5 Total average message delay versus network utilizatin

(2) $T_1=T_2$, $\lambda_3=0$; $\rho_1^E \leq \rho < \rho_2^E$

(3) $T_1=T_2=T_3$; $\rho \geq \rho_2^E$

2.4 Conclusion

In this chapter, we have discussed the optimum route assignment problem. First, we have mathematically modeled a store-and-forward computer communication network as a simple queueing network in which a queueing unit consists of a buffer as a waiting room and a channel as a server. Next, the optimum route assignment problem has been formulated as a problem to find the optimum set of routes on which messages have to be transmitted in order to minimize the total average message delay. The solution, which is referred to as the optimum route assignment theorem, has been found to that problem. Finally, from the analysis and numerical examples, the detouring phenomenon for the optimum route assignment has been compared to that for the equal-delay-principle route assignment. And, it has been found that the commencement of detour for the optimum route assignment appears earlier than the equal-delay-principle route assignment.

CHAPTER 3

ADAPTIVE MESSAGE ROUTING PROCEDURE

3.1 Introduction

The choice of message routing procedure is an important consideration in the design and the operation of store-and-forward computr communication networks. The message routing procedure is defined as an algorithm by which a switching node selects the output channel on which a message (or packet) is transmitted. So far, a number of message routing procedures have been developed by many authors. And, some ways of classifying routing procedures have been considered [42,43,44].

In this section, we shall classify the message routing procedures into the following two main classes.

(1) Nonadaptive message routing procedure

(2) Adaptive message routing procedure

In the former, a switching node determines the route of a message a priori and in time invariant, so the output channel is fixed for the message according to its source node and its destination node. Generally, the nonadaptive procedure provides the optimum routing for a network in steady state as we discussed in the previous chapter.

In the latter, a switching node determines the route of a message according to network conditions such as load and queue. Therefore, the adaptive routing procedure can adjust to changes in the network conditions, and is very useful for real network operations.

In this chapter, the routing procedures are discussed in detail, a new adaptive routing procedure based on the optimum route assignment theorem is proposed, and its efficiency is verified by simulation results.

3.2 Adaptive Message Routing Procedure

Message routing procedure may be defined as the algorithm by which a switching center determines the output channel on which messages are transmitted. In circuit-switching networks, the routing procedure is one of finding a route from its source node to its destination node which is composed of free channels and maintaining this route for the duration of the call. In store-and-forward network, barring failures of nodes or channels, the communication channels are always available for transmitting messages. However, messages are queued in the buffer. Thus, the routing procedure for a store-and-forward network is one of selecting the next output channel by estimating, generally, message delay or queue length in the buffer.

Requirements for the design of routing procedure are as follows:

(1) It should ensure rapid and reliable delivery of messages. Thus, in general, the performance measure for routing procedure is total average message delay. Moreover, looping or ping-pong phenomenon should be prevented.

(2) It should adapt to changes of network topology due to failures of nodes and channels, or insertion and deletion of nodes.

(3) It should adapt to varying traffic load.

(4) It should route messages or packets away from temporarilly

congested nodes within the network.

(5) It should be a simple algorithm with light load on nodal processors.

Classification schemes of routing procedures have been proposed by many authors till now. Here, the routing procedures are classified into two main classes:

(1) Nonadaptive routing procedure

(2) Adaptive routibg procedure

Let us examine this classification.

(1) Nonadaptive routing procedure

A switching node determines routes of messages (or packets) a priori and in time invariant, i.e. the output channel are fixed or determined by a stochastic algorithm. Thus, in general, this procedure cannot adapt to the variation of the network conditions such as channel load, queue length in the buffer or network topology. There are many procedures in this class.

(a) Fixed routing [16,40]

Fixed routing procedure specifies a unique route followed by a message (or packet) which depends only upon the current node at which the message (or packet) is located in the network, and its destination node. Since the routing is fixed, completely reliable nodes and channels are required, except for the occasional retransmission of a message (or packet) due to channel bit errors.

(b) Flooding or selective flooding[41]

A switching node receiveing or originating a message (or packet) transmits a copy of it over "all" output channels or

over "selective" output channels. The switching node transmits a message (or packet) after the node has checked to see it has not previously transmitted the message (or packet), or that it is not the destination node. In this procedure, we have a large volume of traffic, thus it is inefficient.

(c) Split traffic routing procedure [20]

This procedure allows traffic to flow on more than one route between a given source-destination node pair. This splitting is called traffic bifurcation. As an example, assume that two different routes R_1 and R_2 exist. A message (or packet) is routed on R_1 with probability p, or on R_2 with probability 1-p. Therefore, a better balance of traffic can be maintained throughout the network, and smaller average message delay can be achieved as compared to the fixed routing procedure.

(d) Random routing procedure [16,39]

The selection of the next node for a message (or packet) to be transmitted, is made according to some probability distribution over the set of neighboring nodes which are all of nodes connected to the current node or selective nodes. It is highly efficient, but is relatively uneffected by small changes in the network conditions.

(2) Adaptive routing procedure

A switching node selects the route of a message (or packet) according to network conditions such as channel load,queue length in buffer, or network topology.

(a) Ideal observer routing procedure [20]

Each time a new message (or packet) enters a network, a nodal processor computes its route to minimize the travel

time from its source node to its destination node, based on the complete present information about the network condition. However, it is impossible to have the complete information. Therefore, we cannot use this procedure from the operational view point.

(b) Isolated routing procedure [20]

(c) Distributed routing procedure [20]

The isolated procedure and distributed procedure operate in a simillar manner. Each node has a delay table and a routing table as shown in Fig.3.1. The entries of the delay table, which are denoted by $\hat{T}_j(d,L_N)$, are the estimated delays to

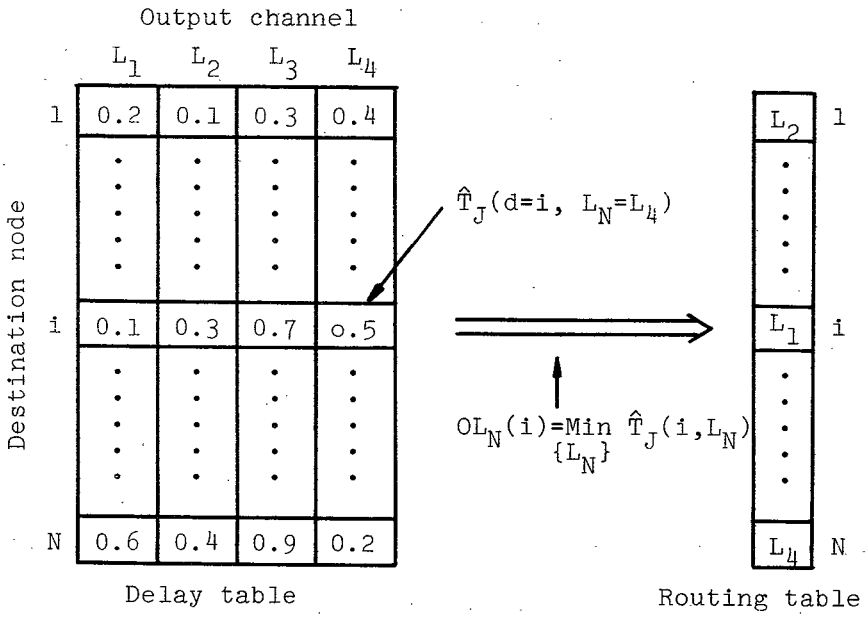

Fig.3.1 Delay table and Routing table

go from the node under consideration (say node J) to some destination node N_d. Then a routing table is formed, for example, by choosing for each row (say the i-th node) that output channel number $OL_N(i)$ whose value in the delay table is minimum as follows:

$$OL_N(i) = \underset{\{L_N\}}{\text{Min.}} \hat{T}_J(i, L_N) \tag{3.1}$$

where $\{L_N\}$ is the set of output channel numbers for node J.

If a node only gains access to the information from the normal packet flowing through it, the procedure is termed "isolated". The examples of this class are as follows:

(i) shortest queue+bias routing[58]

(ii) hot potato routing [59]

(iii) local delay estimated routing [41]

For the distributed routing procedure, the nodes are allowed to exchange routing information by the transmission of special packet. The routing information is exchanged between the nodes periodically or asynchronously. In the ARPA network, the distributed procedure is used.

In general, the nonadaptive procedure is simple, but lacks adaptability, on the other hand, the adaptive procedure is complex, but rich in adaptability. For real operational network, the adaptability seems to be more important requirement than the simplicity, because the nonadaptability to change of the network condition may result in the ruin of communication function of the network. Thus, the adaptive routing procedure is desirable to store-and-forward computer communication networks.

3.3 Adaptive Message Routing Procedure Based on the Optimum Route Assignment Theorem

As we mentioned in the previous section, the conventional adaptive routing procedures use the estimated delay to go from the current node or the source node to the destination node, or the queue length in the buffer, as the routing information. However, the optimum route assignment theorem suggests that by using such routing information, the minimization of the total average message delay cannot be achieved. Furthermore, from that theorem, we find that we can make the total average message delay minimum by using $\partial L/\partial \lambda$ on a channel as routing information. These suggestion can also be confirmed by the following simple example.

In Fig.3.2, we show a network which has two routes, R_1 and R_2,

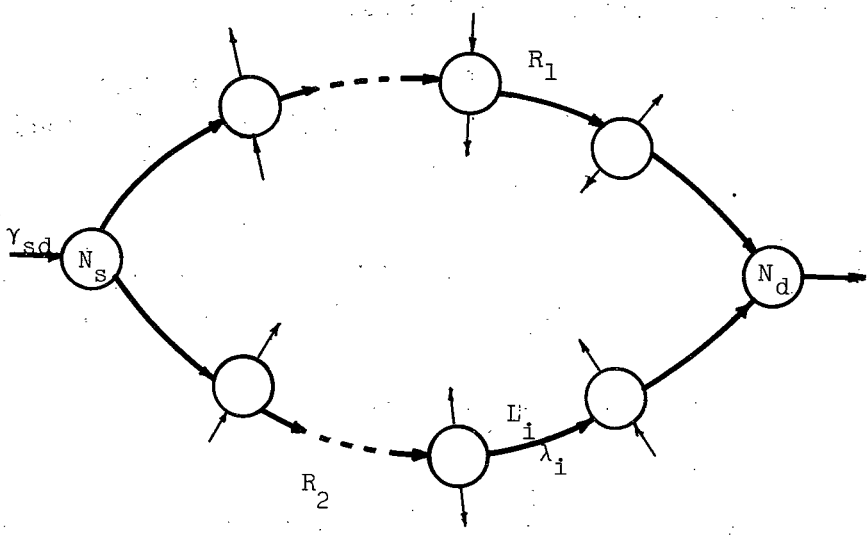

Fig.3.2 Two-route model

from the source node N_s to the destination node N_d. For this network, the total average message delay $T(\gamma)$ can be written by

$$T(\gamma) = \frac{1}{\gamma}\left[\sum_{i|B_i \varepsilon R_1} L_i(\lambda_i) + \sum_{i|B_i \varepsilon R_2} L_i(\lambda_i) + \sum_{i|B_i \notin R_1, R_2} L_i(\lambda_i)\right] \tag{3.2}$$

where $L_i(\lambda_i)$ is the average queue length on the i-th channel, and varies with traffic rate λ_i.

Now, we assume that the traffic amount from N_s to N_d increases by a very small amount $\Delta\gamma_{sd}$, which increases the total avrage message delay. $\Delta\gamma_{sd}$ must be transmitted on either route R_1 or route R_2. For each case, the total average message delay is obtained as follows:

(1) In the case that R_1 is selected,

$$T_1(\gamma+\Delta\gamma_{sd}) = \frac{1}{\gamma+\Delta\gamma_{sd}}\left[\sum_{i|B_i \varepsilon R_1}\left\{L_i(\lambda_i) + \frac{\partial L_i(\lambda_i)}{\partial \lambda_i}\Delta\gamma_{sd}\right\} + \sum_{i|B_i \varepsilon R_2} L_i(\lambda_i) + \sum_{i|B_i \notin R_1, R_2} L_i(\lambda_i)\right] \tag{3.3}$$

(2) In the case that R_2 is selected,

$$T_2(\gamma+\Delta\gamma_{sd}) = \frac{1}{\gamma+\Delta\gamma_{sd}}\left[\sum_{i|B_i \varepsilon R_1} L_i(\lambda_i) + \sum_{i|B_i \varepsilon R_2}\left\{L_i(\lambda_i)\right.\right.$$

$$+\frac{\partial L_i(\lambda_i)}{\partial \lambda_i}\Delta\gamma_{sd}\Big\}+\sum_{i|B_i\notin R_1,R_2} L_i(\lambda_i)\Bigg] \qquad (3.4)$$

Therefore, if

$$\sum_{i|B_i\in R_1}\frac{\partial L_i(\lambda_i)}{\partial \lambda_i} < \sum_{i|B_i\in R_2}\frac{\partial L_i(\lambda_i)}{\partial \lambda_i} \qquad (3.5)$$

then

$$T_1(\gamma+\Delta\gamma_{sd}) < T_2(\gamma+\Delta\gamma_{sd}) \qquad (3.6)$$

On the other hand, it is recognized that the minimization of the total average message delay can be achieved by selecting the route with the minimum sum of the estimated values of $\partial L/\partial\lambda$'s for all channels on that route.

Based on the above discussion, we propose a new adaptive routing procedure, which is belonging to the distributed procedure. In the new adaptive routing procedure, the output channel is computed based on an estimated value of $\partial L/\partial\lambda$. The configuration and operation of this procedure are as follows:

Let consider node I as shown in Fig.3.3.

(1) $\hat{L}(I,J,K)$

This is the sum of the estimated value of $\partial L/\partial\lambda$[†] on the channel from the neighbour node N(I,J)[††] to some node K.

[†] The estimated value of $\partial L/\partial\lambda$ is computed based on the queue length or load of channel.

[††] N(I,J) is the next node for which a message is destined using the output channel J.

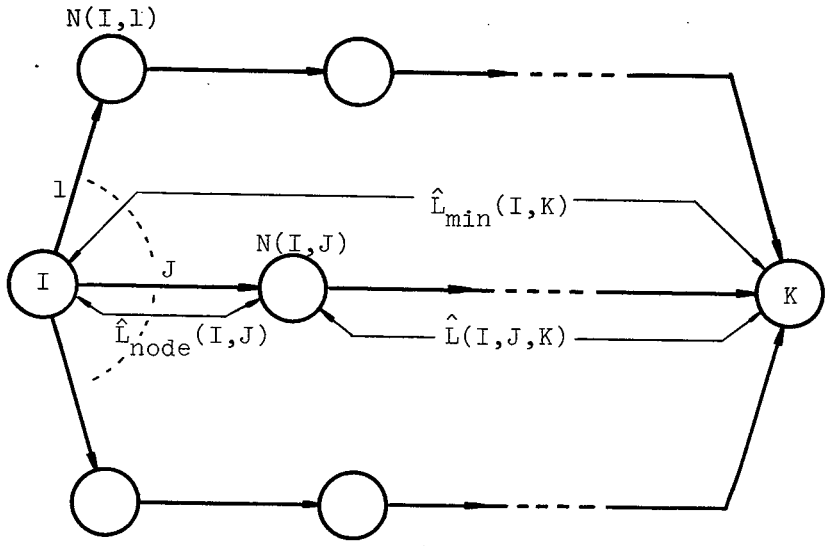

Fig.3.3 Routing information and network

(2) $\hat{L}_{node}(I,J)$

This is the estimated value of $\partial L/\partial \lambda$ on the output channel J.

(3) $\hat{L}_{min}(I,K)$

This is given by

$$\hat{L}_{min}(I,K) = \underset{\{J\}}{\text{Min.}}[\hat{L}(I,J,K) + \hat{L}_{node}(I,J)] \qquad (3.7)$$

(4) Decision of the output channel

Each time a message with destination node K arrives at node I, the nodal processor of node I selects the output channel on which $\hat{L}_{min}(I,K)$ is obtained.

(5) Updating of table $\hat{L}(I,J,K)$

$\hat{L}_{min}(I,K)$ is used as routing information, which is transmitted from each node to its neighbouring nodes

periodically or asynchronouslly. Each time $\hat{L}_{min}(N(I,J),K)$ arrives at node I, the nodal processor of node I updates the table $\hat{L}(I,J,K)$ as follows:

$$\hat{L}(I,J,K) = \hat{L}_{min}(N(I,J),K) \tag{3.8}$$

3.4 Simulation Results and Considerations

In this section, the new adaptive routing procedure proposed in Sec.3.3, is compared with the distributed procedure being used in the ARPA network in which an estimated delay is used as routing information[58]. We consider the ladder network with six nodes and fourteen channels as shown in Fig.3.4.

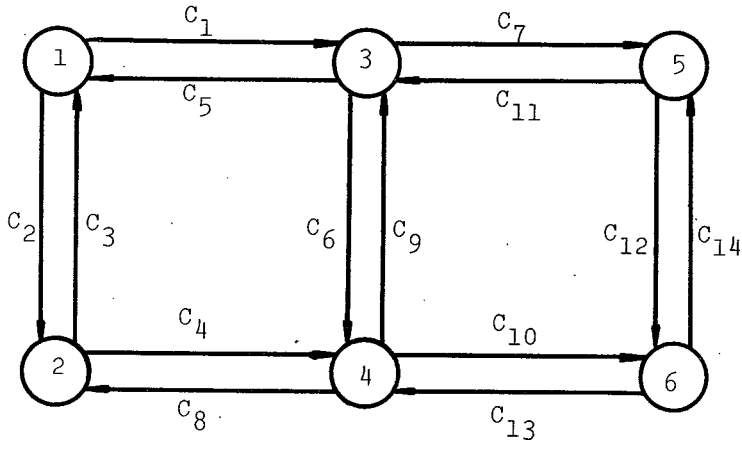

Fig.3.4 Network model for simulation

Furthermore, we assume that message length is exponential with mean 1 kbits, traffic rate is given by

$$[\gamma_{s,d}]=\xi\cdot\begin{bmatrix} 0 & 1.2 & 0.4 & 0.8 & 0.1 & 0.3 \\ 1.2 & 0 & 1.2 & 9.6 & 0.3 & 3.6 \\ 0.4 & 1.2 & 0 & 3.2 & 0.4 & 1.2 \\ 0.8 & 9.6 & 3.2 & 0 & 0.8 & 9.6 \\ 0.1 & 0.3 & 0.4 & 0.8 & 0 & 1.2 \\ 0.3 & 3.6 & 1.2 & 9.6 & 1.2 & 0 \end{bmatrix} \quad \text{(messages/sec)}$$

and the channel capacities are given as follows:

$$C_1=C_2=C_3=C_5=C_6=C_7=C_9=C_{11}=C_{12}=C_{14}=10 \text{ kbits/sec}$$

$$C_4=C_8=C_{10}=C_{13}=30 \text{ kbits/sec}$$

In this model, each channel is modeled as M/M/1, thus $\partial L/\partial\lambda = 1/\mu C(1-\rho)^2$. The estimated values of $\partial L/\partial\lambda$ and channel delay T are computed by observing the average queue length during a routing table updating period T_{ud}[†].

Simulation results[††] are shown in Fig.3.5 and TABLE 3.1. Figure 3.5 shows the comparison of routing procedure performance as a function of traffic rate ξ. It is recognized that the new procedure is superior in total average message delay to the ARPA procedure. Especially, at moderate traffic rate ξ=0.1, 0.3, and 0.5, the saving of total average message delay due to the new procedure is approximately 10 % in magnitude. However, as ξ increases beyond

---

[†] In this simulation, the periodical updating is used,i.e. T_{ud}=const.

[††] The simulation results are obtained by using R-SSQ (Revised System Simulator for Queueing Network) [60].

0.9, the difference of performance may not be observed.

In TABLE 3.1, the average, the variance and the muximum value of the number of channels which a message has passed. For moderate traffic rate, the average and the variance for the new procedure are larger than that for the ARPA procedure. Thus, we may interpret that the new procedure routes messages away from temporarily congested area, which results in smaller total average message delay than the ARPA procedure. However, at about ξ=0.9, that detouring effect is not observed. Now, we note that for T_{ud}=0.1, 1.0 and 2.0 sec, simulation results give simillar performance.

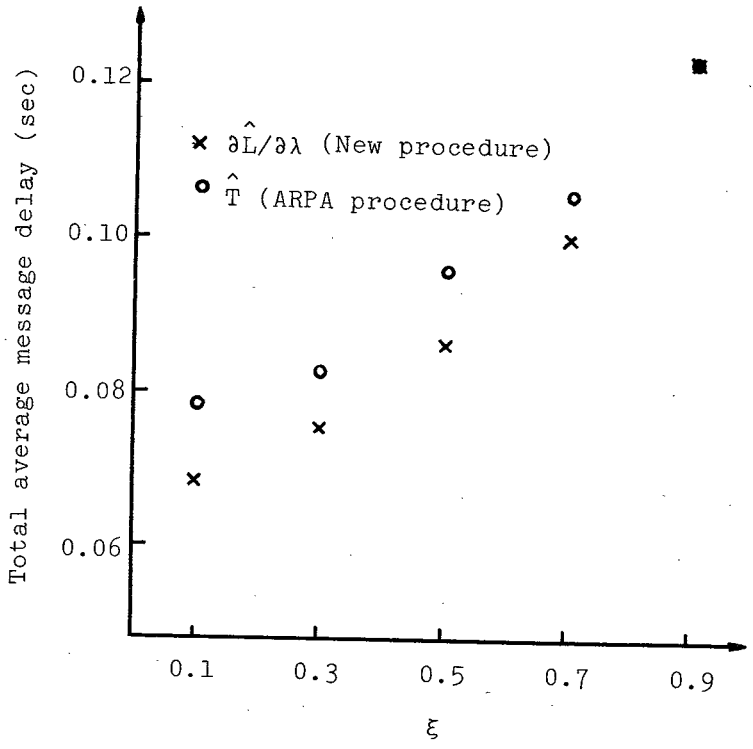

Fig.3.5 Simulated total average message delay

TABLE 3.1 Number of channels which a message has passed

	new procedure			ARPA procedure		
ξ	average	variance	max.	average	variance	max.
0.1	1.308	0.375	5	1.263	0.233	3
0.3	1.303	0.348	6	1.272	0.255	5
0.5	1.325	0.459	21	1.301	0.449	49
0.7	1.359	2.659	273	1.371	0.848	97
0.9	1.368	0.939	110	1.428	0.792	27

3.5 Conclusion

In this chapter, message routing procedures for store-and-forward computer communication networks have been studied.

First, adaptive routing procedures have discussed in detail. Message routing procedures may be classified into two main classes. One of them is nonadaptive routing procedure in which a switching node determines routes of messages (or packets) a priori and in time invariant. The other is adaptive routing procedure in which a switching node determines routes of messages (or packets) according to network conditions. In real network, network conditions vary in time, that is, channel load or number of messages in queue in buffer are not constant in time, and network topology is not fixed due to failures of nodes or channels. Therefore, the adaptive routing procedure is useful for real network.

Second, a new adaptive routing procedure based on the optimum

route assignment theorem has been proposed. Usually, the adaptive routing procedure such as the ARPA procedure, uses estimated message delay or queue length in buffer as routing informations. However, the optimum route assignment theorem suggests that (1) the ARPA procedure cannot make total average message delay minimum, and (2) $\partial L/\partial \lambda$, where L is average queue length, and λ is channel traffic rate, should be used as routing information in order to minimize the total average message delay. In the new procedure, $\partial L/\partial \lambda$ is used as routing information. Its efficiency have been verified by simulation results.

CHAPTER 4

OPTIMUM CHANNEL CAPACITY ASSIGNMENT PROBLEM

4.1 Introduction

Optimum channel capacity assignment problem as first given by Kleinrock is to minimize total average message delay by appropriate assignment of channel capacity. The square root channel capacity assignment given by Kleinrock is the solution to that problem in the case of exponential message length.

In this chapter, the solution to the optimum channel capacity assignment problem is derived, which is referred to as "optimum channel capacity assignment theorem". This theorem gives the necessary and sufficient conditions to minimize the total average message delay in the case of general message length. Furthermore, from numerical results, it is shown that there exists the apparent difference between the behavior of the optimum assignment and that of the most plausible assignment, i.e.the proportional assignment.

4.2 Optimum Channel Capacity Assignment Problem

Optimum channel capacity assignment problem is formulated as follows:

Optimum Channel Capacity Assignment Problem

Given;	network topology
	traffic rate λ_i †
Minimize;	total average message delay T

† If traffic rate γ_{sd} with source node N_s and destination node N_d, and routing of γ_{sd} are fixed, traffic λ_i on channel i is given.

With respect to; $[C_i]$

Under constraint; channel utilization $\rho_i<1$

channel capacity $\sum_i C_i=C$

where C is a fixed total capacity.

4.3 Optimum Channel Capacity Assignment Theorem for General Message Length

The solution to the optimum channel capacity assignment problem is given by the following optimum channel capacity assignment theorem.

Optimum Channel Capacity Assignment Theorem

The solution $C=[C_i]$ to the optimum channel capacity assignment problem is optimum if and only if

$$\frac{\partial L_i}{\partial C_i} = \text{const.} \qquad \text{for all } i \tag{4.1}$$

Proof. The optimum channel capacity assignment problem can be reformulated as follows:

Objective function; $$S(x)=-T=-\sum_i \lambda_i T_i/\gamma=-\sum_i L_i/\gamma \to \max. \tag{4.2}$$

With respect to; $$x=[x_1,x_2,\cdots,x_{N-1}] \tag{4.3}$$

Under constraint; $$K(x)=[g_1(x),g_2(x),\cdots,g_{2N}(x)]\leq 0 \tag{4.4}$$

where x_i is the assignment probability of capacity to channel i,

$$C_i=x_i C \qquad i=1,2,\cdots,N \tag{4.5}$$

$$x_1+x_2+\cdots+x_N=1 \tag{4.6}$$

$$g_i(x)=\frac{\lambda_i}{\mu_i C_i}-1 \qquad i=1,2,\cdots,N \tag{4.7}$$

$$g_{N+i}(x)=-x_i \qquad i=1,2,\cdots,N-1 \tag{4.8}$$

$$g_{2N}(x)=x_1+x_2+\cdots+x_{N-1}-1 \tag{4.9}$$

For the above problem, we use Lagrange function

$$\phi(x,\eta)=S(x)-\eta\cdot K(x) \tag{4.10}$$

where η is Lagrange multiplier given as follows:

$$\eta=(\alpha_1,\alpha_2,\cdots,\alpha_N,\beta_1,\beta_2,\cdots,\beta_{N-1},\delta) \tag{4.11}$$

From Kuhn-Tucker theorem [56], the necessary and sufficient conditions for (x^o,η^o) to maximize S under the constraint Eq.(4.4) are as follows:

(i) Necessary conditions

$$\nabla_x\phi|_{x^o,\eta^o} \le 0, \quad (\nabla_x\phi,x)_{x^o,\eta^o}=0, \quad x^o\ge 0 \tag{4.12}$$

$$\nabla_\eta\phi|_{x^o,\eta^o} \ge 0, \quad (\nabla_\eta\phi,\eta)_{x^o,\eta^o}=0, \quad \eta^o\ge 0 \tag{4.13}$$

(ii) Sufficient conditions

$$\phi(x,\eta^o)\le\phi(x^o,\eta^o)+(\nabla_x\phi|_{x^o,\eta^o},\ x-x^o) \tag{4.14}$$

$$\phi(x^o,\eta)\ge\phi(x^o,\eta^o)+(\nabla_\eta\phi|_{x^o,\eta^o},\ \eta-\eta^o) \tag{4.15}$$

Later on, x^o and η^o will be omitted.

First, we consider the necessary conditions, i.e. Eqs.(4.12) and (4.13)

Since T_i given by Eqs.(2.6) and (2.7) is differentiable with respect to C_i, and the relation between C_i and x_i is given by Eq. (4.5), then T_i is differentiable function with respect to x_i.

Futhermore, the relation between T_i and L_i is given by Eq.(2.4). Therefore, partial derivative of L_i by x_j is given by

$$\frac{\partial L_i}{\partial x_j} = \begin{cases} C\dfrac{\partial L_i}{\partial C_i} & i=1,2,\cdots,N-1, \quad j=i \\ -C\dfrac{\partial L_N}{\partial C_N} & i=N, \quad j=1,2,\cdots,N-1 \\ 0 & \text{elsewhere} \end{cases} \tag{4.16}$$

From Eqs.(4.5) and (4.7),

$$\frac{\partial g_i}{\partial x_j} = \begin{cases} -C\dfrac{\lambda_i}{\mu_i C_i^2} & i=1,2,\cdots,N-1, \quad j=i \\ C\dfrac{\lambda_N}{\mu_N C_N^2} & i=N, \quad j=1,2,\cdots,N-1 \\ 0 & \text{elsewhere} \end{cases} \tag{4.17}$$

from Eq.(4.8),

$$\frac{\partial g_{N+i}}{\partial x_j} = \begin{cases} -1 & i=1,2,\cdots,N-1, \quad j=i \\ 0 & i=1,2,\cdots,N-1, \quad j\neq i \end{cases} \tag{4.18}$$

and, from Eq.(4.9),

$$\frac{\partial g_{2N}}{\partial x_j} = 1 \qquad j=1,2,\cdots,N-1 \tag{4.19}$$

Therefore, the first of the necessary conditions, i.e. Eq.(4.12), is as follows:

$$\frac{\partial \phi}{\partial x_j} = -\frac{C}{\gamma}\frac{\partial L_j}{\partial C_j} + \frac{C}{\gamma}\frac{\partial L_N}{\partial C_N} - \alpha_j C\left(-\frac{\lambda_j}{\mu_j C_j^2} + \frac{\lambda_N}{\mu_N C_N^2}\right) + \beta_j - \delta \leq 0 \qquad (4.20)$$

$$j = 1, 2, \cdots, N-1$$

$$\sum_j x_j \left[-\frac{C}{\gamma}\frac{\partial L_j}{\partial C_j} + \frac{C}{\gamma}\frac{\partial L_N}{\partial C_N} - \alpha_j C\left(-\frac{\lambda_j}{\mu_j C_j^2} + \frac{\lambda_N}{\mu_N C_N^2}\right) + \beta_j - \delta\right] = 0 \qquad (4.21)$$

$$x_j \geq 0 \qquad j = 1, 2, \cdots, N-1 \qquad (4.22)$$

The above three equations must be satisfied simultaneously. Thus, if $x_j > 0$, then

$$-\frac{C}{\gamma}\frac{\partial L_j}{\partial C_j} + \frac{C}{\gamma}\frac{\partial L_N}{\partial C_N} - \alpha_j\left(-\frac{\lambda_j}{\mu_j C_j^2} + \frac{\lambda_N}{\mu_N C_N^2}\right) + \beta_j - \delta = 0 \qquad (4.23)$$

and, if $x_j = 0$, then

$$-\frac{C}{\gamma}\frac{\partial L_j}{\partial C_j} + \frac{C}{\gamma}\frac{\partial L_N}{\partial C_N} - \alpha_j\left(-\frac{\lambda_j}{\mu_j C_j^2} + \frac{\lambda_N}{\mu_N C_N^2}\right) + \beta_j - \delta \leq 0 \qquad (4.24)$$

The partial derivative of the Lagrange function ϕ by the Lagrange multipliers, i.e. α_j, β_j, and δ, are given by as follows:

$$\frac{\partial \phi}{\partial \alpha_j} = -\left(\frac{\lambda_j}{\mu_j C_j} - 1\right) \qquad j = 1, 2, \cdots, N \qquad (4.25)$$

$$\frac{\partial \phi}{\partial \beta_j} = x_j \qquad j = 1, 2, \cdots, N-1 \qquad (4.26)$$

and

$$\frac{\partial \phi}{\partial \delta} = -(x_1+x_2+\cdots+x_{N-1}-1) \tag{4.27}$$

Therefore, the second of the necessary conditions, i.e. Eq.(3.13), means that the optimum solution x must satisfy several equations given by

$$-\left(\frac{\lambda_j}{\mu_j C_j} - 1\right) \geq 0 \qquad j=1,2,\cdots,N \tag{4.28}$$

$$x_j \geq 0 \qquad j=1,2,\cdots,N-1 \tag{4.29}$$

$$-(x_1+x_2+\cdots+x_{N-1}-1) \geq 0 \tag{4.30}$$

$$-\sum_j \alpha_j\left(\frac{\lambda_j}{\mu_j C_j} - 1\right) + \sum_j \beta_j x_j - \delta(x_1+x_2+\cdots+x_{N-1}-1) = 0 \tag{4.31}$$

and

$$\alpha_j \geq 0,\ \beta_j \geq 0,\ \delta \geq 0 \qquad i=1,2,\cdots,N,\ j=1,2,\cdots,N-1 \tag{4.32}$$

From Eqs.(4.28)-(4.32), it is easily recognized that:

(a) If $\lambda_j/\mu_j C_j \to 1$, then $L_j \to \infty$, i.e. $S \to -\infty$. Thus, α_j must be equal to zero.

(b) Since $x_j C = \lambda_j/\mu_j$, $x_j > 0$. Thus, β_j must be equal to zero.

(c) Similarly, $x_N = 1-(x_1+x_2+\cdots+x_{N-1})$ must be larger than zero, thus, $\delta = 0$.

From the above conditions (a), (b), and (c), and Eqs.(4.23) and (4.24), we obtain

$$-\frac{C}{\gamma}\frac{\partial L_j}{\partial C_j} + \frac{C}{\gamma}\frac{\partial L_N}{\partial C_N} = 0 \qquad j=1,2,\cdots,N-1 \tag{4.33}$$

which results in

$$\frac{\partial L_j}{\partial C_j} = \text{const.} \quad \text{for all } j$$

Next, we consider the sufficient conditions, i.e.Eqs.(4.14) and (4.15). These conditions imply that the objective function S must be a continuous, differentiable and concave function of vector x, and the constraint functions must be convex functions. The objective function Eq.(4.2) and the constraint functions Eqs.(4.7), (4.8), and (4.9) clearly satisfy these conditions. Q.E.D.

Assuming that the message length is general, Eq.(4.1) is given by

$$\frac{\rho_i}{C_i}\left[1+(1+\mu_i^2\sigma_i^2)\frac{\rho_i(2-\rho_i)}{2(1-\rho_i)^2}\right]=\text{const. for all } i \quad (4.34)$$

Especially, when the message length is Erlangian with phase k, $\sigma_i^2=1/k\mu_i^2$, then Eq.(4.34) is rewritten by

$$\frac{\rho_i}{C_i}\left[1+(1+\frac{1}{k})\frac{\rho_i(2-\rho_i)}{2(1-\rho_i)^2}\right]=\text{const. for all } i \quad (4.35)$$

For k=1, the message length is exponential, and

$$\frac{\rho_i}{C_i(1-\rho_i)^2} = \text{const.} \quad \text{for all } i \quad (4.36)$$

From Eq.(4.36), the square root channel capacity assignment†

---

† See CHAPTER 5.

given by Kleinrock is easily derived.

For $k=\infty$, the message length is constant, and

$$\frac{\rho_i}{C_i}\left[1+\frac{\rho_i(2-\rho_i)}{2(1-\rho_i)^2}\right]=\text{const.} \quad \text{for all } i \tag{4.37}$$

4.4 Numerical Results and Considerations

We are now ready to make an evaluation of the optimum channel capacity assignment which is derived from the optimum channel capacity assignment theorem described in Eq.(4.1). To do this, we must observe the behavior of total average message delay and channel capacity assignment properties both with the optimum channel capacity assignment and with some other plausible channel capacity assignment. The most plausible and intuitively reasonable assignment is the proportional channel capacity assignment which assigns a fraction of the total capacity C to each channel in direct proportion to the traffic carried by that channel, viz.

$$C_i=\frac{\lambda_i}{\lambda}C \tag{4.38}$$

where

$$\lambda=\sum_i \lambda_i \tag{4.39}$$

In the following numerical results, we assume that the message length is Erlangian with phase k, and $\mu_i=\mu$.

First, we consider the two-channel model as shown in Fig.4.1. In this figure, λ_i is the average traffic rate entering the i-th channel,

$$\lambda_1+\lambda_2=\gamma \tag{4.40}$$

$$\lambda_2=\alpha\lambda \qquad ; \ 0<\alpha<1 \tag{4.41}$$

and the network utilization ρ is given by

$$\rho=\frac{\gamma}{\mu C} \tag{4.42}$$

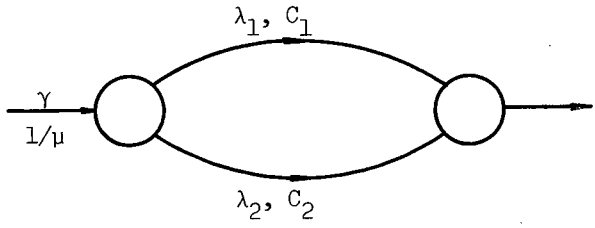

Fig.4.1 Two-channel model

The properties of the channel capacity assignment for both assignments are shown in Fig.4.2. In Fig.4.2, we show the properties of channel capacity assignment probability for the optimum assignment and the proportional assignment, where E_k shows the Erlangian distribution with phase k.

With the optimum assignment, the assignment probability varies according to the variation of the network utilization or the traffic rate γ. On the other hand, it is clear that, with the proportional assignment, the assignment probability is constant independently of the network utilization. The assignment probability x_1 for the first channel, which transmits more traffic than the second channel, with the optimum assignment, is smaller than that with the proportional assignment. The difference between them decreases as the network utilization gets to unity. At the extreme case that $\rho \to 1$, that difference is zero. Furthermore, it is recognized that the difference reduces as α gets large, i.e. the

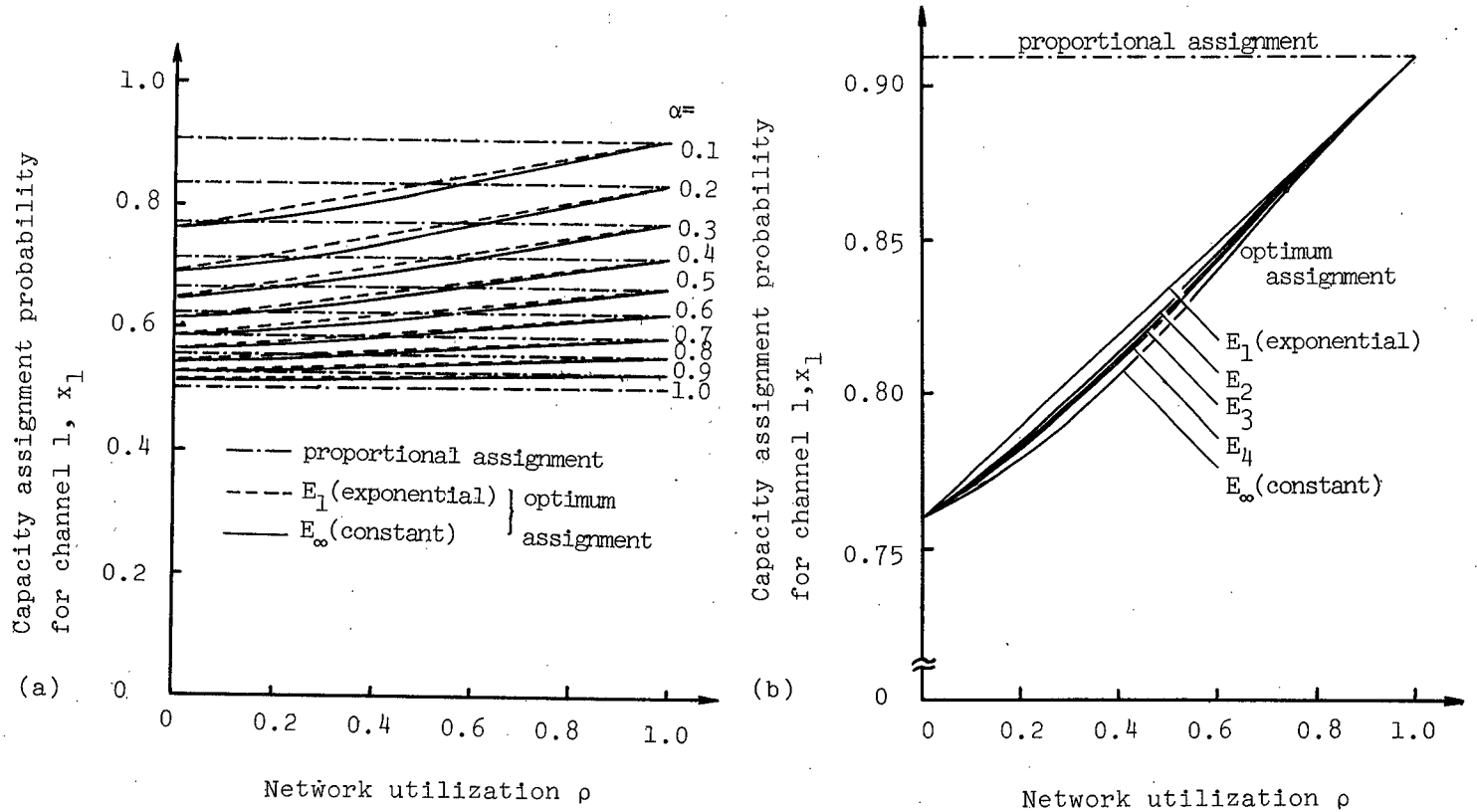

Fig.4.2 Capacity assignment probability for channel 1, x_1 (a) α=0.1,0.2,0.3,0.4,0.5,0.6 0.7,0.8,0.9,1.0;E_1,E_∞ (b) α=0.1;E_1,E_2,E_3,E_4,E_∞

variation between λ_1 and λ_2 gets small, or the phase k of the Erlangian message length gets small, i.e. the variance of message length gets large.

In Fig.4.3, we show the incresing rate of the total average message delay for the proportional assignment to that for the optimum assignment, i.e. $(T_p-T_o)/T_o$, where T_o is the total average message delay for the optimum assignment, and T_p is that for the proportional assignment. The increasing rate gets large as the variance of message length gets large. When the message length is exponential with maximum variance, the increasing rate is maximum, and constant independently of the network utilization ρ. The increasing rate for the Erlangian message length with phase k(>1) is equal to that for the exponential message length at the extreme case that $\rho \to 0$. As ρ increases, it reduces slowly, and it has the minimum value at some network utilization ρ_{min} (in this example, at about $\rho=0.75$). Moreover, it increases rather fast with increasing network utilization beyond the value ρ_{min}. At the extreme case that $\rho \to 1$, it approaches the value of the increasing rate for exponential message length.

Next, we consider a ladder network with six nodes and fourteen channels as shown in Fig.4.4. For the numerical computation, we give relative traffic matrix as shown in Fig.4.5, and give a fixed routing of messages as shown in TABLE 4.1. As the result, the relationship among channel traffics λ_i(i=1,2,$\cdots$,14) is obtained as follows:

$$\lambda_1=\lambda_3=\lambda_7=\lambda_8=\lambda_{12}=\lambda_{13}$$
$$\lambda_2=\lambda_4=\lambda_{10}=\lambda_{11}=\lambda_{14}$$
$$\lambda_6=\lambda_9$$

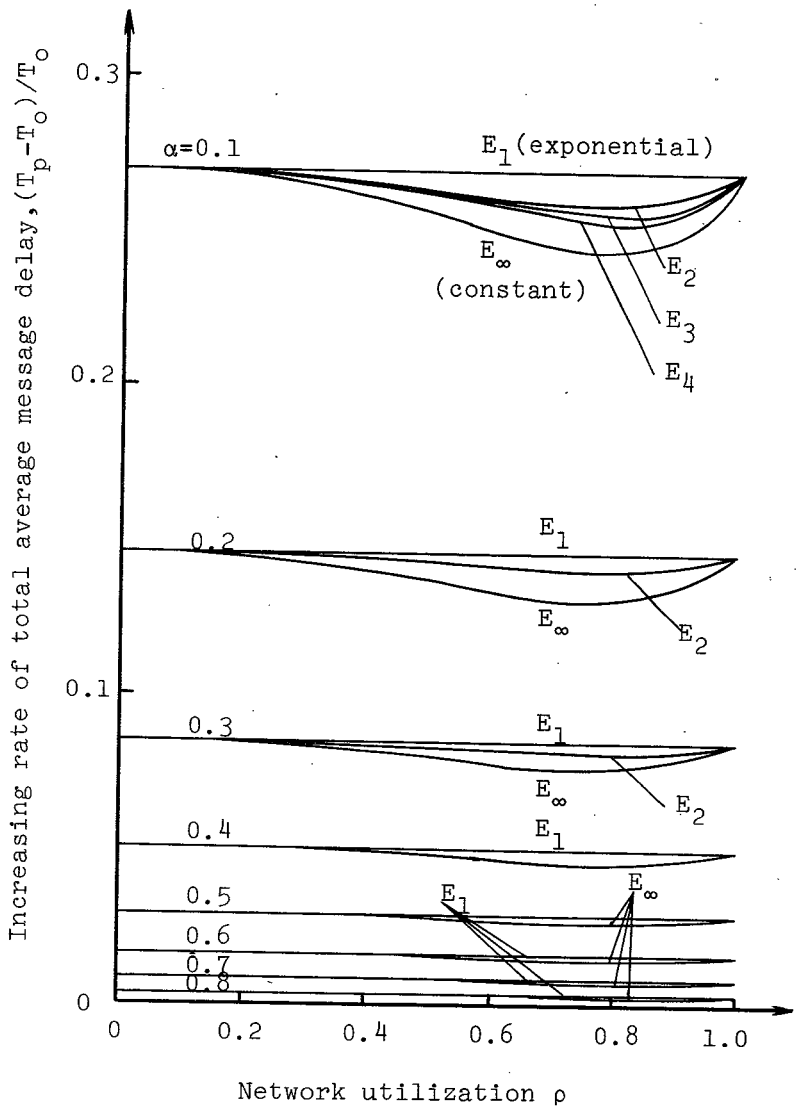

Fig.4.3 Increasing rate of total average message delay for proportional assignment to that for optimum assignment, $(T_p-T_o)/T_o$

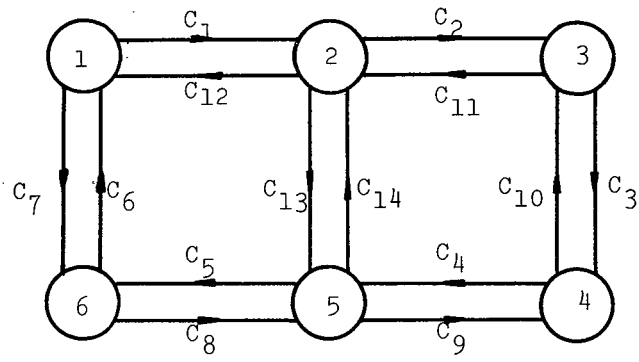

Fig.4.4 Ladder network with six nodes and fourteen channels

Destination node

Source node	1	2	3	4	5	6
1	0	1.0	1.0	1.0	1.0	1.0
2	1.0	0	1.0	1.0	1.0	1.0
3	1.0	1.0	0	1.0	1.0	1.0
4	1.0	1.0	1.0	0	1.0	1.0
5	1.0	1.0	1.0	1.0	0	1.0
6	1.0	1.0	1.0	1.0	1.0	0

Fig.4.5 Relative traffic matrix

$$\lambda_1:\lambda_2:\lambda_6=1.0:0.6:0.2$$

In Fig.4.6, we show the increasing rate of the total average message delay for the proportional channel capacity assignment to that for the optimum channel capacity assignment. It is easily recognized that the behavior of the increasing rate is simillar to that for the two-channel model as shown in Fig.4.3.

TABLE 4.1 ROUTING OF MESSAGES

Source Node \ Destination Node	1	2	3	4	5	6
1	*	c_1	c_1, c_2	c_1, c_2, c_3	c_7, c_8	c_7
2	c_{12}	*	c_2	c_2, c_3	c_{13}	c_{12}, c_7
3	c_{11}, c_{12}	c_{11}	*	c_3	c_3, c_4	c_3, c_4, c_5
4	c_4, c_5, c_6	c_{10}, c_{11}	c_{10}	*	c_4	c_4, c_5
5	c_5, c_6	c_{14}	c_9, c_{10}	c_9	*	c_5
6	c_6	c_6, c_1	c_6, c_1, c_2	c_8, c_9	c_8	*

Therefore, it is found that the numerical results as shown in Figs.4.3 and 4.6 show the basic relation between the total average message delay for the optimum channel capacity assignment and that for the proportional channel capacity assignment.

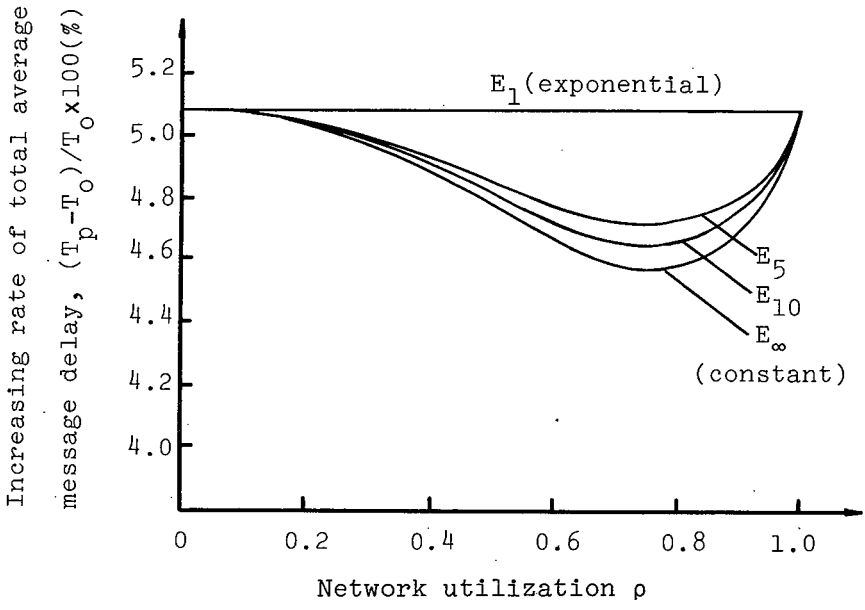

Fig.4.6 Increasing rate of total average message delay for proportional assignment to that for optimum assignment, $(T_p-T_o)/T_o$

4.5 Conclusion

We have discussed the optimum channel capacity assignment problem. The solution to that problem in the case of general message length has been found, which is referred to as the optimum channel capacity assignment theorem. The optimum channel capacity assignment theorem gives the necessary and sufficient conditions to assign

capacity to channels in order to minimize total average message delay under a fixed total capacity constraint.

Moreover, we have compared the optimum channel capacity assignment to the most plausible and reasonable channel capacity assignment, i.e. the proportional channel capacity assignment. From the numerical results, it has been found that (i) the increasing rate of total average message delay for the proportional assignment to that for the optimum assignment increases as the variation among channel traffic rates or the variance of message length gets large, (ii) the increasing rate for exponential message length is maximum, (iii) the increasing rate for Erlangian message length is equal to that for exponential message lengh in the case that the network utilization approaches zero; it decreases slowly as the network utilization increases, it increases rather fast with increasing network utilization beyond certain value of network utilization with the minimum increasing rate, and approaches the value of the increasing rate for exponential message length at the extreme case that the network utilization gets to unity.

CHAPTER 5

EXTENDED OPTIMUM CHANNEL CAPACITY ASSIGNMENT PROBLEMS

5.1 Introduction

The optimum channel capacity assignment problem given by Kleinrock [16], is to find the optimum channel capacity assignment in order to minimize total average message delay under a fixed total capacity constraint. In this chapter, we consider extended optimum channel capacity assignment problems based on Kleinrock's problem. The extended problem formulated by Meister et al.[30] is to find channel capacity assignment in order to reduce variation among channel delays. On the other hand, Kleinrock's problem may be interpreted as a problem to find the channel capacity assignment in order to minimize the total number of messages in the network, i.e. the sum of queueing messages on all channels. Thus, we may formulate another extended optimum channel capacity assignment problem to find the channel capacity assignment in order to reduce variation among the numbers of queueing messages on channels. This formulation is particularly important when the buffer size is a critical factor such as in facsimile network. It will be shown that the assignment for our extended problem has a dual relationship to the assignment for the extended problem by Meister et al.

5.2 Extended Optimum Channel Capacity Assignment Problem for Delay Variation

Meister et al. observed that in minimizing total average message delay T in Kleinrock's optimum channel capacity assignment problem, wide variation was possible among channel delays T_i. As a

result, they formulated an extended optimum channel capacity assignment problem as follows:

Extended Optimum Channel Capacity Assignment Problem [A]

Objective function;
$$\left[\sum_i \frac{\lambda_i}{\gamma} T_i^{\alpha}\right]^{1/\alpha} \rightarrow \min. \tag{5.1}$$

Constraints;
$$\rho_i < 1 \text{ and } \sum_i C_i = C \tag{5.2}$$

where C is a fixed total capacity.

Concerning Problem [A], it is easily recognized that for $\alpha>1$, the variation among channel delays is forced to decrease, and for $0<\alpha<1$, it is forced to increase.† Of cource, when $\alpha=1$, the above problem is Kleinrock's problem whose solution is referred to as the square root channel capacity assignment.

The solution to Problem [A] with a given value α, which we denote by $C_i^{*(\alpha)}$, may be written as

$$C_i^{*(\alpha)} = \frac{\lambda_i}{\mu_i} + C_a \frac{\lambda_i^{1/(1+\alpha)} / \mu_i^{\alpha/(1+\alpha)}}{\sum_j \lambda_j^{1/(1+\alpha)} / \mu_j^{\alpha/(1+\alpha)}} \tag{5.3}$$

where

$$C_a = C - \sum_j \lambda_j/\mu_j \tag{5.4}$$

λ_j/μ_j represents the average traffic rate at which bits will enter the j-th channel, that is, the minimum capacity which must be assigned to the j-th channel. Therefore, the expression for C_a

† In the paper by Meister et al., they didn't discuss the case that α is real. However, Problem [A] and its solution hold good for any positive real value α.

represents the excess capacity which is merely the difference between the fixed total capacity and the sum of minimum capacities assigned to each channel.

Furthermore, using this channel capacity assignment, Eq.(5.3), the channel delay and the total average message delay, which we denote by $T_i^{*(\alpha)}$ and $T^{*(\alpha)}$ respectively, may be obtained as follows:

$$T_i^{*(\alpha)} = \frac{1}{\mu_i C_a} \frac{\sum_j \lambda_j^{1/(1+\alpha)} / \mu_j^{\alpha/(1+\alpha)}}{\lambda_i^{1/(1+\alpha)} / \mu_i^{\alpha/(1+\alpha)}} \tag{5.5}$$

$$T^{*(\alpha)} = \frac{1}{\gamma C_a} \sum_i \frac{\lambda_i^{\alpha/(1+\alpha)}}{\mu_i^{1/(1+\alpha)}} \sum_j \frac{\lambda_j^{1/(1+\alpha)}}{\mu_j^{\alpha/(1+\alpha)}} \tag{5.6}$$

Next, we consider Problem [A] more carefully.

At the case that $\alpha=1$, we can easily obtain the square root channel capacity assignment which is given by

$$C_i^{*(1)} = \frac{\lambda_i}{\mu_i} + C_a \frac{\sqrt{\lambda_i/\mu_i}}{\sum_j \sqrt{\lambda_j/\mu_j}} \tag{5.7}$$

As we mentioned earlier, as α gets large, the variation among channel delays reduces. Especially, when $\alpha \to \infty$, we have

$$C_i^{*(\infty)} = \frac{\lambda_i}{\mu_i} + C_a \frac{1/\mu_i}{\sum_j 1/\mu_j} \tag{5.8}$$

In this extreme case, the channel delay is given by

$$T_i^{*(\infty)} = \frac{1}{C_a} \sum_j \frac{1}{\mu_j} \tag{5.9}$$

From Eq.(5.9), it is clear that each T_i is the same, that is, there is no variation among them. And, Equation (5.6) reduces to

$$T^{*(\infty)} = \frac{1}{\gamma C_a} \sum_i \lambda_i \sum_j \frac{1}{\mu_j} \tag{5.10}$$

At the other extreme case that $\alpha \to 0$, we have

$$C_i^{*(0)} = \frac{\lambda_i}{\mu_i} + C_a \frac{\lambda_i}{\Sigma \lambda_j} \tag{5.11}$$

$$T_i^{*(0)} = \frac{1}{\mu_i C_a} \frac{\Sigma \lambda_j}{\lambda_i} \tag{5.12}$$

and

$$T^{*(0)} = \frac{1}{\gamma C_a} \sum_i \frac{1}{\mu_i} \sum_j \lambda_j \tag{5.13}$$

In this case, it is recognized that the excess capacity is assigned to each channel in direct proportion to the message arrival rate (messages/sec) at that channel.

Next, we assume that $\mu_i = \mu$ for all i. This assumption is reasonable for message-switching networks such as computer communication networks. By making the above assumption, we reduce Eqs.(5.3)-(5.6) and (5.10)-(5.13) to the following equations;

For any positive real value α, we have

$$C_i^{*(\alpha)} = \frac{\lambda_i}{\mu} + C_a \frac{\lambda_i^{1/(1+\alpha)}}{\sum_j \lambda_j^{1/(1+\alpha)}} \tag{5.14}$$

$$C_a = C - \sum_j \frac{\lambda_j}{\mu} \tag{5.15}$$

$$T_i^{*(\alpha)} = \frac{1}{\mu C_a} \frac{\sum_j \lambda_j^{1/(1+\alpha)}}{\lambda_i^{1/(1+\alpha)}} \tag{5.16}$$

and

$$T^{*(\alpha)} = \frac{1}{\mu\gamma C_a} \sum_i \lambda_i^{\alpha/(1+\alpha)} \sum_j \lambda_j^{1/(1+\alpha)} \tag{5.17}$$

In the case that $\alpha \to \infty$, we have

$$C_i^{*(\infty)} = \frac{\lambda_i}{\mu} + \frac{C_a}{N} \tag{5.18}$$

$$T_i^{*(\infty)} = \frac{N}{\mu C_a} \tag{5.19}$$

and

$$T^{*(\infty)} = \frac{N}{\mu\gamma C_a} \sum_i \lambda_i \tag{5.20}$$

In the case that $\alpha \to 0$, we have

$$C_i^{*(0)} = \frac{\lambda_i}{\mu} + C_a \frac{\lambda_i}{\sum_j \lambda_j} = C \frac{\lambda_i}{\sum_j \lambda_j} \tag{5.21}$$

$$T_i^*(0) = \frac{1}{\mu C_a} \frac{\sum_j \lambda_j}{\lambda_i} \tag{5.22}$$

and

$$T^*(0) = \frac{N}{\mu\gamma C_a} \sum_j \lambda_j \tag{5.23}$$

This assignment gives capacity to a channel in direct proportion to the traffic amount (λ_i/μ bits/sec) carried by that channel. Thus, it is called by the proportional channel capacity assignment.

5.3 Extended Optimum Channel Capacity Assignment Problem for Queue Variation

The well known Little's formula [55] says that average queue length L_i is merely a product of message arrival rate λ_i and channel delay T_i. Therefore, it is easily recognized that the total average message delay is directly proportional to the sum of average queue lengths in all channels, i.e. the total number of messages within the network. Thus, Kleinrock's problem may be interpreted as an optimum channel capacity assignment problem whose solution is the channel capacity assignment minimizing the total number of messages within the network. Here, we consider the variation among the queue lengths L_i, and formulate another extended optimum channel capacity assignment problem different from Problem [A] as follows:

Extended Optimum Channel Capacity Assignment Problem [B]

Objective function;

$$\left[\sum_i L_i^{\alpha} \right]^{1/\alpha} \rightarrow \min. \tag{5.24}$$

Constraint ; $\rho_i<1$ and $\sum_i C_i=C$ (5.25)

The optimum channel capacity assignment, which we denote by $C_i^{(\alpha)}$, is given by

$$C_i^{(\alpha)}=\frac{\lambda_i}{\mu_i}+C_a\frac{(\lambda_i/\mu_i)^{\alpha/(1+\alpha)}}{\sum_j(\lambda_j/\mu_j)^{\alpha/(1+\alpha)}} \tag{5.26}$$

and the average channel delay and the total average message delay, which we denote by $T_i^{(\alpha)}$ and $T^{(\alpha)}$ respectively, are given by

$$T_i^{(\alpha)}=\frac{1}{\mu_i C_a}\frac{\sum_j(\lambda_j/\mu_j)^{\alpha/(1+\alpha)}}{(\lambda_i/\mu_i)^{\alpha/(1+\alpha)}} \tag{5.27}$$

$$T^{(\alpha)}=\frac{1}{\gamma C_a}\sum_i\left(\frac{\lambda_i}{\mu_i}\right)^{1/(1+\alpha)}\sum_j\left(\frac{\lambda_j}{\mu_j}\right)^{\alpha/(1+\alpha)} \tag{5.28}$$

Proof. Problem [B] can be reformulated as follows:

Objective function; $f(C')=\sum_i L_i^{\alpha}\rightarrow\min.$ (5.29)

Constraint†; $g(C')=\sum_i C_i'-C_a=0$ (5.30)

---

† In Eq.(5.31), we allow C_i' to take zero at which case $\rho_i=1$. Thus the constraint equation Eq.(5.31) is different from the original constraint that $\rho_i<1$ (Eq.(5.25)). However, when $\rho_i\rightarrow 1$, $L_i\rightarrow\infty$, which shows that the objective function given by Eq.(5.29) can never be minimum for $C_i'=0$. Thus, it is reasonable that Eq.(5.25) is replaced with Eqs.(5.30) and (5.31).

$$C' \geq 0 \tag{5.31}$$

where $C'=(C_1', C_2', \cdots, C_N')$ and

$$C_i' = C_i^{(\alpha)} - \lambda_i/\mu_i \tag{5.32}$$

Concerning the reformulated problem, the objective function $f(C')$ is a continuous, differentiable and convex function with respect to vector C'. And, both the constraint functions Eqs.(5.30) and (5.31) are linear functions with respect to C', that is, special forms of convex functions. For the proof of the above problem, we use Lagrange function given by

$$\phi(C', u) = f(C') + u \cdot g(C') \tag{5.33}$$

where u is a Lagrange multiplier. From Kuhn-Tucker theorem [56], the necessary and sufficient conditions for $(\hat{C}', \hat{u})$ to minimize $f(C')$ under the constraint given by Eqs.(5.30) and (5.31) are as follows:

(i) necessary conditions

$$\nabla_{C'} \phi(\hat{C}', \hat{u}) \geq 0 \tag{5.34}$$

$$\hat{C}' \cdot \nabla_{C'} \phi(\hat{C}', \hat{u}) = 0 \tag{5.35}$$

$$\nabla_u \phi(\hat{C}', \hat{u}) = 0 \tag{5.36}$$

(ii) sufficient condition

$\phi(\hat{C}', \hat{u})$ is a convex function with respect to both C' and u.

Since $f(C')$ and $g(C')$ are convex, $\phi(C', u)$ satisfies the sufficient condition.

Next, we consider the necessary conditions. On the assumption that the message length is exponential, the average queue length L_i is given by

$$L_i = \frac{\lambda_i}{\mu_i C_i'} \tag{5.37}$$

Therefore, partially differentiating Eq.(5.33) with respect to vector C', and evaluating at $(\hat{C}', \hat{u})$, we may rewrite Eqs.(5.34) and (5.35) as follows:

$$\left[-\alpha\left(\frac{\lambda_i}{\mu_i \hat{C}_i'}\right)^{\alpha-1} \frac{\lambda_i}{\mu_i \hat{C}_i'^2} + \hat{u}\right] \geq 0 \tag{5.38}$$

$$\sum_i \hat{C}_i' \left[-\alpha\left(\frac{\lambda_i}{\mu_i \hat{C}_i'}\right)^{\alpha-1} \frac{\lambda_i}{\mu_i \hat{C}_i'^2} + \hat{u}\right] = 0 \tag{5.39}$$

Furthermore, partially differentiating Eq.(5.33) with respect to u, and evaluating at $(\hat{C}', \hat{u})$, we may rewrite Eq.(5.36) as follows:

$$\sum_i \hat{C}_i' - C_a = 0 \tag{5.40}$$

The necessary conditions mean that the optimum solution $(\hat{C}', \hat{u})$ must satisfy Eqs.(5.38), (5.39), and (5.40) simultaneously. And, $L_i \to \infty$ for $\hat{C}_i' \to 0$, thus it must hold that $\hat{C}_i' > 0$.

Therefore, from Eqs.(5.38), (5.39), and (5.40), we obtain

$$-\alpha\left(\frac{\lambda_i}{\mu_i \hat{C}_i'}\right)^{\alpha-1} \frac{\lambda_i}{\mu_i \hat{C}_i'^2} + \hat{u} = 0 \tag{5.41}$$

or

$$\hat{C}_i' = \left(\frac{\alpha}{\hat{u}}\right)^{1/(1+\alpha)} \left(\frac{\lambda_i}{\mu_i}\right)^{\alpha/(1+\alpha)} \tag{5.42}$$

Substituting Eq.(5.42) into Eq.(5.32), we find

$$\sum_i \hat{C}_i' = \left(\frac{\alpha}{\hat{u}}\right)^{1/(1+\alpha)} \sum_i \left(\frac{\lambda_i}{\mu_i}\right)^{\alpha/(1+\alpha)} = C_a \tag{5.43}$$

from which

$$\left(\frac{\alpha}{\hat{u}}\right)^{1/(1+\alpha)} = \frac{C_a}{\sum_i \left(\frac{\lambda_i}{\mu_i}\right)^{\alpha/(1+\alpha)}} \tag{5.44}$$

From Eqs.(5.42) and (5.44), we obtain

$$\hat{C}_i' = C_a \frac{(\lambda_i/\mu_i)^{\alpha/(1+\alpha)}}{\sum_j (\lambda_j/\mu_j)^{\alpha/(1+\alpha)}} \tag{5.45}$$

Substituting Eq.(5.45) into Eq.(5.32), we arrive at

$$C_i^{(\alpha)} = \frac{\lambda_i}{\mu_i} + C_a \frac{(\lambda_i/\mu_i)^{\alpha/(1+\alpha)}}{\sum_j (\lambda_j/\mu_j)^{\alpha/(1+\alpha)}} \tag{5.46}$$

Finally, substituting Eq.(5.46) into Eqs.(2.6) and (2.4), we can easily obtain Eqs.(5.27) and (5.28). Q.E.D.

Next, we consider properties of Problem [B], and the relation between Problem [A] and Problem [B].

It is clear that for $\alpha=1$, Kleinrock's problem and its solution, i.e. the square root channel capacity assignment, are obtained.

(1) Concerning the total average message delay, the following equation holds good.

$$T^{(\alpha)}=T^{(1/\alpha)} \tag{5.47}$$

The variation among the average queue lengths L_i decreases by raising L_i to α-th power. On the contrary, the variation increases by raising L_i to $(1/\alpha)$-th power. However, from Eq.(5.47), we find that the total average message delay T is same for both cases.

As we mentioned above, the variation among the L_i decreases with increasing α. In particular, let us examine the case $\alpha \to \infty$.

(2) For $\alpha \to \infty$, we have

$$C_i^{(\infty)}=\frac{\lambda_i}{\mu_i}+C_a\frac{\lambda_i/\mu_i}{\sum_j \lambda_j/\mu_j} \tag{5.48}$$

This assignment gives capacity to a channel in direct proportion to the traffic amount (bits) carried by that channel. Commonly, it is known as the proportional channel capacity assignment. From Eqs.(5.37) and (5.48), the average queue lengths $L_i^{(\infty)}$ is given by

$$L_i^{(\infty)}=\frac{\Gamma}{C_a}\sum_j\frac{\lambda_j}{\mu_j} \tag{5.49}$$

From Eq.(5.49), it is recognized that, in the extreme case that $\alpha \to \infty$, we have no variation among the L_i.

From Eq.(5.48), the channel delay $T_i^{(\infty)}$ is obtained as follows:

$$T_i^{(\infty)}=\frac{1}{\lambda_i C_a}\sum_j\frac{\lambda_j}{\mu_j} \tag{5.50}$$

And, from Eqs.(5.50) and (2.3), the total average message delay is obtained as follows:

$$T^{(\infty)} = \frac{N}{\gamma C_a} \sum_j \frac{\lambda_j}{\mu_j} \tag{5.51}$$

(3) At the other extreme case that $\alpha \to 0$, we have

$$C_i^{(0)} = \frac{\lambda_i}{\mu_i} + \frac{C_a}{N} \tag{5.52}$$

From Eq.(5.27), the channel delay $T_i^{(0)}$ is obtained as follows:

$$T_i^{(0)} = \frac{N}{\mu_i C_a} \tag{5.53}$$

And, from Eq.(5.28), total average message delay $T^{(0)}$ is given as follows:

$$T^{(0)} = \frac{N}{\gamma C_a} \sum_i \frac{\lambda_i}{\mu_i} \tag{5.54}$$

At that extreme case, the assignment gives each channel its minimum required amount (λ_i/μ_i) plus a constant additional amount

Next, we assume that $\mu_i = \mu$. As we mentioned earlier, that assumption is reasonable. On that assumption, Eqs.(5.26), (5.27), and (5.28) reduce to as follows:

$$C_i^{(\alpha)} = \frac{\lambda_i}{\mu} + C_a \frac{\lambda_i^{\alpha/(1+\alpha)}}{\sum_j \lambda_j^{\alpha/(1+\alpha)}} \tag{5.55}$$

$$T_i^{(\alpha)} = \frac{1}{\mu C_a} \frac{\sum_j \lambda_j^{\alpha/(1+\alpha)}}{\lambda_i^{\alpha/(1+\alpha)}} \tag{5.56}$$

$$T^{(\alpha)} = \frac{I}{\gamma \mu C_a} \sum_i \lambda_i^{1/(1+\alpha)} \sum_j \lambda_j^{\alpha/(1+\alpha)} \tag{5.57}$$

At the case that $\alpha \to \infty$, we have

$$C_i^{(\infty)} = \frac{\lambda_i}{\mu} + C_a \frac{\lambda_i}{\sum_j \lambda_j}$$

$$= C \frac{\lambda_i/\mu}{\sum_j \lambda_j/\mu} \tag{5.58}$$

$$T_i^{(\infty)} = \frac{1}{C_a} \frac{\sum_j \lambda_j}{\lambda_i} \tag{5.59}$$

and

$$T^{(\infty)} = \frac{N}{\gamma \mu C_a} \sum_j \lambda_j \tag{5.60}$$

And, at the case that $\alpha \to 0$, we have

$$C_i^{(0)} = \frac{\lambda_i}{\mu} + \frac{C_a}{N} \tag{5.61}$$

$$T_i^{(0)} = \frac{N}{\mu C_a} \tag{5.62}$$

and

$$T^{(0)}=\frac{N}{\gamma\mu C_a}\sum_j \lambda_j \tag{5.63}$$

On this assumption that $\mu_i=\mu$, some close relations between Problem [A] and Problem [B] can be found.

First, we consider the optimum cahnnel capacity assignments for Problem [A] and Problem [B].

(4) The channel capacity assignment to decrease the variation among the queue lengths L_i in Problem [B], is the same as that to increase the variation among channel delays T_i in Problem [A], that is,

$$C_i^{(\alpha)}=C_i^{*(1/\alpha)} \tag{5.64}$$

From Eq.(5.64), it is recognized that if we decrease the variation among the L_i, the variation among the T_i increases, on the contrary, if we increase the variation among the L_i, the variation among the T_i decreases.

Concerning the total average message delay for these two problems, the following relation is found.

(5) The total average message delay given by the capacity assignment to reduce the variation among the L_i in Problem [B], is equal to that given by the capacity assignment to reduce the variation among the T_i in Problem [A], that is,

$$T^{(\alpha)}=T^{*(\alpha)} \tag{5.65}$$

It is amazing that, though the capacity assignment for Problem [A] is different from that for Problem [B], they have the same total average message delay.

5.4 Numerical Results and Considerations

The first numerical result which shows the effect of α in Problem [B], is obtained by using a simple two-channel model as shown in Fig.5.1.

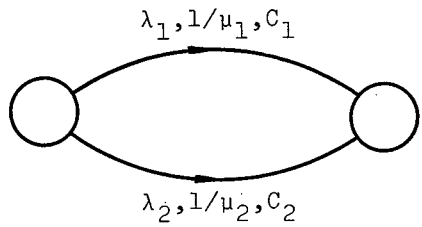

Fig.5.1 Two-channel model

In Fig.5.1, we have two channels from the first node to the second node. The i-th channel carries traffic λ_i (messages/sec), and the quantitative relation between λ_1 and λ_2 is as follows:

$$\lambda_1 = \xi \cdot \lambda_2 \tag{5.66}$$

Furthermore, we assume the rate of the average queue length on the first channel to that on the second channel is given by

$$L_1^{(\alpha)} / L_2^{(\alpha)} = \xi^{1/(1+\alpha)} \tag{5.67}$$

The behavior of this relation is shown in Fig.5.2, and the average queue lengths on both channels are shown in Fig.5.3, where ξ=0.1 and 0.5.

As shown in these figures, it is easily recognized that as α gets large, $L_1^{(\alpha)}$ increases and $L_2^{(\alpha)}$ decreases. As a result, the variation among the average queue lengths decreases, as we expected.

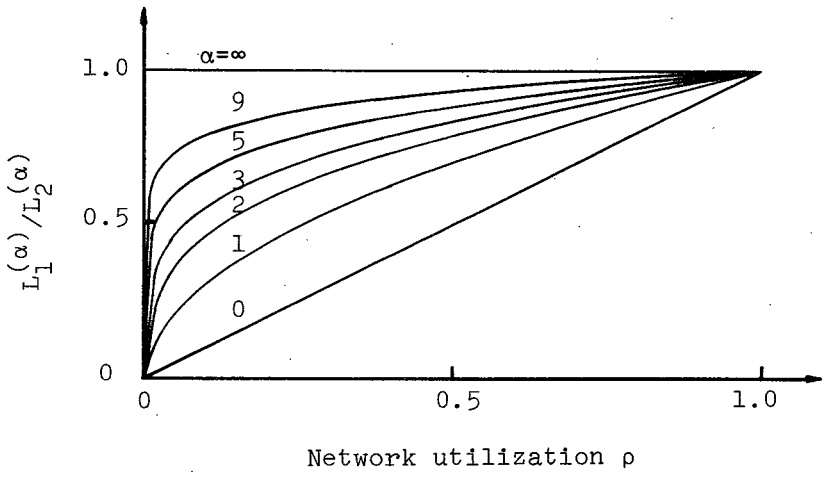

Fig.5.2 Rate of average queue length on the first channel to that on the second channel

It may also be seen that the reducing rate of the variation increases, as the ratio ξ decreases, i.e. the variation among λ_1 and λ_2 increases.

Now, let us consider the design of buffer size. Usually, the buffer size, which we denote by B [messages], is determined by the following equation.

$$\sum_{k=B}^{\infty} P(k) \leq \varepsilon \tag{5.68}$$

where P(k) is the steady state probability having k messages in the queue, and ε is a given block probability.

Since the channel is mathematically modeled as a queueing unit M/M/1, P(k) is given by

$$P(k)=(1-\rho)\rho^k \tag{5.69}$$

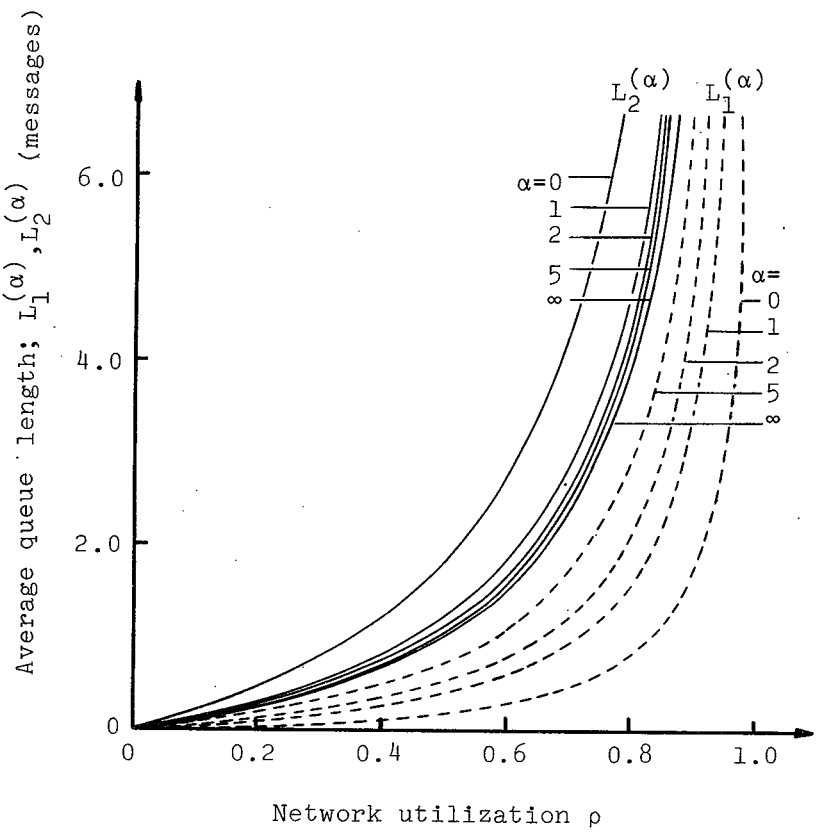

Fig.5.3 (a) Average queue length versus network utilizatin; ξ=0.1

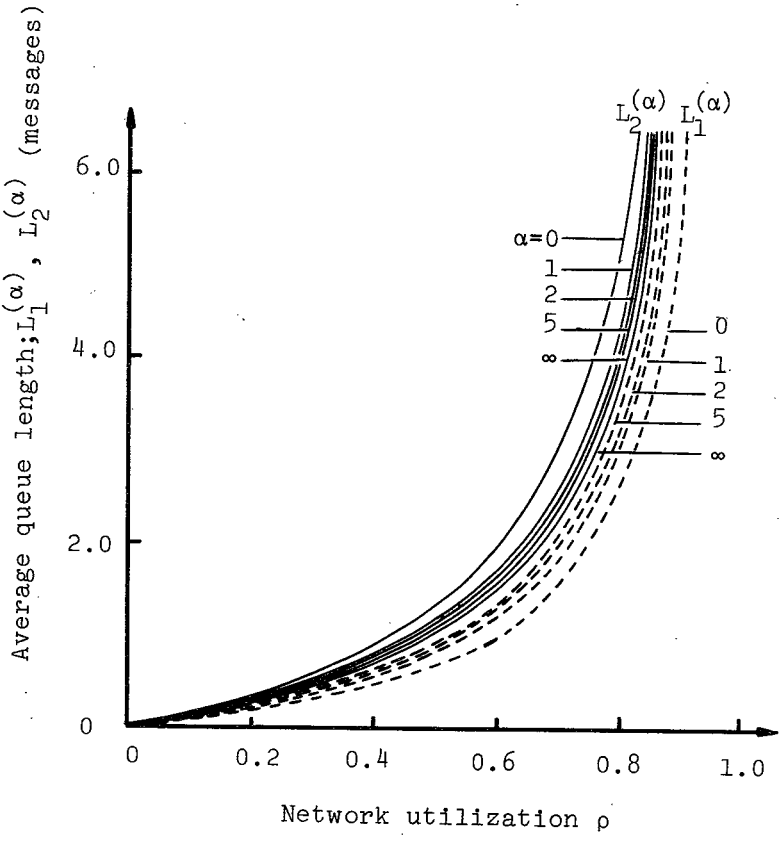

Fig.5.3 (b) Average queue length versus network utilization

ξ=0.5

where ρ is the channel utilization.

From Eqs.(5.68) and (5.69), we find that

$$B = \frac{\log \varepsilon}{\log \rho} \tag{5.70}$$

In Fig.5.4, we show the buffer sizes for the first channel and the second channel, which we denote by $B_1^{(\alpha)}$ and $B_2^{(\alpha)}$ respectively, where $\varepsilon=10^{-10}$†.

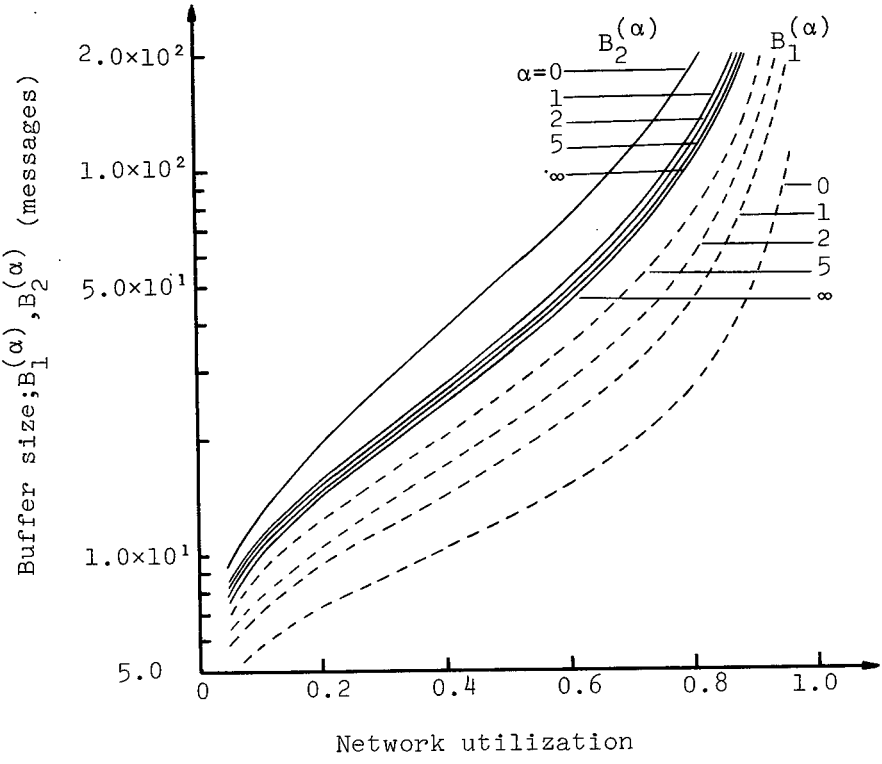

Fig.5.4 (a) Buffer size versus network utilization;ξ=0.1

† For $\varepsilon=10^{-b}$, the buffer size is b/10 times as great as that for $\varepsilon=10^{-10}$.

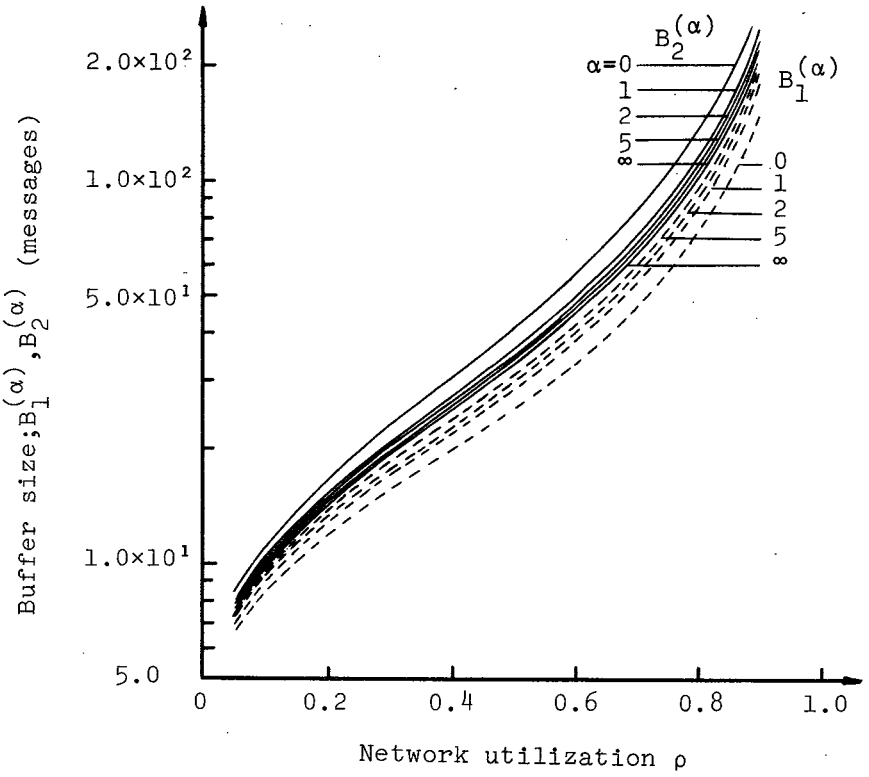

Fig.5.4 (b) Buffer size versus network utilization;ξ=0.5

As shown in Fig.5.4, we can decrease the difference between $B_1^{(\alpha)}$ and $B_2^{(\alpha)}$ by increasing value of α, which may also be deduced from Fig.5.3. The reduction of the difference is due to the decreasing of $B_2^{(\alpha)}$ and the increasing $B_1^{(\alpha)}$. Furthermore, as ξ gets large, the difference decreases.

Now, it is costly to implement many buffers with various sizes for cahnnels in a network. Thus, it is desirable to standardize the buffer size. For the above two-channel model, $B_2^{(\alpha)}(>B_1^{(\alpha)})$ may be considered as the standardized buffer size . Therefore, it is able to make the standardized buffer size small by getting α large.

Next, we consider a ladder network with six nodes and fourteen channels as shown in Fig.5.5.

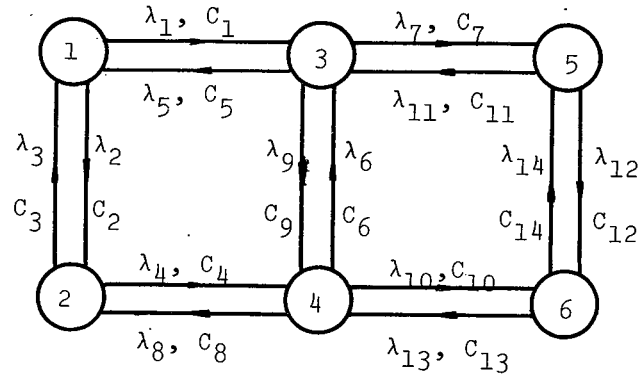

Fig.5.5 Ladder network

For the numerical computations, we give a relative traffic matrix as follows:

Destination node

Source node	1	2	3	4	5	6
1	0	1.0	1.0	1.0	1.0	1.0
2	1.0	0	1.0	1.0	1.0	1.0
3	1.0	1.0	0	1.0	1.0	1.0
4	1.0	1.0	1.0	0	1.0	1.0
5	1.0	1.0	1.0	1.0	0	1.0
6	1.0	1.0	1.0	1.0	1.0	0

and give a fixed routing of messages as shown in TABLE 5.1. As a result, the quantitative relation among the traffics λ_i is as follows:

$$\lambda_1=\lambda_3=\lambda_7=\lambda_8=\lambda_{12}=\lambda_{13}$$

$\lambda_2=\lambda_4=\lambda_5=\lambda_{10}=\lambda_{11}=\lambda_{14}$

$\lambda_6=\lambda_9$

$\lambda_1:\lambda_2:\lambda_6=1.0:0.6:0.2$

TABLE 5.1 FIXED ROUTING OF MESSAGES

Source Node \ Destination Node	1	2	3	4	5	6
1	*	c_2	c_1	c_2,c_4	c_1,c_2	c_1,c_7,c_{12}
2	c_3	*	c_3,c_1	c_4	c_3,c_1,c_7	c_4,c_{10}
3	c_5	c_5,c_2	*	c_6	c_7	c_7,c_{12}
4	c_8,c_3	c_8	c_9	*	c_{10},c_{14}	c_{10}
5	c_{11},c_5	c_{12},c_{13},c_8	c_{11}	c_{12},c_{13}	*	c_{12}
6	c_{13},c_8,c_3	c_{13},c_8	c_{14},c_{11}	c_{13}	c_{14}	*

In TABLE 5.2, we show the total average message delay given by the optimum channel capacity assignment for Problem [B]. As a matter of course, for $\alpha=1$, the capacity assignment is the square root channel capacity assignment which gives the minimum value of total average message delay. And, it is also recognized that as α gets large, both $T^{(\alpha)}$ and $T^{(1/\alpha)}$ increase. Of course, $T^{(\alpha)}=T^{(1/\alpha)}$, and $T^{(\alpha)}=T^{*(\alpha)}$.

TABLE 5.2 TOTAL AVERAGE MESSAGE DELAY (sec)

ρ \ α	1	2	4	8	16	32	64	128	∞
		$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{128}$	0
0.1	0.247	0.248	0.251	0.254	0.256	0.258	0.259	0.259	0.259
0.3	0.317	0.319	0.323	0.327	0.330	0.331	0.332	0.333	0.333
0.5	0.444	0.447	0.452	0.458	0.462	0.464	0.465	0.466	0.467
0.7	0.740	0.744	0.753	0.763	0.769	0.773	0.775	0.777	0.778
0.9	2.222	2.232	2.260	2.288	2.308	2.320	2.326	2.330	2.333

In Fig.5.6, we show the increasing rate of total average message delay for $\alpha=1$, to that for positive real value α. The increasing rate may be written as $(T^{(\alpha)}-T^{(1)})/T^{(1)}$. From Fig.5.6, it is recognized that as α increases, the increasing rate increases rather fast at the case that $0<\alpha<8$, and more slowly at the case that $\alpha>8$. However, we may find that the increasing rate is not so large.

In Fig.5.7 and 5.8, we show the variation among the channel delays, which we denote by σ_T, and the variation among the

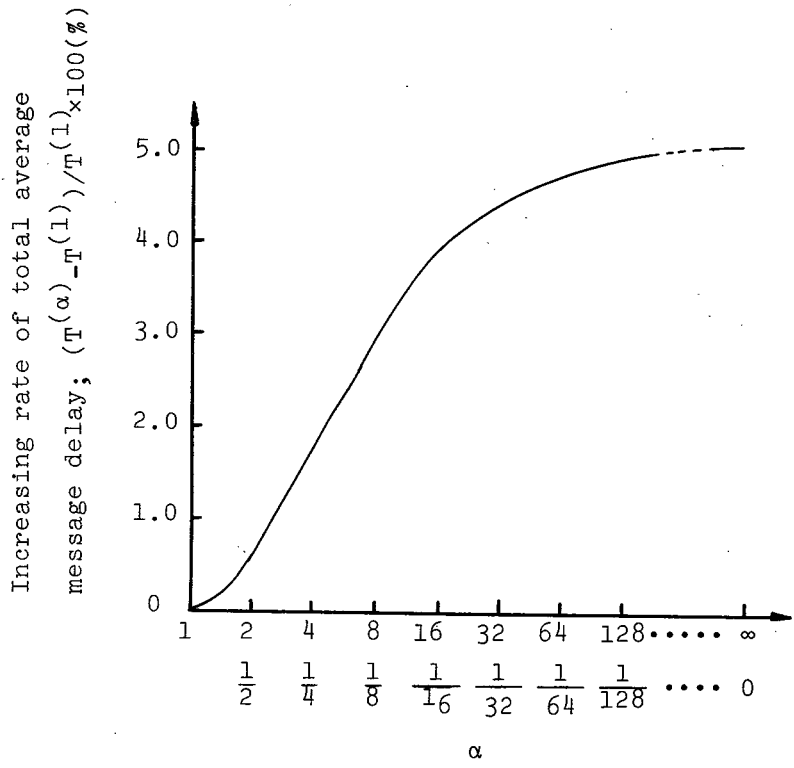

Fig.5.6 Increasing rate of total average message delay

queue lengths, which we denote by σ_L. These variations are defined as follows:

$$\sigma_T=\left[\sum_i \frac{1}{N}(T_i-\overline{T})^2\right]^{1/2} \tag{5.71}$$

$$\sigma_L=\left[\sum_i \frac{1}{N}(L_i-\overline{L})^2\right]^{1/2} \tag{5.72}$$

where

$$\overline{T}=\sum_i \frac{T_i}{N} \tag{5.73}$$

$$\overline{L}=\sum_i \frac{L_i}{N} \tag{5.74}$$

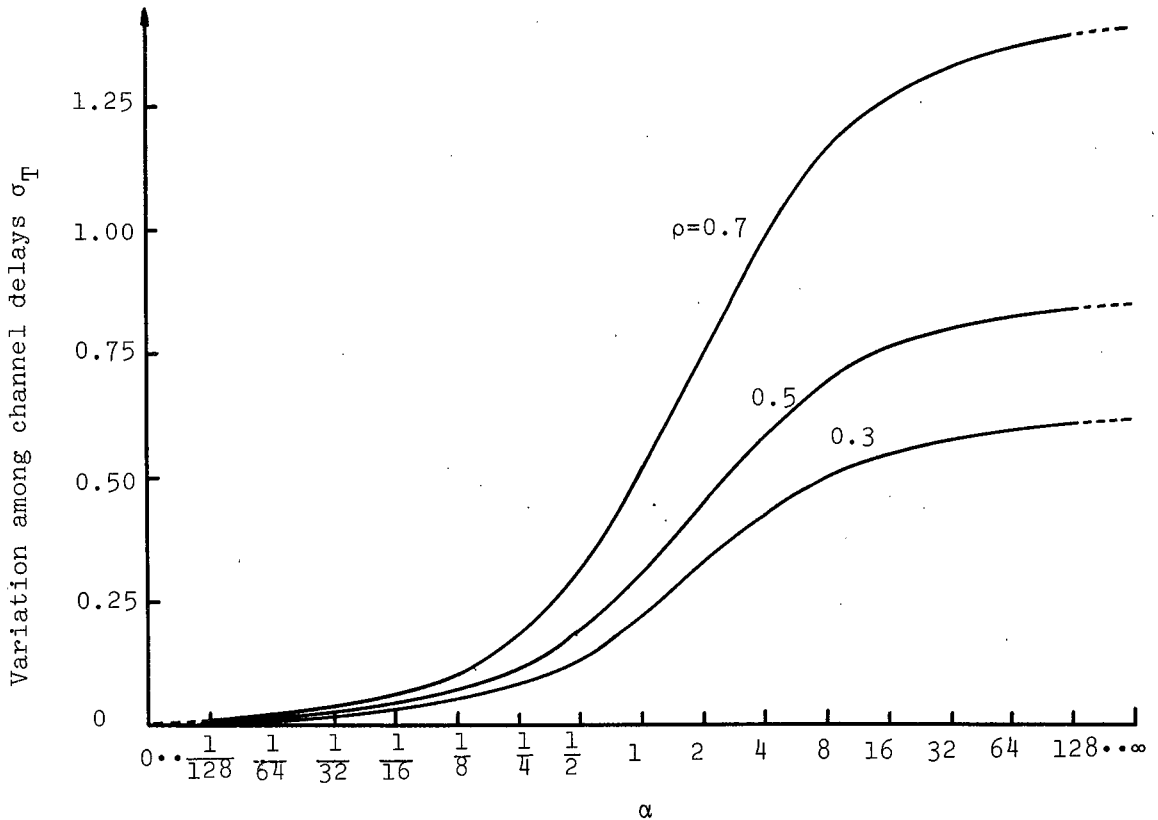

Fig.5.7 Variation among channel delays

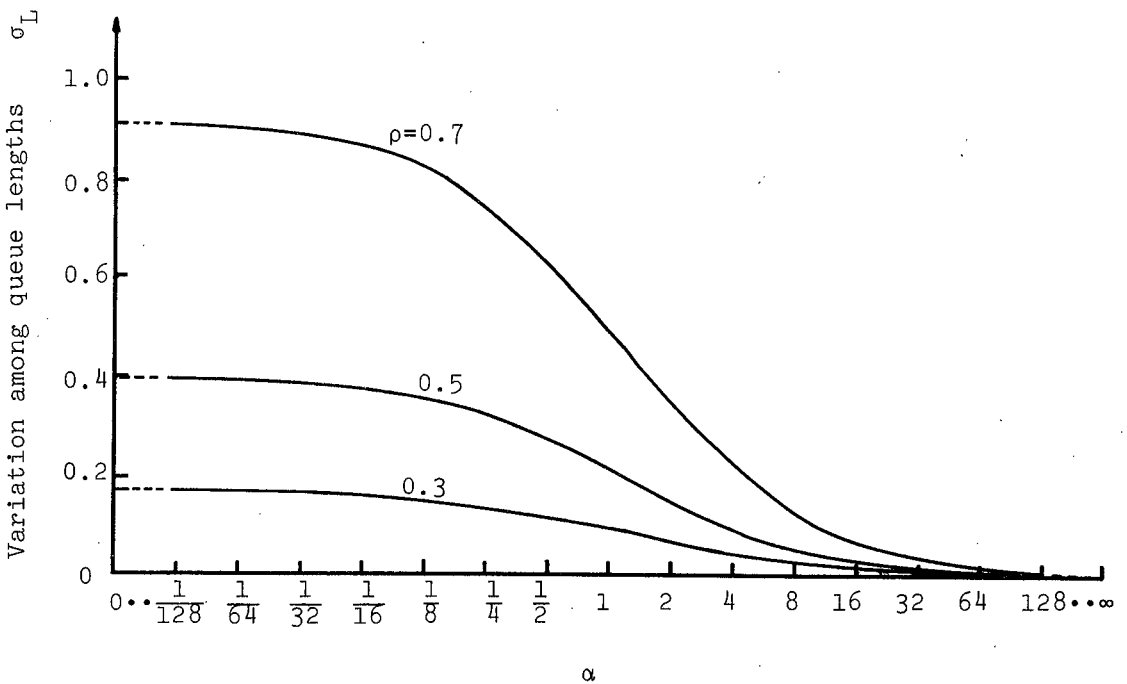

Fig.5.8 Varation among queue lengths

As shown in Fig.5.7, σ_T increases as α gets large, which we expected earlier. Moreover, it is recognized that the increasing amount of σ_T is large for large network utilization ρ. At the extreme case that $\alpha \to 0$, in which the excess capacity is divided into equal amount for each channels, we have no variation among channel delays. And, as shown in Fig.5.8, σ_L decreases as α increases, which is mentioned earlier. Furthermore, we find that the increasing amount of σ_L increases as network utilization increases. At the extreme case that $\alpha \to \infty$, which gives the proportional channel capacity assignment, we have no variation among queue lengths. Therefore, it is obvious that the channel capacity assignment for Problem [A] has a dual relation to the assignment for Problem [B].

5.5 Conclusion

In summary, a new extended optimum channel capacity assignment problem has been formulated for store-and-forward communication networks. The optimum channel capacity assignment problem as first given by Kleinrock is to achieve the minimum total average message delay, which may be also interpreted as a problem to minimize the number of messages within the network. The new extended problem is to find the channel capacity assignment to reduce variation among queue lengths. The solution for the extended problem results in:(1)Total average message delay for the channel capacity assignment reducing variation among the queue lengths, is equal to the assignment increasing that variation. (2)The proportional channel capacity assignment has no variation among the queue lengths. (3) The channel capacity assignment which gives each channel its

required capacity plus a constant additional capacity, has the maximum variation among the queue lengths.

Furthermore, by assuming that all of channels have the same average message length, it has been found that there exists a dual relation between the new extended problem and the other extended problem reducing variation among channel delays given by Meister et al. as follows:(1) The channel capacity assignment making the variation among the queue lengths to the α degree, is the same as the channel capacity assignment making the variation among the channel dealys to the $1/\alpha$ degree. (2) The former assignment is different from that making the variation among the channel delays to the α degree, but both have the same total average message delay.

CHAPTER 6

CONCLUSIONS

Some important objectives of this thesis were to gain useful techniques applicable to store-and-forward computer communication networks.

The several significant results in this thesis are summarized as follows:

(1) Optimum route assignment problem

The optimum route assignment theorem has been obtained, which gives the necessary and sufficient conditions to find the optimum route assignment minimizing total average message delay.

For a simple multiple-channel model, the optimum route assignment has been compared with the equal-delay-principle route assignment.

As a result, it has been found that as traffic rate increases, the commencement of detour for the optimum assignment appears earlier than that for the equal-delay-principle assignment.

(2) Adaptive routing procedure

A new adaptive routing procedure based on the optimum route assignment theorem has been proposed, in which an estimated value of $\partial L/\partial \lambda$ is used as routing informations, where L is the average queue length, and λ is the average traffic rate.

From simulation results, it has been verified that the new procedure can achieve smaller total average message delay than ARPA procedure.

(3) Optimum channel capacity assignment problem

The optimum channel capacity assignment theorem has been obtained, which gives the necessary and sufficient conditions

to find the optimum channel capacity assignment minimizing total average message delay in the case of general message length.

(4) Extended optimum channel capacity assignment problems

An extended optimum channel capacity assignment problem has been formulated and solved. It is a channel capacity assignment problem to reduce variation among queue lengths. It has been found that this extended problem has a dual relation with the extended problem given by Meister et al., which gives a channel capacity assignment reducing variation among channel delays.

In this thesis, we have developed several optimization problems on store-and-forward computer communication networks which deal with one kind of messages with one average message lengh. Therefore, the optimization problems for the network which deal with many kind of messages, for example, interactive message and file message, are left unsolved for future research.

REFERENCES

[1] L.G.Roberts and B.D.Wessler," Computer Network Development to Achieve Resource Sharing," Spring Joint Computer Conf., 1970

[2] Mckay,et al.," IBM Computer Network/440." Courant Computer Science Symposium 3, December 1970

[3] S.F.Mendicino," OCTPUS: The Lawrence Radiation Laboratory Network," Courant Computer Science Symposium 3, December 1970

[4] W.D.Farmer and E.E.Newhall," An Experimental Distributed Switching System to Handle Bursty Computer Traffic," Proc. A.C.M.Symposium, October 1969

[5] J.R.Pierce," Network for Block Switching of Data," Bell Syst. Tech. J., 51, No. 6, July 1972

[6] F.E.Heart, R.E.Kahn, S.M.Ornstein, W.R.Crowther, and D.C. Walden," The Interface Message Processor for the ARPA Computer Network," Spring Joint Computer Conf., AFIPS Conf. Proc., May 1970

[7] C.S.Carr, S.D.Crocker, and V.G.Cerf," HOST-HOST Communication Protocol in the ARPA Network," Spring Joint Computer Conf., AFIPS Conf. Proc., May 1970

[8] S.M.Ornstein, F.E.Heart, W.R.Crowther, H.K.Rising, S.B.Russel and A.Michel," Terminal IMP for ARPA Computer Network," Spring Joint Computer Conf., AFIPS Conf. Proc., 1972

[9] L.Kleinrock," Analysis and Simulation Methods in Computer Network Design," Spring Joint Computer Conf., AFIPS Conf. Proc., 1970

[10] D.W.Davies et al .," A Digital Communication Network for Computer Giving Rapid Response at Remote Terminal," ACM Symposium on Operating System Principles, October 1967

[11] D.W.Davies," Packet Switching in a Public Data Network," IFIP Congress,1971

[12] H.Miyahara, T.Hasegawa, and Y.Teshigawara," A Comparative Analysis of Switching Methods in Computer Communication Networks," Proc., International Conference on Communications, June 1975

[13] R.D.Rosner," Packet Switching and Circuit Switching: A Comparison," Proc., National Telecommunications Conf., Atlanta, Georgia, November 1975

[14] K.Kummerle," Multiplexer Performance for Integlated Line and Packet Switched Traffic," Proc., International Conference on Computer Communication, 1974

[15] H.Okada and Y.Tezuka," Block Switching System," Trans. IECE, Japan, 59-A, No.4, April 1976 (in Japanese)

[16] L.Kleinrock," Communication Nets: Stochastic Message Flow and Delay," McGraw-Hill, 1964

[17] P.J.Burke," The Output of a Queueing System," Operations Research, Vol.4, 1956

[18] J.R.Jackson," Network of Waiting Lines," Operations Research, Vol.5, 1974

[19] H.Miyahara, H.Sanada, and T.Hasegawa," A Method for Determining Transmission Line and Buffer Capacities in Store-and-Forward Switched Networks," Trans. IECE, Japan, 60-A, No.2, February 1977 (in Japanese)

[20] G.L.Fultz," Adaptive Routing Techniques for Message Switching Computer Communication Networks," Doctor Dissertation of UCLA, 1972

[21] I.Rubin," Communication Networks: Message Path Delays," IEEE Trans. Inform. Theory, Vol.IT-20, No.6, November 1974

[22] I.Rubin," Message Path Delays in Packet-Switching Communication Networks," IEEE Trans. Communications, Vol. COM-23, No.2, February 1975

[23] I.Rubin," An Approximate Time-Delay Analysis for Packet-Switching Communication Networks," IEEE Trans. Communications, Vol. COM-24, No.2, February 1976

[24] H.Okada," Analysis of Intervening Probability Distribution in Packet Switching Computer Networks," Trans. IECE., Japan, 55-A, Vol.12, December 1972 (in Japanese)

[25] H.Okada," Analysis of Intervening Probability in Packet Switching Networks with Transmission Precedence," Trans. IECE, Japan, Vol. 59-A, No.12, December 1976 (in Japanese)

[26] O.Hashida," Analysis of Intervening Packets in a Packet Switched Network," Trans. IECE, Japan, Vol.59-A, Vol.2, February 1976

[27] D.Doll," Efficient Allocation of Resources in Centralized Computer-Communication Network Design," SEL. Tech. Rept. 36, University of Michigan, 1969

[28] H.Frank, I.T.Frisch, and W.Chou," Topological Considerations in the Design of the ARPA Computer Network," Spring Joint Computer Conf., AFIPS Conf. Proc., May 1970

[29] H.Frank, I.T.Frisch, R.Van Slyke, and W.S.Chou," Optimal Design of Centralized Computer Networks," Networks, 1, 1971

[30] B.Meister, H.R.Müller, and H.R.Rubin," New Optimization Criteria for Message-Switching Networks," IEEE Trans. Communications, Vol. COM-19, No.3, June 1971

[31] B.Meister, H.R.Müller, and H.R.Rubin," On the Optimization of Message-Switching Networks," IEEE Trans. Communications, Vol. COM-20, No.1, February 1972

[32] H.Frank and I.T.Frisch," Communication, Transmission and Transportation Networks," Addison-Wesley, Ready, Mass., 1971

[33] L.K.Ford,Jr., and D.R.Filherson," Maximal Flow through a Network," Can. J. Math, 8, 1956

[34] B.Rothford and I.T.Frisch," On The 3-Commodity Flow Problem," SIAM J. Appl. Math. 17, 1969

[35] H.Sanada and Y.Tezuka," Route Assignment on Computer Network," Pacific Area Computer Communication Network System Symposium, August 1975

[36] D.G.Cantor and M.Gerla," Optimal Routing in a Packet-Switching Computer Network," IEEE Trans. Computers, Vol. C-23, No.10, October 1974

[37] L.Fratta, M.Gerla, and L.Kleinrock," The Flow Deviation Method; An Approach to Store-and-Forward Communication Network Design," Networks, 3, 1973

[38] M.Schwartz and C.K.Cheung," The Gradient Projection Algorithm for Multiple Routing in Message-Switched Networks," IEEE Trans. Communications, Concise Paper, April 1976

[39] K.T.Prosser," Routing Procedures in Communications Network -Part I ; Random Procedure," IRE Trans. Communication Systems, Vol. CS-10, December 1962

[40] K.T.Prosser," Routing Procedures in Communications Network -Part II ; Directory Procedures," IRE Trans. Communication Systems, Vol. CS-10, December 1962

[41] B.W.Boehm and R.L.Mobley," Adaptive Routing Technique for Distributed Communication System," IEEE, Trans. Communications, Vol. COM-17, No.3, June 1969

[42] G.L.Fultz and L.Kleinrock," Adaptive Routing Techniques for Store-and-Forward Computer-Communication Network," IEEE I.C.C. 1971

[43] J.M.McQwillan," Adaptive Routing Algorithms for Distributed Computer Networks," Bolt Beranek and Newman Inc., Cambridge, MA, Report No. 2831, May 1974

[44] H.Rubin," On Routing and " Delta Routing ": A Taxonomy and Performance Comparison of Techniques for Packet-Switched Networks," IEEE Trans. Communications, Vol. COM-24, No. 1, January 1976

[45] A.Butrimenko," Routing Technique for Message Switching Networks with Message Outdating," Symposium on Computer-Communications Networks and Teletraffic Polytechnic Institute of Brooklyn, April 1972

[46] R.L.Pickholtz and C.Mcloy,Jr.," Effcts of a Priority Discipline in Routing for Packet-Switched Networks," IEEE Trans. Communications, Vol. COM-24, No. 5, May 1976

[47] M.C.Pennoti and M.Schwartz," Congestion Control in Store-and-Forward Tandem Links," IEEE Trans. Communications, Vol. COM-23 No. 12, December 1975

[48] H.Sanada, T.Nishizawa and Y.Tezuka," Analysis of Blocking Phenomena on Computer Network," Seventh Hawaii International Conference on System Sciences Technical Report CN 74-14, January 1974

[49] R.E.Kahn and W.R.Croether," Flow Control in a Resource-Sharing Network," IEEE Trans. Vol. COM-20, No. 3, June 1972

[50] J.Herrman," Flow Control in the ARPA Network," Computer Networks, 1, 1976

[51] D.W.Davies," The Control of Congestion in Packet Switching Network," IEEE Trans. Communication, Vol. COM-20, No. 3, June 1972

[52] W.L.Price," Simulation Studies of an Isarithmically Controlled Store and Forward Data Communication Network," Information Processing 74, 1974

[53] W.L.Price," Simulation of Packet Switching Networks Controlled on Isarithmic Principle," Proc. of Third ACM/ IEEE Data Communications Symposium, November 1973

[54] H.Okada, M.Suzuki, T.Sunouchi and Y.Tezuka," Analysis of the Isarithmic Flow Control Method in Packet Switching Computer Networks," Trans. IECE, Vol. 59-A, No. 3, March 1976

[55] T.Honma," Queueing Theory," Rikogakusha, 1966 (in Japanese)

[56] H.W.Kuln and A.W.Tucker," Nonlinear Programming," Proc. the Second Barkley Symposium on Mathematical Statics and Probability, Barkley, California, 1950

[57] Wadrop," Some Theoretical Aspect of Road Traffic Research," Proc. Institute of Civil Engineers, 1952

[58] L.Kleinrock," Queueing Systems," A Wiley-Interscience Publication, New York, 1975

[59] S.P.Boehm and P.Baran," On Distributed Communication-II Digital Simulation of Hot-Potato Routing in a Broadband Distributed Communication Networks," RAND Corp., Rep. RM 3101-PR, August 1964

[60] Y.Ouchi," Technique of Computer Communication Simulator," Graduation Thesis, Osaka University, March 1976

[61] M.Komatsu, Y.Hayasida, J.Shiomi, Y.Tezuka," On Access Methods for a Circular Data Network," Papers of Technical Group on Switching, IEEE, Japan, SE74-54, 1974 (in Japanese)

[62] M.Komatsu, H.Nakanishi and Y.Tezuka," Traffic Analysis of a Loop Network," Papers of Technical Group on Switching., IEEE, Japan, SE75-5, 1975 (in Japanese)

[63] M.Komatsu, H.Nakanishi and Y.Tezuka," Traffic Analysis of a Two-Way Loop Network," IECE, Proc. of Annual Convention, March 1975 (in Japanese)

[64] J.Ishii, M.Komatsu, H.Sanada and Y.Tezuka," Route Assignment Theorem on Computer Network," Papers of Technical Group on Switching, IECE, Japan, SE75-75, 1975 (in Japanese)

[65] M.Komatsu, H.Nakanishi, H.Sanada and Y.Tezuka," Adaptive Routing Procedure Based on Route Assignment Theorem for Computer Network [1]," Papers of Technical Group of Switching, IECE, Japan, SE75-76, 1975 (in Japanese)

[66] J.Ishii, M.Kimatsu, H.Sanada and Y.Tezuka," Optimum Route Assignment Theorem on Store-and-Forword Switching Data Communication Network," IECE, Japan, Proc. of the Annual Convention, March. 1976 (in Japanese)

[67] M.Komatsu, H.Nakanishi, H.Sanada and Y.Tezuka," Adaptive Routing Procedure based on the Optimum Route Assignment Theorem," IECE, Japan, Proc. of the Annual Convention. March 1976 (in Japanese)

[68] M.Komatsu, H.Nakanishi, H.Sanada and Y.Tezuka," On Optimal Capacity Assignment Problems in the Case of General Service Distribution," Papers of Technical Group on Switching, IECE, Japan, SE-76-38, 1976, (in Japanese)

[69] M.Komatsu, H.Nakanishi, H.Sanada and Y.Tezuka," Optimal Route Assignment Problems for Data Communication Networks with General Service Distribution," Papers of Technical Group on Switching, IECE, Japan, SE-76-38, 1976 (in Japanese)

[70] M.Komatsu, Y.Ouchi, H.Nakanishi, H.Sanada and Y.Tezuka, " Adaptive Routing Procedure Based on Route-Assignment Theorem for Computer Network (2)," Papers of Technical Group on Switching, IECE, Japan, SE-76-70, 1976 (in Japanese)

[71] M.Komatsu, Y.Ouchi, H.Nakanishi, H.Sanada and Y.Tezuka, " Optimum Route Controlling Method on Computer Networks," Papers of Technical Group on Computer Network, (PS, Japan) CN9-6, 1977 (in Japanese)

[72] M.Komatsu, H.Nakanishi, H.Sanada and Y.Tezuka," Minimum Average Delay Theorems for Computer Networks," Technology Reports of the Osaka University, Vol. 27, No. 1375, 1977

[73] M.Komatsu, Y.Ouchi, H.Nakanishi, H.Sanada and Y.Tezuka, " Optimum Route Assignment Theorem and Its Application to Adaptive Routing Procedure in Message-Switching Network," Trans. IECE, Vol. 60-B, No. 12, December 1977 (in Japanese)

[74] M.Komatsu, H.Nakanishi, H.Sanada and Y.Tezuka," Extension of Optimum Channel Capacity Assignment Problem on Message-Switching Network," IECE, Japan, Proc. of Annual Convention, March 1978 (in Japanese to be published)

[75] R.T.Rockfeller," Convex Analysis;" Princeton, New Jersey, Princeton University Press, 1970

APPENDIX A

GENERAL INDEPENDENCE ASSUMPTION TESTS FOR A NETWORK

Fultz introduced the general independence assumption, and showed its validity for a simple model with three nodes. In this appendix, its validity is verified for a more complex network model with six nodes and fourteen channels as shown in Fig.A.1,

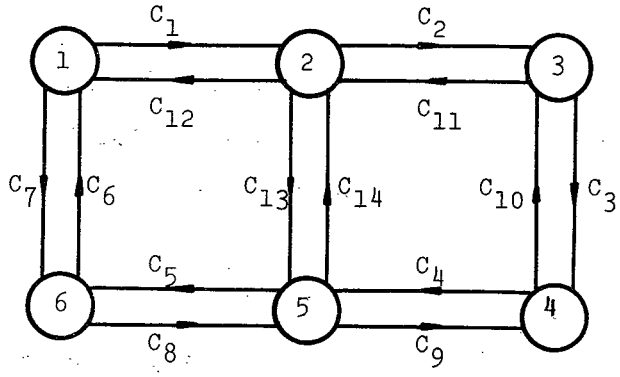

Fig.A.1 Model for general independence assumption test

In this model, it is assumed that the distribution of message length is Erlangian with phase k=1,3, and ∞, which we denote by E_1, E_2, and E_∞ respectively, the channel capacities are given by

$$C_i = \begin{cases} 10 \text{ kbits/sec} & ;i=1,2,\cdots,6 \\ 6 \text{ kbits/sec} & ;i=7,8,\cdots,12 \\ 2 \text{ kbits/sec} & ;i=13,14 \end{cases}$$

and the relative traffic matrix is given as follws:

destination node

source node	1	2	3	4	5	6
1	0	1	1	1	1	1
2	1	0	1	1	1	1
3	1	1	0	1	1	1
4	1	1	1	0	1	1
5	1	1	1	1	0	1
6	1	1	1	1	1	0

Furthermore, a fixed routing for messages is shown in Fig.A.2. In this model, each channel may be mathematically modeled as $M/E_k/1$, thus the channel delay T_i is given by

$$T_i = \frac{1}{\mu C_i} + \left(1+\frac{1}{k}\right)\frac{\rho_i}{2\mu C_i(1-\rho_i)} \qquad (A.1)$$

From Eqs.(2.3) and (A.1), the total average message delay is found as follows:

$$T = \sum_{i=1}^{14} \left[\rho_i + \left(1+\frac{1}{k}\right)\frac{\rho_i^2}{2(1-\rho_i)}\right]/\gamma \qquad (A.2)$$

Figure A.3 shows a comparison of the total average message delay obtained from Eq.(A.2) and that obtained from simulation. As shown in Fig.A.3, it is confirmed that there is adequate agreement between simulation data and theoretical result.

Destination Node

Source Node	1	2	3	4	5	6
1	*	c_1	c_1, c_2	c_1, c_2, c_3	c_7, c_8	c_7
2	c_{12}	*	c_2	c_2, c_3	c_{13}	c_{12}, c_7
3	c_{11}, c_{12}	c_{11}	*	c_3	c_3, c_4	c_3, c_4, c_5
4	c_4, c_5, c_6	c_{10}, c_{11}	c_{10}	*	c_4	c_4, c_5
5	c_5, c_6	c_{14}	c_9, c_{10}	c_9	*	c_5
6	c_6	c_6, c_1	c_6, c_1, c_2	c_8, c_9	c_8	*

Fig.A.2 Fixed routing for general independence assumption test

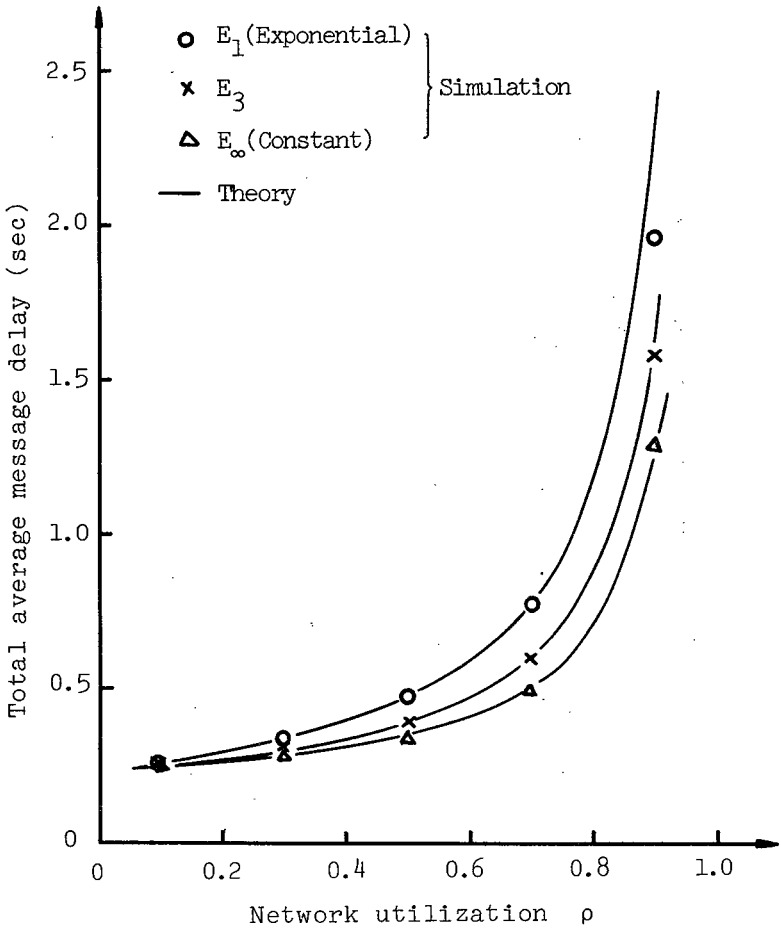

Fig.A.3 Comparison between simulated total average message delay and theoretical result.

APPENDIX B

PROOF OF EQ.(2.3)

Average message delay for messages with source node N_s and destination node N_d, which we denote by Z_{sd}, is given by

$$Z_{sd}=\frac{1}{\gamma_{sd}}\sum_{k=1}^{n_{sd}}\gamma_{sd}^{(k)}Z_{sd}^{(k)} \tag{B.1}$$

where

n_{sd} ;number of routes from N_s to N_d

$\gamma_{sd}^{(k)}$;average number of messages transmitted on the k-th route $R_k(s,d)$ from N_s to N_d

$Z_{sd}^{(k)}$;average delay for messages transmitted on the route $R_k(s,d)$ from N_s to N_d

and

$$\gamma_{sd}=\sum_{k=1}^{n_{sd}}\gamma_{sd}^{(k)} \tag{B.2}$$

By defining that

$$\delta_{sd}^{(k)}(i)=\begin{cases}1 & ;B_i\varepsilon R_k(s,d)\\ 0 & ;B_i\not\varepsilon R_k(s,d)\end{cases} \tag{B.3},$$

Equation (B.1) may be rewritten by follows:

$$Z_{sd}=\frac{1}{\gamma_{sd}}\sum_{k=1}^{n_{sd}}\gamma_{sd}^{(k)}\sum_{i=1}^{N}\delta_{sd}^{(k)}(i)T_i \tag{B.4}$$

where

B_i ; the i-th channel in the network

N ; number of lines in the network

T_i ;average channel delay for B_i

On the other hand, the total average message delay is obtained by

$$T = \sum_{s,d} \frac{\gamma_{sd}}{\gamma} Z_{sd} \tag{B.5}$$

Substituting Eq.(B.4) into (B.5), we obtain

$$T = \sum_{s,d} \frac{\gamma_{sd}}{\gamma} \frac{1}{\gamma_{sd}} \sum_{k=1}^{n_{sd}} \gamma_{sd}^{(k)} \sum_{i=1}^{N} \delta_{sd}^{(k)}(i)\, T_i \tag{B.6}$$

We may change the order of summation for triple sum and regroup terms such that

$$T = \sum_{i=1}^{N} \frac{T_i}{\gamma} \sum_{s,d} \sum_{k=1}^{n_{sd}} \delta_{sd}^{(k)}(i)\, \gamma_{sd}^{(k)} \tag{B.7}$$

Since

$$\lambda_i = \sum_{s,d} \sum_{k=1}^{n_{sd}} \delta_{sd}^{(k)}(i)\, \gamma_{sd}^{(k)} \tag{B.8}$$

we arrive at

$$T = \sum_{i=1}^{N} \frac{\lambda_i}{\gamma} T_i$$

APPENDIX C

PROOF THAT EQ.(2.11) IS A CONCAVE FUNCTION OF VECTOR $x=[x_j^{sd}]$

From Eqs.(2.6),(2.7), and (2.4), it is easily recognized that L_i is a convex function of λ_i, which may be written as follows:

$$L_i\Big|_{\lambda_i=\alpha\lambda+(1-\alpha)\lambda'} \le \alpha L_i\Big|_{\lambda_i=\lambda} + (1-\alpha)L_i\Big|_{\lambda_i=\lambda'} \qquad (C.1)$$

Since the relationship between λ_i and x_j^{sd} is given by Eq.(2.9), we have

$$\begin{aligned} & L_i\Big|_{x=\alpha y+(1-\alpha)z} \\ &= L_i\Big|_{\lambda_i=\sum_{j,s,d|B_i\epsilon R_j(s,d)} [\alpha y_j^{sd}+(1-\alpha)z_j^{sd}]\gamma_j^{sd}} \\ &= L_i\Big|_{\lambda_i=[\alpha\sum_{j,s,d|B_i\epsilon R_j(s,d)} y_j^{sd}\gamma_j^{sd} \\ &\qquad +(1-\alpha)\sum_{j,s,d|B_i\epsilon R_j(s,d)} z_j^{sd}\gamma_j^{sd}]} \qquad (C.2) \end{aligned}$$

Similarly, we have

$$L_i\Big|_{x=y} = L_i\Big|_{\lambda_i=\sum_{j,s,d|B_i\epsilon R_j(s,d)} y_j^{sd}\gamma_j^{sd}} \qquad (C.3)$$

$$L_i\Big|_{x=z} = L_i\Big|_{\lambda_i=\sum_{j,s,d|B_i\epsilon R_j(s,d)} z_j^{sd}\gamma_j^{sd}} \qquad (C.4)$$

From Eqs.(C.1),(C.2),(C.3), and (C.4), it is found that

$$L_i\Big|_{x=\alpha y+(1-\alpha)z} \le \alpha L_i\Big|_{x=y} + (1-\alpha)L_i\Big|_{x=z} \qquad (C.5)$$

Equation (C.5) implies that L_i is a convex function with respect to $x=[x_j^{sd}]$. Thus, the linear sum $\Sigma L_i/\gamma$, which has positive coefficients $1/\gamma$, is also convex [75]. This is the same that Eq.(2.11) is a concave function of x. Q.E.D.