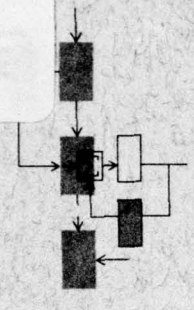
May, 1976

Report ESL-R-661

ARPA Contract N00014-75-C-1183



DIFFUSION MODELS FOR COMPUTER-COMMUNICATIONS NETWORK

José H. Barberá

Electronic Systems Laboratory *Decisions and Control Sciences Group*

MASSACHUSETTS INSTITUTE OF TECHNOLOGY, CAMBRIDGE, MASSACHUSETTS 02139

Department of Electrical Engineering and Computer Science

May, 1976

Report ESL-R-661

DIFFUSION MODELS FOR COMPUTER-COMMUNICATIONS NETWORKS

by

José H. Barberá

This report is based on the unaltered thesis of José H. Barberá, submitted in partial fulfillment of the requirements for the degree of Master of Science at the Massachusetts Institute of Technology in June, 1976. The research was conducted at the Decision and Control Sciences Group of the M.I.T. Electronic Systems Laboratory with partial support provided by the Advanced Research Project Agency of the Department of Defense under Contract N00014-75-C-1183.

Electronic Systems Laboratory
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER	2. GOVT ACCESSION NO. / 3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) (6) Diffusion Models for Computer-Communication Networks.	5. TYPE OF REPORT & PERIOD COVERED (9) RESEARCH rept.,
	6. PERFORMING ORG. REPORT NUMBER (14) ESL-R-661
7. AUTHOR(s) (10) José H. Barberá	8. CONTRACT OR GRANT NUMBER(s) (15) N00014-75-C-1183, ARPA Order [redacted]-3045, 5-7-75 ONR N00014-75-C-1183
9. PERFORMING ORGANIZATION NAME AND ADDRESS Massachusetts Institute of Technology Electronic Systems Laboratory Cambridge, Mass. 02139	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Program Code No. 5T10 ONR Identifying No. 049-383
11. CONTROLLING OFFICE NAME AND ADDRESS Defense Advanced Research Projects Agency 1400 Wilson Boulevard Arlington, Virginia 22209	12. REPORT DATE (11) May 1976 13. NUMBER OF PAGES 128
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) Office of Naval Research Information System Program Code 437 Arlington, Virginia 22217 (12) 140p.	15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Computer-Communication Networks
Diffusion Model
Message Routing
Load Sharing

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The diffusion theory is used to model a computer-communication network in which messages flow from one computer center to another. The idea is to approximate the various queueing processes that occur in the system (of discrete nature themselves) as continuous-state processes. The messages waiting at the queues to be transmitted are considered of small duration so that in the limit the flow of messages is continuous.

With these ideas a general model for routing of messages in a computer network is established and expressions for the diffusion parameters (drift and

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

127 200

covariance per unit time) are derived in terms of the network traffic. The mean length of the queues can thus be calculated and procedures to minimize the system overall queue size may be applied.

Examples for simple networks are shown. One of them corresponds to a load-sharing computer system and it is indicated how the general diffusion methods derived earlier for message routing, can be used.

Finally, a comparison is made between the expressions obtained by diffusion techniques and those corresponding to the classical exponential M/M/1 queue.

DIFFUSION MODELS FOR COMPUTER-COMMUNICATIONS NETWORKS

by

José Heredia Barberá

Ingeniero de Telecomunicación

E.T.S.I.T., Universidad Politécnica de Madrid

(1971)

SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June, 1976

Signature of Author .....
Department of Electrical Engineering and Computer Science, May 7 1976

Certified by .....
Thesis Supervisor

Accepted by .....
Chairman, Departmental Committee on Graduate Students

II

DIFFUSION MODELS FOR

COMPUTER-COMMUNICATIONS NETWORKS

by

José Heredia Barberá

Submitted to the Department of Electrical Engineering and Computer Science on May 7, 1976 in partial fulfillment of the requirements for the Degree of Master of Science.

ABSTRACT

The diffusion theory is used to model a computer-communication network in which messages flow from one computer center to another. The idea is to approximate the various queueing processes that occur in the system (of discrete nature themselves) as continuous-state processes. The messages waiting at the queues to be transmitted are considered of small duration so that in the limit the flow of messages is continuous.

With these ideas a general model for routing of messages in a computer network is established and expressions for the diffusion parameters (drift and covariance per unit time) are derived in terms of the network traffic. The mean length of the queues can thus be calculated and procedures to minimize the system overall queue size may be applied.

Examples for simple networks are shown. One of them corresponds to a load-sharing computer system and it is indicated how the general diffusion methods derived earlier for message routing, can be used.

Finally, a comparison is made between the expressions obtained by diffusion techniques and those corresponding to the classical exponential M/M/1 queue.

THESIS SUPERVISOR: Adrian Segall
TITLE: Assistant Professor of Electrical Engineering and Computer Science

Acknowledgements:

I would like to thank my thesis supervisor, Professor Adrian Segall, for suggesting the topic and for his guidance along the course of this thesis.

This work was supported by the Fundación del Instituto Tecnológico de Postgraduados of Spain and by the Advanced Research Project Agency of the Department of Defense (monitored by ONR) under Contract number N00014-75-C-1183.

TABLE OF CONTENTS

LIST OF FIGURES

1.- INTRODUCTION

1.1.- General considerations

A computer-communication system consists of several computers interconnected by communication channels. It is usually referred as a network in which the nodes represent the computers and the links represent the interconnecting channels. Messages are originated at some node and have to reach some other destination node according to some routing strategy which will try to use the network in an efficient way.

The computer network considered here is assumed to operate in the "store and forward" mode: a message generated at a computer center will be directed into the appropiate outgoing channel according to the routing policy and will be transmitted over this channel if it is free for transmission. If the channel is busy, the message will be stored at the node in some buffer joining other possible waiting messages. When the channel becomes free one of the wating messages is transmitted through the channel according to some queueing priority basis. This will be assumed "first-come, first-served" (FCFS) as it is usually referred in queueing literature. [7, 12, 22]

The queue of messages at each node constitutes a queueing process of discrete nature $n_i(t)$ such that $n_i(t) = A_i(t) - D_i(t)$ where $A_i(t)$ and $D_i(t)$ represent respectively the arrival and departure processes at node i, namely the number of arrivals and departures at the queue i up to the time t.

In order to provide mathematical tractability a model for the network of queues has to be established.

The type of queue depends on the statistics of the interarrival and service times. The simplest type of queues is the M/M/1 queue (*). This means that the interarrival and service times are independent and obey an exponential distribution or equivalently that the arrival and service rates follow a Poisson distribution. Because of the Poisson property (see [17] for example) the expressions of the system dynamics are easy to obtain and the steady-state distribution of the queue length p_n is quite straightforward [7] :

$$p_n = (1 - \rho)\,\rho^n \quad ; \quad n = 0, 1, 2, \ldots \qquad (1\text{-}1a)$$

$$\rho = \lambda/\mu \quad < \quad 1 \qquad (1\text{-}1b)$$

where λ and μ are respectively the arrival and service constant rates expressed in messages/unit time.

The condition $\lambda < \mu$ is necessary to assure that the steady-state is reached and the process does not blow up.

The expression (1-1) allows to calculate the average queue length

---

(*) In queue literature it is usual to denominate a queue by the symbols A/B/X/Y/Z where A indicates the interarrival-time distribution B the service-time distribution, X the number of parallel servers Y the restriction on the queue length capacity and Z the queue discipline. Often the last two symbols are omitted and it is understood that $Y = \infty$, that is no restrictions on the maximun queue length and Z = FCFS ("first come first served"). The symbols used for A and B are: D for deterministic, M for exponential, E_k for erlagian type k, G for general and GI for general independent. (Reference [7])

$\bar{n} = \sum_{n} n\, p_n$. From the same expression, the waiting time distribution (including also the time spent in the service, i.e. transmission through the channel) can be obtained and from this the average waiting time can be drawn. An alternate way is using Little's formula [15] which states that the average number of customers in a queueing system is equal to the average arrival rate of customers to that system, times the average time spent in the system:

$$\bar{n} = E[T] \qquad (1\text{-}2)$$

so that $E[T]$ can be calculated yielding

$$E[T] = \frac{1}{\mu - \lambda} \qquad (1\text{-}3)$$

In the computer network $E[T]$ is the average delay a message suffers going from one node to another and includes the average waiting time at the entering node plus the average transmission time in the channel.

When a network of queues is considered, the messages arriving at a new node along their path lose the independence property mentioned above because of the strong dependency between interarrival times and lengths of adjacents messages. For example if a message at node i has a length of s_1 seconds and arrives at node j at time t_1, it is clear that during the interval $(t_1, t_1 + s_1)$ no messages can arrive at j from i since they are transmitted in a sequential order, and therefore the independence assumption is no longer valid. This makes the analysis very complicated from a mathematical point of view. Kleinrock [11] overcomes this difficulty by introducing the "independence assumption" which specifies that each time the messages enter a new node they are assigned new independent lengths

(exponentially distributed).

With this assumption an expression for the average delay over the entire network can be found in terms of the average delay in each link. The desired routing strategy is that which minimizes the average delay and procedures to obtain the optimal strategy have been derived for example in [1].

1.2.- Existing Models for Networks of Queues

One of the earliest models was established by Jackson [8]. He considered an arbitrary network of N nodes each of them consisting of r_i servers with constant exponential mean service time μ_i. Messages arrive at each node i from outside the network according to an homogeneous Poisson process of rate λ_i. Each message upon being served, is directed to some other node j according to some probability θ_{ij} or leaves the system with probability $1 - \sum_j \theta_{ij}$. The transition probabilities θ_{ij} are assumed corresponding to a 1st order Markov chain. The total arrival rate at each node i Γ_i consists of the sum of the arrival rate from outside the network (Poisson) and the arrival rates from other nodes within the network:

$$\Gamma_i = \lambda_i + \sum_{j=1}^{N} \Gamma_j \, \theta_{ji} \qquad (1\text{-}4)$$

Jackson showed that when $\Gamma_i < r_i \mu_i$ for all i as far as the steady state is concerned, the network behaves as if all nodes were independent Poisson processes $M/M/r_i$ with rate Γ_i. Therefore the steady-state joint

distribution can be expressed as the product of the corresponding marginal distributions, and expressions for the queue lengths and average delays can be easily found.

In a subsequent work [9] Jackson considered a more general network in which the arrival process still being Poisson, is allowed to have a rate dependent on the total number of customers in the network. Each node has r_i servers and the service time is exponentially distributed with mean dependent on the number of customers at that node. Still the joint distribution factors into the product of the marginal ones and each node can be treated independently.

A modification of the Jackson's model was considered by Gordon and Newell [6]. They consider a closed system of queues which handles a finite and fixed number of customers. This model can be made equivalent to that of Jackson by assuming $\lambda_i = 0$ for all i and $\sum_{j=1}^{N} \theta_{ij} = 1$.

More general models that allow a service time discipline not necesarily exponential, have been considered and explicit solutions have been obtained. [20].

In all cases the main difficulty comes from the fact that there is a very large system of equations due to the enormous number of states.

In order to overcome these difficulties and break away from the sometimes too simplistic models that assume exponential service time, different approaches to network analysis have been made. Thus for example Kingman [10] has shown in his treatement of heavy traffic theory that properties of nearly saturated queues are rather insensitive to the specific form of arrival or service distribution.

The heavy traffic approximation is based upon the central limit theorem. This leads to the idea of approximating a discrete-state processes by continuous-state ones which have been called diffusion processes and will be explained in a subsequent section.

The idea of approximating a discrete-state process by a diffusion one is not new. (See for example Feller [3]). Nonetheless is rather recent. Thus for example, Newell [16] gives an extensive treatement of queues with time dependent arrival rate by using the diffusion approximation. Gaver [4] applies diffusion approximation techniques to the waiting time in a M/G/1 queue. He shows that the waiting time is exponential in the diffusion approximation provided the system was initially empty. An asymptotic approximation is supplied for the mean waiting time near saturation and comparisons are made with the exact solutions provided by the classical methods (see [7] for example). The results show to be rather accurate for those conditions.

Gaver and Shedler [5] have applied the diffusion approximation to evaluate the CPU utilization of a multiprogrammed system represented by a cyclic queueing model. Solutions appear to be easier than the classical ones and yet the accuracy seems quite adequate for the case studied.

Kobayashi [13, 14] has analyzed a system of queues by diffusion methods. His model is based on the Markovian model of Jackson (open networks) and Gordon and Newell (closed networks) which we mentioned earlier. In the first paper [13] and based upon central limit theorem arguments he finds the steady-state distribution of a single queue, and a system of queues (open and closed) assuming independent identically distributed interarrival

and service times with general distributions. In the second paper [14] and using diffusion methods too, the transient behavior of those systems is analyzed. The analysis provides an estimation of the transient period which shows to be shorter as the system is less congested. A comparison of results with those exactly known by classical methods is given in [19] and they show to be rather accurate for utilization factors near 1.

1.3.- Objetives of this thesis

As it was pointed out in the preceeding section when the number of states of a Markov model becomes very large, although finite, the search of solutions appears quite cumbersome. The procedure of approximating the discrete-state process by a diffusion process can be therefore useful because mathematical methods associated with a continuous space are very often more easily treated than those in a discrete-state space. In a computer-communication network this is the case when the number of computer centers is relatively large.

The purpose of this thesis is to consider a general type of computer network and by using diffusion methods find a model for analysis of the behavior of the network. Then a strategy for routing messages throughout the system in an efficient way is to be found. In order to make an optimal use of the network, messages shall reach their destination as soon as possible and thus the performance criteria for routing will be to minimize the overall average delay on the entire network.

The ideas for the model set up (Section 4.1) resemble some those

established by Segall [23] and have been taken from this reference. The mentioned paper deals with dynamic routing in computer networks and avoids the "independence assumption" that was mentioned earlier although the model of [23] assumes a deterministic scheme with known traffic inputs whereas here the model is stochastic in the sense that the inputs are only known in terms of their statistics.

2.- THE DIFFUSION PROCESS

2.1.- The random walk process [2]

It is introduced here the concept of random walk as a discrete-state discrete-time Markov process for the diffusion process can be drawn from it in the limit. Consider the time divided in intervals of duration $\Delta t : 0, \Delta t, 2\Delta t, \ldots n\Delta t \ldots$ and the state divided in intervals of length $\theta : 0, \theta, 2\theta, \ldots k\theta \ldots$. At time $t = 0$ the state is $x_0 = k_0\theta$. At time $t = \Delta t$ the state can jump one step θ upwards with probability p, one step downwards with probability q or remain the same with probability $1 - p - q$. No other transitions are allowed. In each interval of time later the same jumps with the same probabilities can happen and are independent of the previous jumps. This is graphically shown in Fig.2.1 and can be considered as the motion of a particle in a one-dimensional space. If the particle continues to move indefinitely the random-walk is said to be unrestricted. Nonetheless it is frequently necessary to have the motion restricted in some way, usually by the presence of "barriers". For example a random walk starting at the origin can be restricted to move between an upper barrier $a > 0$ and a lower barrier at $b < 0$. Several types of barriers are encountered. One of these is an "absorbing barrier": when it is reached the particle stays there and the motion ceases. Another type is a "reflecting barrier" defined as a state that when crossed in a given direction, say downwards, holds the particle

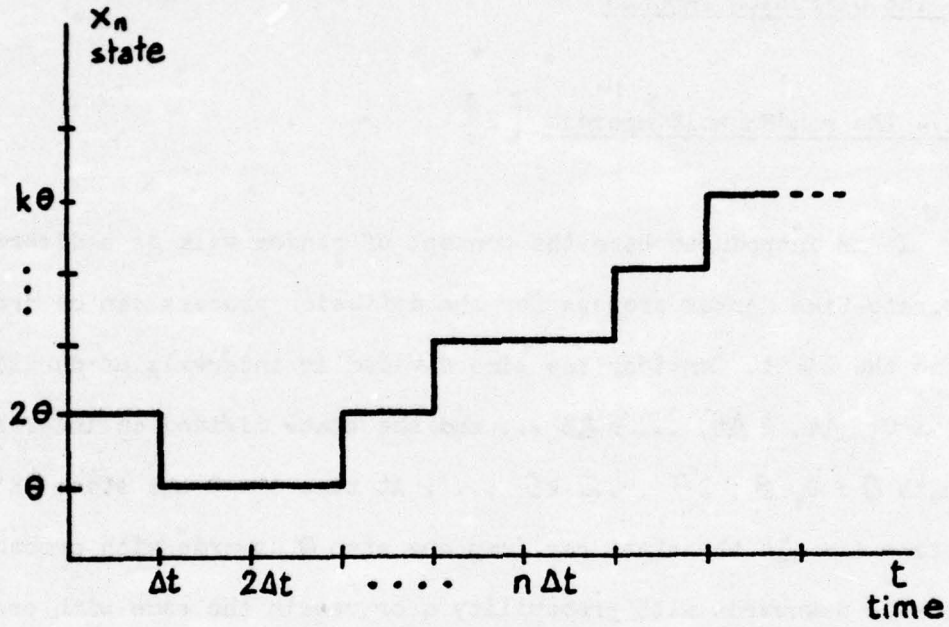


Fig. 2.1 A random walk

until a positive jump occurs and brings the particle out of the barrier resuming the motion.

We shall examine here the properties of a random walk with reflecting barriers. It is of interest to determine the steady-state or equilibrium distribution of the state as t goes to infinity.

Clearly the dynamics of the random walk exposed above are governed by the following equation:

$$P_{k\theta}(n\,\Delta t + \Delta t) = P_{k\theta}(n\,\Delta t)(1 - p - q) + P_{(k-1)\theta}(n\,\Delta t)\,p + P_{(k-1)\theta}(n\,\Delta t)\,q \qquad (2-1)$$

where:

$$P_{k\theta}(n\,\Delta t) = \left\{ \begin{array}{l} \text{Prob. of being at state } k\theta \text{ at time } n\Delta t \text{ given that the} \\ \text{initial state was } k_0\theta \end{array} \right\}$$

2.2.- The diffusion process as a limit of a random walk

Consider the random walk of section 2.1. Assume that θ and Δt go to zero so that $n\,\Delta t \rightarrow t$ and $k\theta \rightarrow x(t)$. The resultant process $x(t)$ becomes a continuous-sate continuous-time Markov process. It is shown in Fig. 2.2

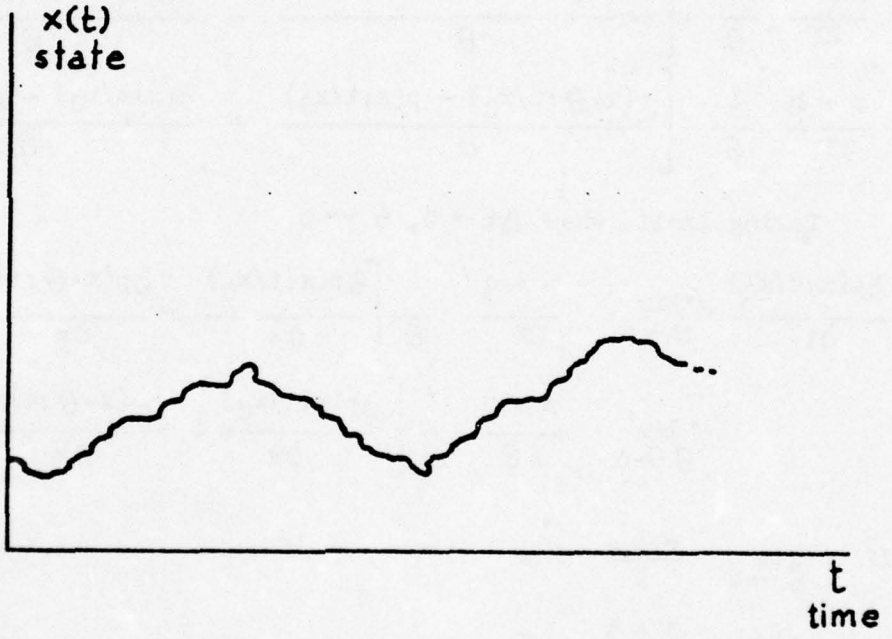


Fig 2.2: Diffusion process as a limit of a random walk

Equation (2-1) can be rearranged as:

$$P_{k\theta}(n\,\Delta t + \Delta t) - P_{k\theta}(n\,\Delta t) = \frac{p+q}{2}\left\{\left[P_{(k+1)\theta}(n\,\Delta t) - P_{k\theta}(n\,\Delta t)\right] - \left[P_{k\theta}(n\,\Delta t) - P_{(k-1)\theta}(n\,\Delta t)\right]\right\} + \frac{q-p}{2}\left\{\left[P_{(k+1)\theta}(n\,\Delta t) -\right.\right.$$

$$- p_{k\theta}(n\Delta t)\Big] + \Big[p_{k\theta}(n\Delta t) - p_{(k-1)\theta}(n\Delta t)\Big]\Big\} \qquad (2\text{-}2)$$

When Δt becomes very small, so that $n\Delta t \to t$, $k\theta \longrightarrow x(t)$ and $k_0\theta \to x_0$ the state probability: $p_{k\theta}(n\Delta t) \longrightarrow p(x;t/x_0)$.

Let $\Delta t = K\theta^2$ (K being a constant) and divide both sides of (2-2) by Δt. Then

$$\frac{p(x;t+\Delta t/x_0) - p(x;t/x_0)}{\Delta t} =$$

$$= \frac{p+q}{2K}\,\frac{1}{\theta}\cdot\left[\frac{p(x+\theta;t/x_0) - p(x;t/x_0)}{\theta} - \frac{p(x;t/x_0) - p(x-\theta;t/x_0)}{\theta}\right] -$$

$$- \frac{p-q}{2K}\,\frac{1}{\theta}\cdot\left[\frac{p(x+\theta;t/x_0) - p(x;t/x_0)}{\theta} + \frac{p(x;t/x_0) - p(x-\theta;t/x_0)}{\theta}\right]$$

Taking limits when $\Delta t \to 0$, $\theta \to 0$

$$\frac{\partial p(x;t/x_0)}{\partial t} = \lim_{\theta\to 0}\ \frac{p+q}{2K}\,\frac{1}{\theta}\left[\frac{\partial p(x;t/x_0)}{\partial x} - \frac{\partial p(x-\theta;t/x_0)}{\partial x}\right] -$$

$$- \lim_{\theta\to 0}\ \frac{p-q}{K\theta}\,\frac{1}{2}\left[\frac{\partial p(x;t/x_0)}{\partial x} + \frac{\partial p(x-\theta;t/x_0)}{\partial x}\right] \qquad (2\text{-}4)$$

If $\lim_{\theta\to 0}\ \dfrac{p+q}{K} = \alpha$ (2-4a)

and $\lim_{\theta\to 0}\ \dfrac{p-q}{K\theta} = \beta$ (2-4b)

α and β being constants then (2-3) becomes

$$\frac{\partial p(x;t/x_0)}{\partial t} = \frac{1}{2}\alpha\,\frac{\partial^2 p(x;t/x_0)}{\partial x^2} - \beta\,\frac{\partial p(x;t/x_0)}{\partial x} \qquad (2\text{-}5)$$

which is the diffusion equation [2].

For the conditions (2-4) to be satisfied, the probabilities p and q must be taken as:

$$p = K\ \frac{1}{2}\ (\alpha + \beta\theta) \quad = \frac{K}{2}\ (\alpha + \frac{\beta}{\sqrt{K}}\ \sqrt{\Delta t}) \qquad (2\text{-}6)$$

$$q = K\ \frac{1}{2}\ (\alpha - \beta\theta) \quad = \frac{K}{2}\ (\alpha - \frac{\beta}{\sqrt{K}}\ \sqrt{\Delta t}) \qquad (2\text{-}7)$$

that is the probabilities of jumping upwards and downwards have to be nearly the same, the difference being a term that tends to zero as $\sqrt{\Delta t}$.

Notice that the parameters β and α are respectively the incremental mean and variance of the process $x(t)$ per unit time since

$$E\left[x(t + \Delta t) - x(t)/x(t)\right] = \theta\cdot(p - q) \longrightarrow (K\theta^2) = \beta\cdot\Delta t$$

$$\mathrm{Var}\left[x(t + \Delta t) - x(t)/x(t)\right] = \theta^2(p + q) - \theta^2(p - q)^2 \longrightarrow$$

$$\longrightarrow \theta^2 K(\alpha - \beta^2\theta^2) \approx \alpha\cdot\Delta t$$

that is

$$\beta = \lim_{\Delta t \to 0} \frac{E\left[x(t + \Delta t) - x(t)/x(t)\right]}{\Delta t} = \lim_{\Delta t \to 0} \frac{E\left[\Delta x(t)/x(t)\right]}{\Delta t} \qquad (2\text{-}8)$$

$$\alpha = \lim_{\Delta t \to 0} \frac{\mathrm{Var}\left[x(t + \Delta t) - x(t)/x(t)\right]}{\Delta t} = \lim_{\Delta t \to 0} \frac{\mathrm{Var}\left[\Delta x(t)/x(t)\right]}{\Delta t} \qquad (2\text{-}9)$$

In general the parameters α and β can be dependent on the state $x(t)$.

We can consider as well a multidimensional diffusion process $\underline{x}(t)$ with vector mean per unit time $\underline{\beta}$ and covariance matrix per unit time $\underline{\Lambda} = [\alpha_{ij}]$. In this case the diffusion equation relates the derivatives of the multidimensional p.d.f. $p(\underline{x}; t/\underline{x}_0)$ and the expression (2-7) becomes

$$\frac{\partial p(\underline{x}; t/\underline{x}_0)}{\partial t} = \sum_{i=1}^{m} \sum_{j=1}^{m} \frac{1}{2}\ \alpha_{ij} \frac{\partial^2 p(\underline{x}; t/\underline{x}_0)}{\partial x_i\ \partial x_j} - \sum_{i=1}^{m} \beta_i \frac{\partial p(\underline{x}; t/\underline{x}_0)}{\partial x_i} \qquad (2\text{-}10)$$

where m is the dimension of the process and

$$\beta_i = \lim_{\Delta t \to 0} \frac{E\left[\Delta x_i(t) / \underline{x}(t)\right]}{\Delta t} \qquad (2\text{-}11$$

$$\alpha_{ij} = \lim_{\Delta t \to 0} \frac{\mathrm{Cov}\left[\Delta x_i(t) \;\; \Delta x_j(t) / \underline{x}(t)\right]}{\Delta t} \qquad (2\text{-}12)$$

$$i = 1, 2, \ldots m$$

The probability of each process i ($i = 1,2,\ldots m$) jumping upwards p_i and downwards q_i have to satisfy now the conditions (2-6) and (2-7) for each process $x_i(t)$ to be considered as diffusion, and similarly for each individual process it must be $\Delta t = K_i \;\; \theta_i^{\,2}$ where the constant K_i are in principle different.

In the computer-network system we are interested the state of each process represents the total number of messages that are waiting to be transmitted at an specific queue. For each queue i there is a lower reflecting barrier at $x_i = 0$ because the number of messages in a queue cannot be negative. If there is no restriction on the queue length, then there is no upper barrier. Practically the size of the queues are limited by the length of the buffers where the messages are stored. Therefore an upper barrier is to be assumed too which if reached indicates that the buffer is full.

3.- SOLUTION OF THE DIFFUSION EQUATION

The solution of (2-5) or (2-10) depends upon the conditions imposed on $\underline{x}(t)$. If $\underline{x}(t)$ is allowed to take any value: $-\underline{\infty} < \underline{x}(t) < +\underline{\infty}$ the joint p.d.f. ($p(\underline{x}\,; t/\underline{x}_0)$ solution of the diffusion equation is the corresponding to a multidimensional Wiener process with mean $\underline{\beta}t$ and covariance matrix $\underline{\Lambda}t$:

$$p(\underline{x}\,; t/\underline{x}_0) = \frac{\exp\left[-\frac{1}{2}(\underline{x} - \underline{x}_0 - \underline{\beta}t)^T (\underline{\Lambda}t)^{-1} (\underline{x} - \underline{x}_0 - \underline{\beta}t)\right]}{(2\pi)^{m/2} \left|\underline{\Lambda}t\right|^{1/2}} \qquad (3\text{-}1)$$

(Taking in (3-1) the derivates $\partial/\partial x_i$, $\partial^2/\partial x_i\,\partial x_j$ and $\partial/\partial t$, it can be easily seen that (3-1) satisfies expression (2-10).

Observe that (3-1) has no steady-state solution when $t \to \infty$.

If one reflecting barrier is considered, say at $x = 0$, the solution of the scalar diffusion equation (2-5) with initial condition $p(x;0/x_0) = \delta(x - x_0)$ can be found by using the method of images [24] and it is [2]:

$$p(x\,; t/x_0) = \frac{1}{\sqrt{2\pi\alpha t}} \left\{ \exp\left[-\frac{(x - x_0 - \beta t)^2}{2\alpha t}\right] + \exp\left(-\frac{2|\beta|}{\alpha} x_0\right) \cdot \right.$$

$$\left. \cdot \exp\left[-\frac{(x + x_0 - \beta t)^2}{2\alpha t}\right] \right\} + \frac{2|\beta|}{\alpha} \exp\left(-\frac{2|\beta|}{\alpha} x\right) \cdot$$

$$\cdot \Phi\left(\frac{x + x_0 + \beta t}{\sqrt{\alpha t}}\right) \qquad (3\text{-}2)$$

where

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_z^{\infty} e^{-u^2/2}\, du$$

The first term of (3-2) corresponds to the transient period and the second one is the steady-state term. By inspection of (3-2) it is easy to see that when $t \longrightarrow \infty$ the first term of (3-2) vanishes and the second one becomes

$$\lim_{t \to \infty} p(x;\, t/x_0) = p(x) = \begin{cases} 0 \quad ; & \beta > 0 \\ \\ \dfrac{2|\beta|}{\alpha} \exp\left(-\dfrac{2|\beta|}{\alpha} x\right) & ; \beta < 0 \end{cases} \qquad (3\text{-}3)$$

$$\text{for } x \geq 0$$

that is an steady-state solution exist for negative drift β which physically means that the service process has to be faster than the arrival process so that the length of the queue does not become infinite.

If two barriers are considered at $x = 0$ and $x = a > 0$ for example then the equation (2-5) for the scalar diffusion with initial condition $p(x;\, 0/x_0) = \delta(x - x_0)$ and boundaries $0 \leq x(t) \leq a$, can be solved by the separation of variables method and the solution is: [25]

$$p(x;t/x_0) = \frac{\frac{\beta}{2\alpha} \exp\left|\frac{\beta}{2\alpha} x\right|}{\exp\left|\frac{\beta}{2\alpha} a\right| - 1} + \exp\left(-\frac{\beta^2}{8\alpha} t\right) \exp\left(\frac{\beta}{\alpha}(x - x_0)\right) \cdot$$

$$\cdot \frac{4}{a} \sum_{n=1}^{\infty} \exp\left(-\lambda_n^2 \frac{\alpha}{2} t\right) \frac{\lambda_n^2}{\lambda_n^2 + (\beta/\alpha)^2} Y_n(x)\, Y_n(x_0) \qquad (3\text{-}4)$$

$$\text{for } 0 \leq x \leq a \; ; \quad t > 0$$

and $\lambda_n = n \frac{\pi}{a}$; $n = 0, \pm 1, \pm 2, \ldots$

$$Y_n(x) = \cos \lambda_n x + \frac{\beta/\alpha}{\lambda_n} \sin \lambda_n x$$

Regardless of the sign of β there is a steady-state solution when $t \to \infty$. This distribution denoted by

$$p(x) = \lim_{t \to \infty} p(x; t/x_0)$$

can be obtained by taking limits in (3-4)

$$p(x) = \frac{\frac{\beta}{2\alpha} \exp\left(\frac{\beta}{2\alpha} x\right)}{\exp\left(\frac{\beta}{2\alpha} a\right) - 1} \quad ; \quad 0 \leqslant x \leqslant a \; ; \qquad (3\text{-}5)$$

For the multidimensional process, define a vector

$$\underline{\gamma} = 2 \underline{\Lambda}^{-1} \underline{\beta} \qquad (3\text{-}6)$$

and then the steady-state distribution of the process $\underline{x}(t)$ can be obtained by equating to zero the right hand side of (2-10).

This gives [14]

$$p(\underline{x}) \equiv \lim_{t \to \infty} p(\underline{x}; t/\underline{x}_0) = \prod_{i=1}^{m} p_i(x_i) \qquad (3\text{-}7a)$$

where

$$p_i(x) = \frac{\gamma_i e^{\gamma_i x}}{e^{\gamma_i a_i} - 1} \quad ; \; 0 \leqslant x_i \leqslant a_i \qquad (3\text{-}7b)$$

4.- DIFFUSION MODEL, FOR MESSAGE ROUTING IN A COMPUTER NETWORK

4.1.- The general Model

Let us consider a general network of N computer centers (nodes). The same notation as in [23] will be followed. Each node can be connected to any of the remaining (N-1) nodes in both directions. We have then a possible total number of communication lines N(N-1).

The nodes will be represented by letters i (i=1,2, ...,N) and the branches by pairs (i,j) (i,j = 1,2, ... N ; $i \neq j$). Call

I(i) the set of branches entering node i

D(i) the set of branches coming out from node i

At each node there can be a maximun of (N-1) queues where messages with destination any of the other remaining nodes wait to be transmitted. Clearly the total number of queues in the whole system M is such that $M \leqslant N(N-1)$.

The queueing processes are of discrete nature themselves as was pointed out in a preceding section. The diffusion model that is stablished here makes the approximation of considering them as continuous processes. In order to do this, the messages (that in principle have different lengths) are assumed to be divided in small packets of duration Δt units of time.

The time will be assumed divided in intervals $(t, t + \Delta t)$ small enough so that the following events will occur:

$a_{ij}(t)$ = Prob. that a message of length Δt with final destination node j arrives at node i from outside the network.

$1 - a_{ij}(t)$ = Prob that no message of length Δt with final destination j arrives at node i from outside the network

$$i,j = 1, 2, \ldots N \quad ; i \neq j$$

Therefore during $(t, t + \Delta t)$ only these former events can occur. The probability that more than one message comes is zero.

$u_{ij}^{k}(t)$ = Prob. that a message of length Δt with final destination k is transmitted from node i to node j.

$1 - u_{ij}^{k}(t)$ = Prob. that no message of length Δt with final destination k is transmitted from node i to node j

$$i,j = 1,2, \ldots N \quad ; \quad i \neq j$$

c_{ij} = Prob. that no message of length Δt is transmitted succesfully from node i to node j.

$$i,j = 1,2, \ldots N \quad ; \quad i \neq j$$

Observe that c_{ij} corresponds to the capacity of the link (i,j) expressed in terms of probabilities rather than in traffic units.

In Fig 4.1 a diagram of such a general network is depicted.

The capacities c_{ij} are fixed for each channel (i,j). The incoming traffic probabilities $a_{ij}(t)$ are quantities that depend on the amount of users' demand at the specific time t. We shall consider this demand rate to be stationary so that it will not be dependent on time but a constant a_{ij}.

The outcoming traffic probabilities $u_{ij}^{k}(t)$ are the quantities we want to find according to the input traffic and the network fixed capacities so that the system performance is satisfied, according to some criteria as we shall see later. For the same reason as $a_{ij}(t)$, the probabilities $u_{ij}^{k}(t)$ will be independent of time.

From the previous definitions, notice that each channel (i,j) can

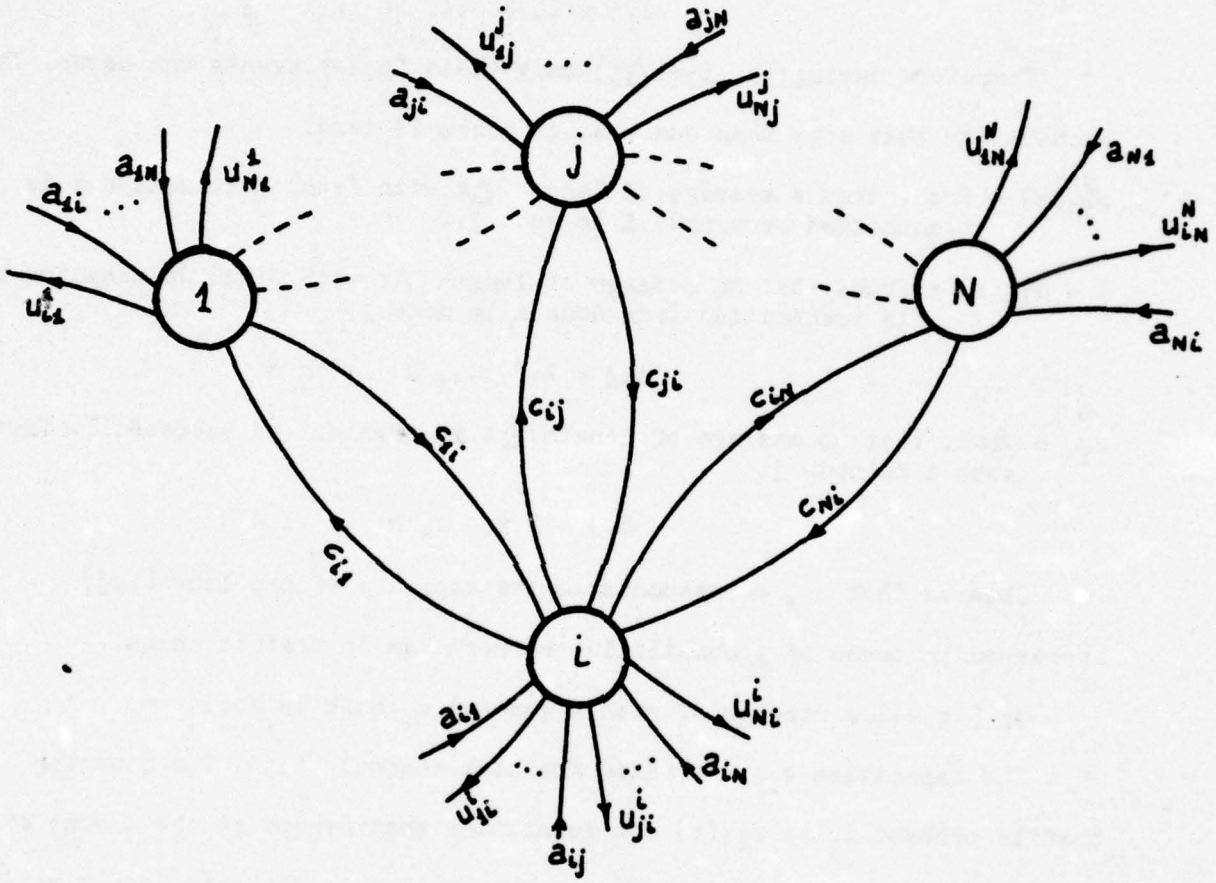


Fig 4.1: General configuration of a computer-communication network.

handle messages with different final destination. Thus the total traffic the channel (i,j) carries is composed of all the messages going during $(t, t + \Delta t)$ from node i to node j whatever the final destination is. In order for the transmission to be succesful this traffic has to be less than or equal to the capacity c_{ij}, that is:

$$\sum_{k \neq i} u^k_{ij} \leqslant c_{ij} \leqslant 1 \qquad (4\text{-}1a)$$

and clearly $$u_{ij}^{k} \geq 0 \quad ; \forall (i,j) \qquad (4\text{-}1b)$$

The constraint (4-1a) is necessary to have an errorless transmission.

Now consider the queueing process representing the number of messages $n_{ij}(t)$ that are waiting at node i to be transmitted to node j at the time t. If finite capacities are assumed for the buffers, that is N_{ij} is the maximun number of messages that can be stored at i to be transmitted to j, then the ratio

$$x_{ij}(t) = \frac{n_{ij}(t)}{N_{ij}} \qquad (4\text{-}2)$$

represents the normalized queueing process so that

$$0 \leqslant x_{ij}(t) \leqslant 1 \quad ; \qquad i,j = 1,2, \ldots N \; ; \quad i \neq j \qquad (4\text{-}3)$$

When the message lengths Δt become very small it is clear that $x_{ij}(t)$, representing the amount of messages filling the buffer with respect to its total capacity, becomes of continuous nature and can be approximated by a diffusion process with two reflecting barriers at x = 0 and x = 1 as indicated in (4-3). The lower barrier represents the queue completely empty and the upper one representing the queue full.

At each node we can have at most N - 1 queues and in the whole system the maximun number of queues is N(N-1). This can be visualized in Fig 4.2 for N = 3 in which node 1 has been magnified to indicate all the queueing processes that take place.

The switch S_{12} represents that the channel c_{12} can handle messages from $x_{12}(t)$ with probability u_{12}^{2} and messages from $x_{13}(t)$ with probability u_{12}^{3} such that $u_{12}^{2} + u_{12}^{3} \leqslant c_{12}$.

Similarly for the other switches: messages travel from node 2 to node

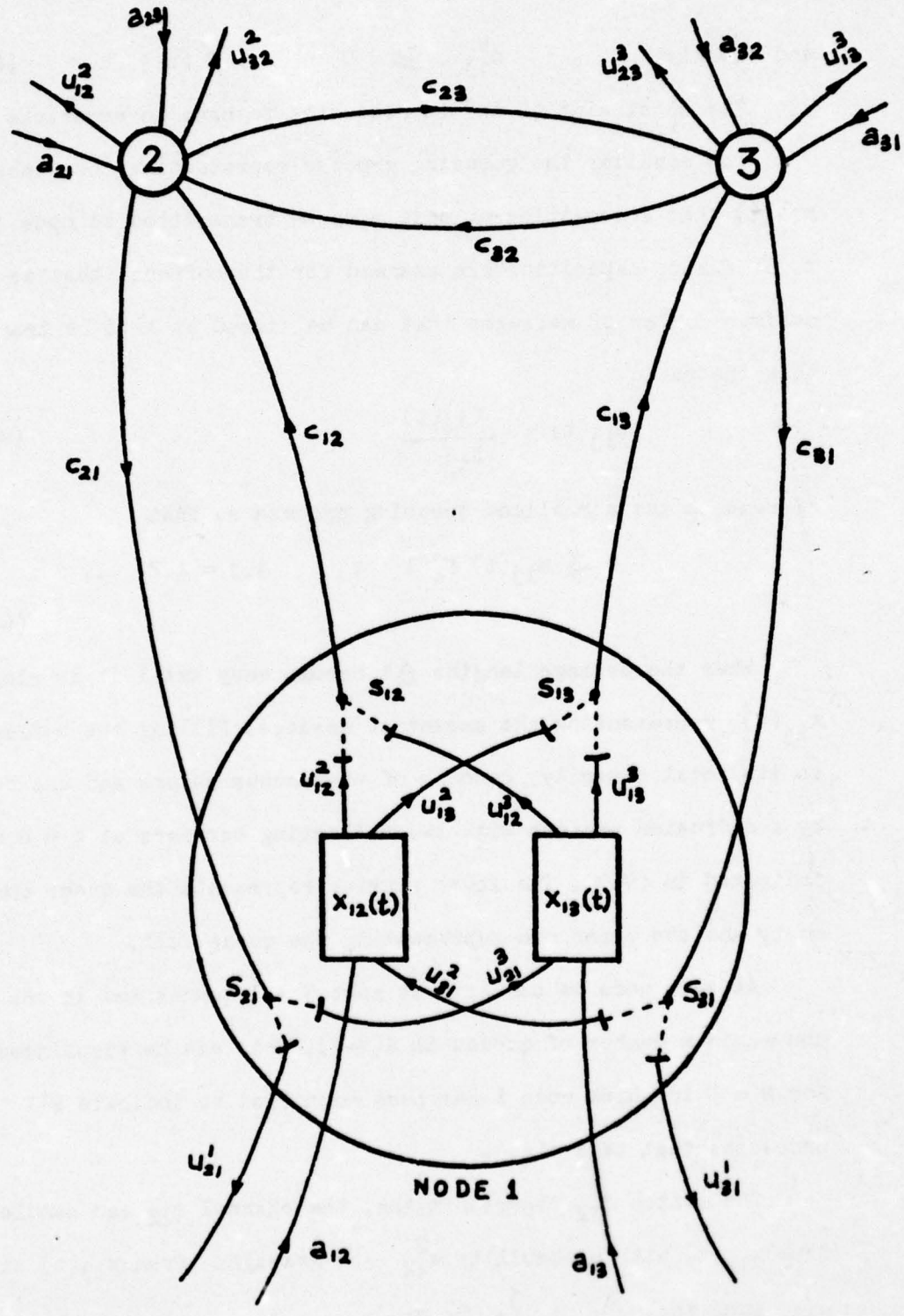


Fig. 4.2. Detail of the queueing process at one node

1 with probability u_{21}^{3} entering queue $x_{13}(t)$ and waiting to be transmitted to its final destination node 3.

4.2.- Diffusion approximation for the routing model

We saw in section 3 the steady-state solution of the M-dimensional diffusion equation (2-12).

The case we are dealing with is $M \leqslant N(N-1)$ and the barriers for all queues are 0 and 1.

The joint p.d.f. of the system is given by (3-7) with $a_i = 1$ for all i.

It was established in section 1.3 that the performance criteria will be to minimize the overall delay messages experience (on the average) when they are transmitted from their origin to their final destination.

An equivalent condition is to minimize the average queue size of the whole system. To see this, consider one queue $n_{ij}(t)$. If we call $A_{ij}(t)$ the number of arrivals at node i with destination j during (0,t) and $D_{ij}(t)$ the number of departures from this queue in (0,t) then

$$n_{ij}(t) = A_{ij}(t) - D_{ij}(t) \qquad (4-3)$$

represents the total number of messages waiting at that queue at time t. The quantity

$$\int_0^t n_{ij}(\tau)\, d\tau \qquad (4-4)$$

is the total time all messages have spent in that queue during (0,t).

The average delay per message at that queue will be:

$$E\left[T_{ij}\right] = \frac{\int_0^t n_{ij}(\tau)d\tau}{A_{ij}(t)} \qquad (4\text{-}5)$$

and the average number of messages waiting at the queue:

$$E\left[n_{ij}(t)\right] = \frac{\int_0^t n_{ij}(\tau)d\tau}{t} \qquad (4\text{-}6)$$

Therefore from (4-5) and (4-6) we see that minimizing the average delay is equivalent to minimizing the average queue length or normalizing acording to (4-2), minimizing $E\left[x_{ij}(t)\right]$.

In the discussion of section 3 about the diffusion process we implicitely assumed that there were no idle periods, that is when the process reaches the lower barrier it does not stay on there but jumps up. This has not to be the case in general because with some probability there will be certain idle periods in which the queue will be empty. This probability of idle periods can be expressed in terms of the utilization factor [12] which is defined as the ratio of the rate at which the jobs enter the system to the maximun rate (capacity) at which the system can perform this work. Calling this factor $\rho\,(<1)$ the probability of idle period will be $1-\rho$.

Therefore the p.d.f. (3-7b) shoud be modified to include the effect of idle periods and this can be taken into account by splitting the p.d.f. into two parts. One representing the probability of empty queue (an impulse of weight $1-\rho$) and the other representing the continuous distribution when the queue is not empty

$$p_i(x) = (1-\rho_i)\,\delta(x) + \rho_i \frac{\gamma_i e^{\gamma_i x}}{e^{\gamma_i}-1} \quad ; \quad 0 \leq x \leq 1 \qquad (4\text{-}7)$$

which is represented in Fig 4.3.

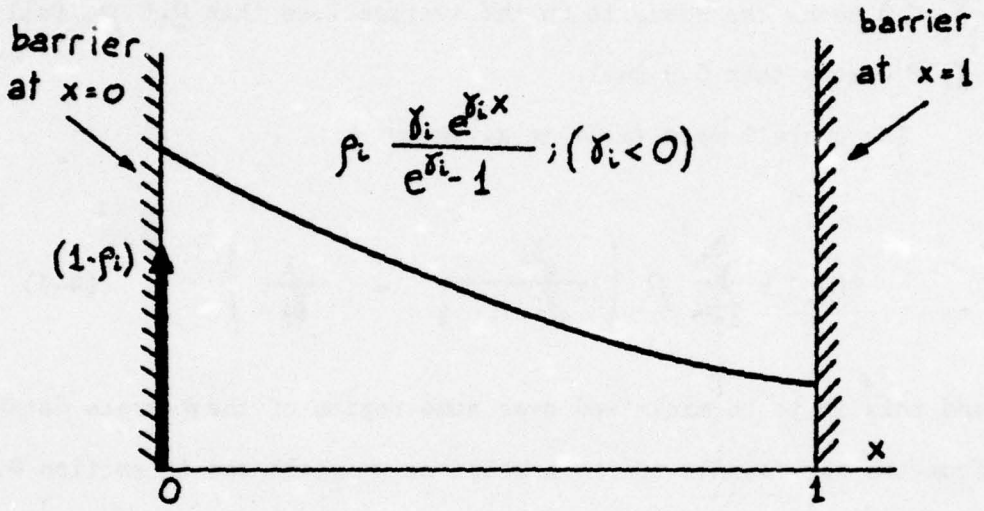


Fig. 4.3 Steady-State p.d.f. of a diffusion process

The expected queue length for the i-th queue is calculated from (4-7)

$$E(x_i) = \int_0^1 x_i \, p_i(x_i) \, dx_i = \rho_i \left(\frac{e^{\gamma_i}}{e^{\gamma_i} - 1} - \frac{1}{\gamma_i} \right) \qquad (4-8)$$

In Fig. 4.3 the expression (4-8) is represented. As it can be seen $0 < E(x_i) < \rho_i$. In the same figure it is represented the function $-1/\gamma_i$ ($\gamma_i < 0$) which would be the mean value in the case of no upper barrier. We can see that for values of γ_i less than -3 both curves are very close. The presence of the upper barrier prevents

the queue length from increasing without limit and therefore it does not become unbounded at $\gamma_i = 0$.

The significance of the parameters γ_i can be seen from Fig. 4.4: $\gamma_i < 0$ means the queue is on the average less than 0.5 ρ_i full and $\gamma_i > 0$ more than 0.5 full.

The overall mean value is given by

$$F(\underline{\gamma}) = \sum_{i=1}^{M} \rho_i \left(\frac{e^{\gamma_i}}{e^{\gamma_i} - 1} - \frac{1}{\gamma_i} \right) \qquad (4\text{-}9)$$

and this is to be minimized over some region of the M-space determined from the constraints of the problem as we shall see in section 4.5.

4.3.- Calculation of the diffusion parameters

We want to find the components of the vector mean per unit time $\underline{\beta}$ and the elements of the covariance matrix per unit time $\underline{\Lambda}$ defined in section 2 (Eqs. (2-11) and (2-12)).

Consistent with the notation in section 4.1 let us redefine $\underline{\beta}$ and $\underline{\Lambda}$ as:

$$\beta_{ij} = \lim_{\Delta t \to 0} \frac{E\left[x_{ij}(t + \Delta t) - x_{ij}(t) \,\middle|\, \underline{x}(t) \right]}{\Delta t} = \qquad (4\text{-}10)$$

$$= \lim_{\Delta t \to 0} \frac{E\left[\Delta x_{ij}(t) \,\middle|\, \underline{x}(t) \right]}{\Delta t}$$

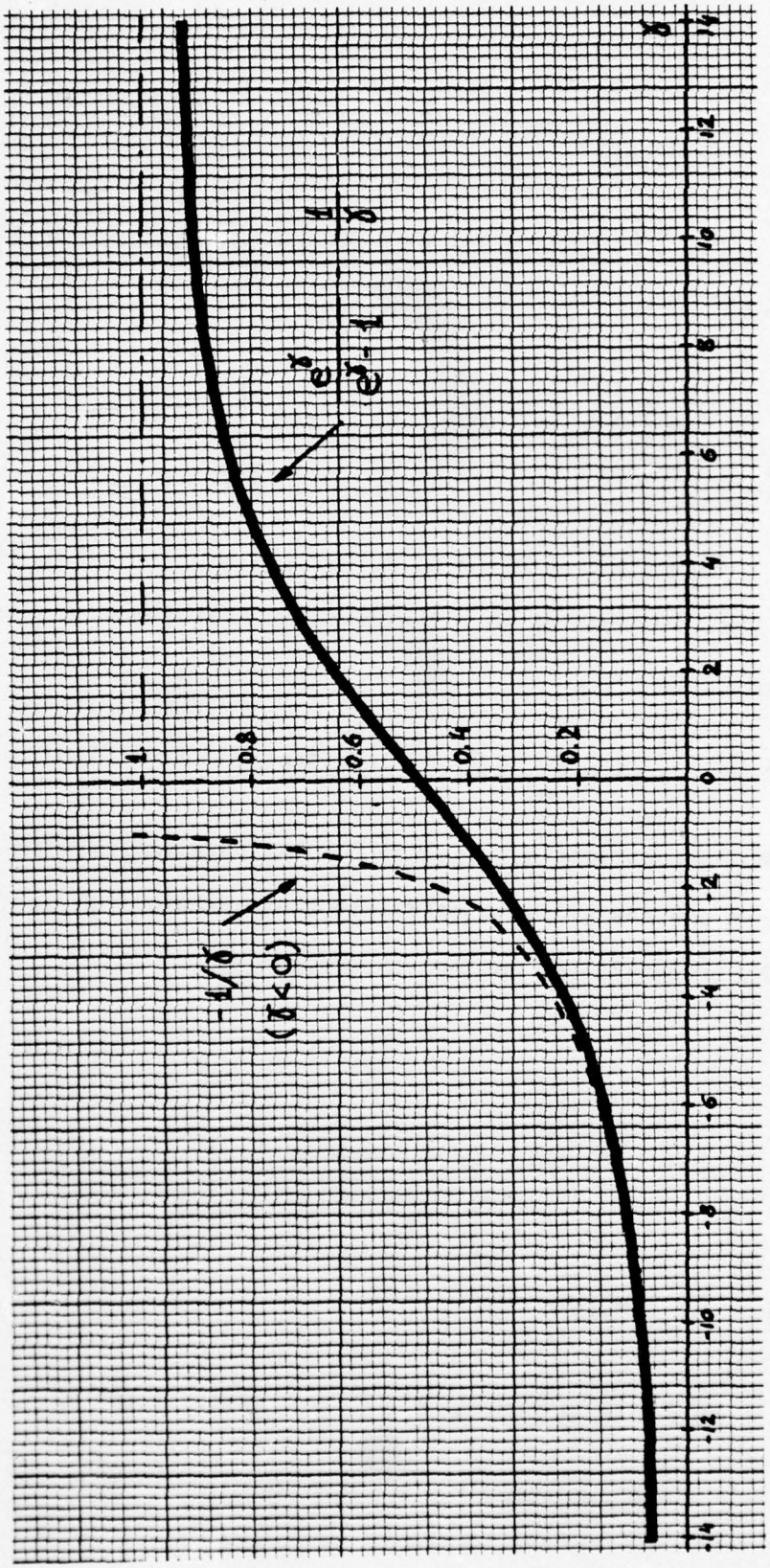


Fig.4.4. Average length of a diffusion queue in steady state

$$\alpha_{(ij),(kl)} =$$

$$= \lim_{\Delta t \to 0} \frac{\mathrm{Cov}\left\{ \left[x_{ij}(t+\Delta t) - x_{ij}(t) \right] \left[x_{kl}(t+\Delta t) - x_{kl}(t) \right] \;\middle|\; \underline{x}(t) \right\}}{\Delta t} =$$

$$= \lim_{\Delta t \to 0} \frac{\mathrm{Cov}\left[\Delta x_{ij}(t) \;\; \Delta x_{kl}(t) \;\middle|\; \underline{x}(t) \right]}{\Delta t} \qquad (4\text{-}11)$$

$$(i,j),\ (kl) = 1, 2, \ldots\ldots\ M \quad ; \quad M \leq N(N-1)$$

When Δt tends to zero the jumps at each queue θ_{ij} tend to zero too according to

$$\Delta t = K_{ij} \, (\theta_{ij})^2 \quad , \quad \forall (ij) \qquad (4\text{-}12)$$

where the constants K_{ij} account for the possible different buffer lengths.

We are now going to calculate the parameters β_{ij}, $\alpha_{(ij),(kl)}$ in terms of the probabilities defined in section 4.1.

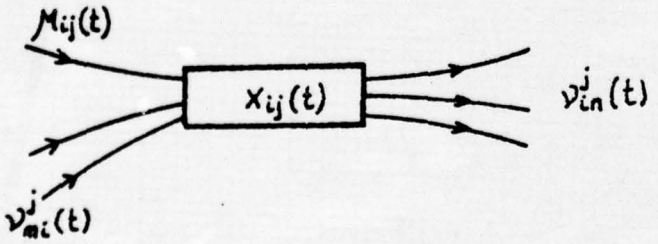


Fig 4.5: Arrival and departure processes in a queue

Consider the queue $x_{ij}(t)$ (Fig. 4.5)

then

$$\Delta x_{ij}(t) = \mu_{ij}(t) + \sum_{\substack{m \in I(i) \\ m \neq j}} \nu^j_{mi}(t) - \sum_{n \in D(i)} \nu^j_{in}(t) \qquad (4\text{-}13)$$

where for all t, $\mu_{ij}(t)$, $\nu^j_{mi}(t)$, $\nu^j_{in}(t)$ are Bernoulli independent random variables that can take the following values:

$$\mu_{ij}(t) = \theta_{ij} \quad \text{with probability } a_{ij} \qquad (4\text{-}14)$$
$$= 0 \quad \text{with probability } (1 - a_{ij})$$

$$\nu^j_{mi}(t) = \theta_{ij} \quad \text{with probability } u^j_{mi} \qquad (4\text{-}15)$$
$$= 0 \quad \text{with probability } (1 - u^j_{mi})$$

$$\nu^j_{in}(t) = \theta_{ij} \quad \text{with probability } u^j_{in} \qquad (4\text{-}16)$$
$$= 0 \quad \text{with probability } (1 - u^j_{in})$$

according to what was established in section 4.1

Calculation of the incremental mean coefficients

The mean value of $\Delta x_{ij}(t)$ is from (4-14) - (4-16)

$$E\left[\Delta x_{ij}(t) \mid \underline{x}(t)\right] = E\left[\Delta x_{ij}(t)\right] =$$
$$= \theta_{ij}\left(a_{ij} + \sum_{\substack{m \in I(i) \\ m \neq j}} u^j_{mi} - \sum_{n \in D(i)} u^j_{in}\right) \qquad (4\text{-}17)$$

Substituting $\theta_{ij} = K_{ij}^{-(1/2)} \sqrt{\Delta t}$ and taking limits in (2-13) when $\Delta t \to 0$, we obtain

$$\beta_{ij} = \lim_{\Delta t \to 0} \frac{E\left[\Delta x_{ij}(t)\right]}{\Delta t} = \tag{4-18}$$

$$= \lim_{\Delta t \to 0} \frac{a_{ij} + \sum_{\substack{i \in I(i) \\ m \neq j}} u^j_{mi} - \sum_{n \in D(i)} u^j_{in}}{K_{ij}^{1/2} \cdot \Delta t}$$

for all pairs (ij) = 1,2, ... M ; $M \leq N(N-1)$

Calculation of the covariance matrix elements

--- Diagonal elements : from (4-13). since $\Delta x_{ij}(t)$ is a sum of independent random variables we have:

$$\mathrm{Var}\left[\Delta x_{ij}(t)\right] = \mathrm{Var}\left[\mu_{ij}(t)\right] + \sum_{\substack{m \in I(i) \\ m \neq j}} \mathrm{Var}\left[\nu^j_{mi}(t)\right] + \tag{4-19}$$

$$+ \sum_{n \in D(i)} \mathrm{Var}\left[\nu^j_{in}(t)\right]$$

and substituting the value of the variance corresponding to a Bernoulli random variable we obtain:

$$\mathrm{Var}\ \Delta x_{ij}(t) = \theta^2_{ij}\left(a_{ij}(1-a_{ij}) + \sum_{\substack{m \in I(i) \\ m \neq j}} u^j_{mi}(1-u^j_{mi}) + \right. \tag{4-20}$$

$$\left. + \sum_{n \in D(i)} u^j_{in}(1-u^j_{in}) \right)$$

--- Off-diagonal elements: Consider two different queues $x_{ij}(t)$ and $x_{k.l}(t)$ (Fig. 4.6)

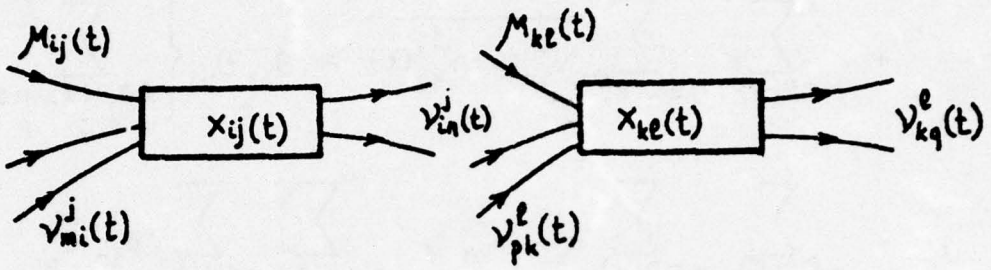


Fig. 4.6: Arrival and departure processes involved in two queues

$$\Delta x_{ij}(t) = \mu_{ij}(t) + \sum_{\substack{m \in I(i) \\ m \neq j}} \nu^j_{mi}(t) - \sum_{n \in D(i)} \nu^j_{in}(t)$$

$i \neq j$

$$\Delta x_{kl}(t) = \mu_{kl}(t) + \sum_{\substack{p \in I(k) \\ p \neq l}} \nu^l_{pk}(t) - \sum_{q \in D(k)} \nu^l_{kq}(t)$$

$k \neq l$

Since $\mu_{ij}(t)$ and $\mu_{kl}(t)$ are independent:

$$\mathrm{Cov}\left[\mu_{ij}(t)\ \mu_{kl}(t)\right] = \mathrm{Cov}\left[\mu_{ij}(t)\ \nu^l_{pk}(t)\right] = \mathrm{Cov}\left[\nu^j_{mi}(t)\mu_{kl}(t)\right] = 0$$

then

$$\mathrm{Cov}\left[\Delta x_{ij}(t)\ \Delta x_{kl}(t)\right] = \sum_{\substack{m \in I(i) \\ m \neq j}} \sum_{\substack{p \in I(k) \\ p \neq l}} \overline{\nu^j_{mi}(t)\ \nu^l_{pk}(t)} -$$

$i \neq j \quad ; \quad k \neq l$

$$- \sum_{\substack{m \in I(i) \\ m \neq j}} \sum_{q \in D(k)} \overline{\nu^j_{mi}(t)\, \nu^l_{kq}(t)} - \sum_{n \in D(i)} \sum_{\substack{p \in I(k) \\ p \neq l}} \overline{\nu^j_{in}(t)\, \nu^l_{pk}(t)} +$$

$$+ \sum_{n \in D(i)} \sum_{q \in D(k)} \overline{\nu^j_{in}(t)\, \nu^l_{kq}(t)} = \theta_{ij}\, \theta_{kl} \Bigg\{ \sum_{\substack{m \in I(i) \\ m \neq j}} \sum_{\substack{p \in I(k) \\ p \neq l}} u^j_{mi} u^l_{pk} -$$

$$- \sum_{\substack{m \in D(i) \\ m \neq j}} \sum_{q \in D(k)} u^j_{mi}\, u^j_{kq} - \sum_{n \in D(i)} \sum_{\substack{p \in I(k) \\ p \neq l}} u^j_{in}\, u^l_{pk} +$$

$$+ \sum_{n \in D(i)} \sum_{q \in D(k)} u^j_{in}\, u^l_{kq} \Bigg\} \qquad (4\text{-}21)$$

Consider each term in (4-21) separately (see Fig. 4.7 as an example for $N = 4$, $M = N(N-1) = 12$)

1) $m \in I(i)$, $m \neq j$; $p \in I(k)$, $p \neq l$

$$\overline{\nu^j_{mi}(t)\ \nu^l_{pk}(t)} = \begin{cases} u^j_{mi}\, u^l_{pk} \ ; \ m,i \neq p,k & \text{(a)} \\ 0 \ ; \ m,i = p,k \ \text{ and } \ j \neq l & \text{(b)} \\ \text{cannot be that; } m,i = p,k \ \text{ and } \ j = l & \text{(c)} \end{cases}$$

The above expressions are obtained because:

(a) $\nu^j_{mi}(t)$ and ν^l_{pk} are independent (queues at the same different nodes)

(b) Messages with different destination ($j \neq l$) cannot go over the same

channel (m,i = p,k) at the same time t (see Fig 4.7, both queues are in the same node i = k)

(c) This case corresponds to the $\mathrm{Var}\left[\nu_{mi}^{j}(t)\right]$ and was calculated previously in (4-19). Observe that the terms corresponding to (4-22a) when substracting the products of the means to obtain the covariance, will yield

$$\overline{\nu_{mi}^{j}(t)\ \nu_{pk}^{l}(t)} - \overline{\nu_{mi}^{j}(t)}\ \ \overline{\nu_{pk}^{l}(t)} = u_{mi}^{j}\ u_{pk}^{l} - u_{mi}^{j} \cdot u_{pk}^{l} = 0$$

and the terms corresponding to (4-22b) will yield

$$\overline{\nu_{mi}^{j}(t)\ \ \nu_{pk}^{l}(t)} - \overline{\nu_{mi}(t)}\ \ \overline{\nu_{pk}^{l}(t)} = 0 - u_{mi}^{j}\ \ u_{pk}^{l}$$

2) $m \in I(i)$, $m \neq j$ $\qquad q \in D(k)$

$$\overline{\nu_{mi}^{j}(t)\ \nu_{kq}^{l}(t)} = \begin{cases} u_{mi}^{j}\ u_{kq}^{l} & ;\ m,i \neq k,q \qquad (a) \\ 0 & ;\ m,i = k,q \text{ and } j \neq l \qquad (b) \\ u_{mi}^{j} & ;\ m,i = k,q \text{ and } j = l \qquad (c) \end{cases} \qquad (4\text{-}23)$$

The equations (4-23) are obtained because:

(a) and (b) the same reason as in (4-22)

(c) it is the same random variable $\nu_{mi}^{j}(t)$ corresponding to two different queues $x_{ij}(t)$ and $x_{kl}(t)$ which are in nodes i and k respectively. See for example Fig 4.7: Consider the processes $x_{12}(t)$ and $x_{32}(t)$ belonging to nodes 1 and 3 respectively (i=1, k=3, j=l=2)

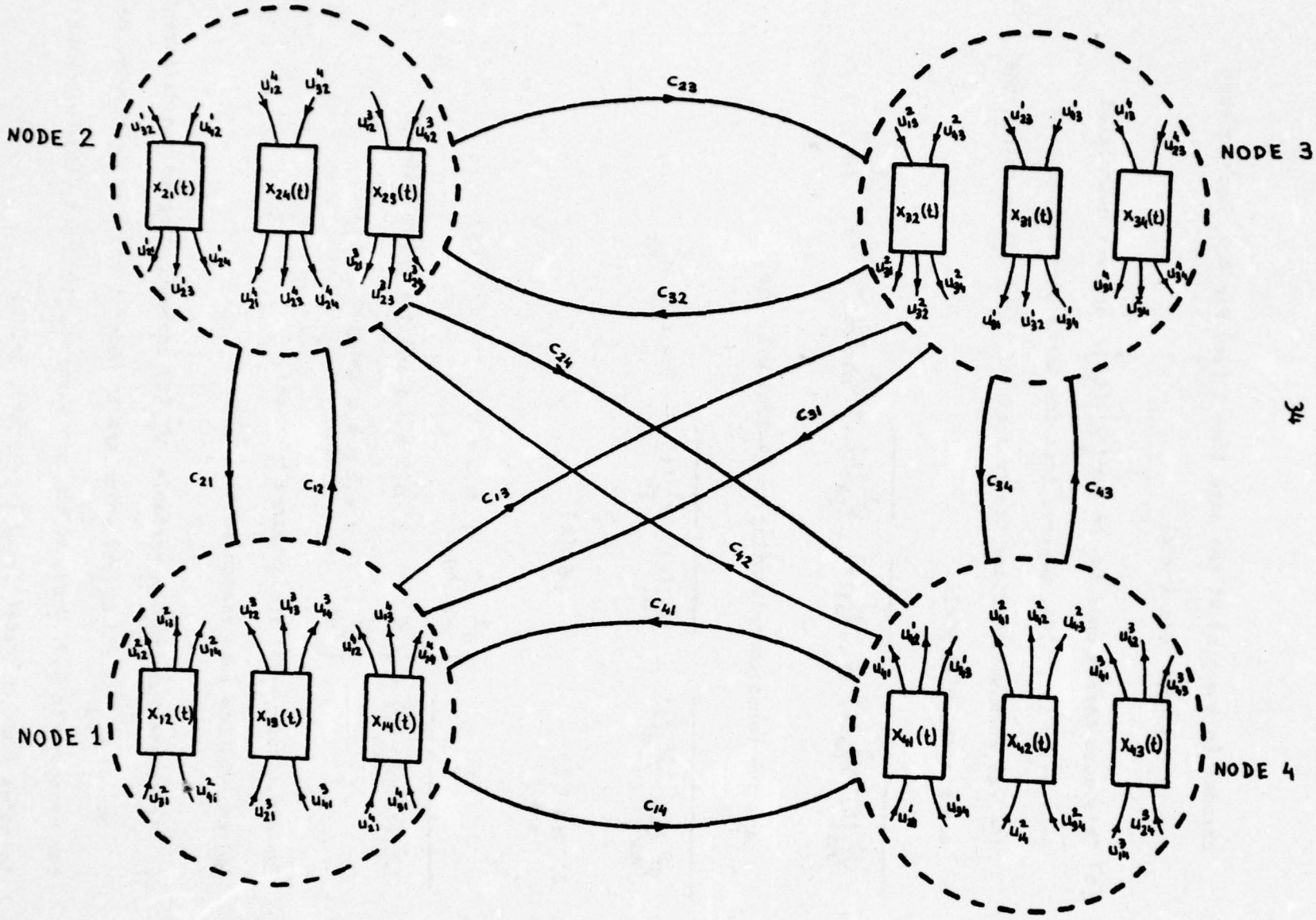


Fig.4.7. Detail of all queueing process in a network of four nodes.

When m=k=3 and q=i=1 the term

$$\overline{\nu_{31}^2(t)}\ \overline{\nu_{31}^2(t)} = \overline{\left[\nu_{31}^2(t)\right]^2} = u_{31}^2$$

as indicated by (4-23c)

The term corresponding to (4-23c) in the covariance expression will be

$$\overline{\left[\nu_{mi}^j(t)\right]^2} - \overline{\left[\nu_{mi}^j(t)\right]^2} = u_{mi}^j - (u_{mi}^j)^2 = u_{mi}^j\,(1 - u_{mi}^j)$$

3) $n \in D(i)$; $p \in I(k)$, $p \neq 1$

This case is exactly like 2). Just interchange this subscripts

$i \longleftrightarrow k$, $m \longleftrightarrow p$, $q \longleftrightarrow n$, $j \longleftrightarrow 1$

4) $n \in D(i)$; $q \in D(k)$

This case is similar to 1), The only difference is that it takes the processes coming out from node i and node k whereas in 1) we had the processes entering nodes i an k, Therefore the corresponding expression is

$$\overline{\nu_{in}^j(t)\ \nu_{kq}^1(t)} = \begin{cases} u_{in}^j\ u_{kq}^1 & ;\ i,n \neq k,q \quad (a) \\ 0 & ;\ i,n = k,q \text{ and } j \neq 1 \quad (b) \\ \text{cannot be that } i,n = k,q \text{ and } j = 1 & \quad (c) \end{cases}$$

(4-24)

According to this we can have several cases:

A) Both queues are in the same node i = k. In this case only the first and fourth term of (4-20) enter in the covariance (relating the inputs and outputs respectively):

$$\underset{i\neq j\neq l}{\text{Cov}}\left[\Delta x_{ij}(t)\ \Delta x_{il}(t)\right] = -\sum_{\substack{m\in I(i)\\ m\neq j\neq i\neq l\\ (N-3)\text{terms}}} u^{j}_{mi}\ u^{l}_{mi} - \sum_{\substack{n\in D(i)\\ n\neq i\\ (N-1)\text{ terms}}} u^{j}_{in}\ u^{l}_{in} \qquad (4\text{-}25)$$

B) Both queues are m different nodes $i \neq k$. In this case the 2nd. and 3rd. term of (4-20) appear. There are two possibilities:

B-1) Both queues have the same destination : $j = l$

$$\underset{i\neq k\neq j}{\text{Cov}}\left[\Delta x_{ij}(t)\ \Delta x_{kj}(t)\right] = -u^{j}_{ik}\left(1-u^{j}_{ik}\right) - u^{j}_{ki}\left(1-u^{j}_{ki}\right) \qquad (4\text{-}26)$$

B-2) The queues have different destinations: $j \neq l$

$$\underset{i\neq j\ ;\quad k\neq l}{\text{Cov}}\left[\Delta x_{ij}(t)\ \Delta x_{kl}(t)\right] = u^{j}_{ik}\ u^{l}_{ik} + u^{l}_{ki}\ u^{j}_{ki} \qquad (4\text{-}27)$$

but observe that if $k \neq j$

$$\text{Cov}\left[\Delta x_{ij}(t)\ \Delta x_{jl}(t)\right] = u^{j}_{ij}\ u^{l}_{ij} \quad (\text{because } u^{j}_{ji} = 0)$$

and if $i = l$

$$\text{Cov}\left[\Delta x_{ij}(t)\ \Delta x_{ji}(t)\right] = u^{i}_{ki}\ u^{i}_{ki} \quad (\text{because } u^{i}_{ik} = 0)$$

4.4 Conditions for the diffusion to be valid

Going back to the expression (4-18), if β_{ij} has to have a finite value it is necessary that the numerator of that expression tends to zero as $(\Delta t)^{1/2}$.

Recall what was established in section 2.2. concerning to the conditions for diffusion: for the variance and mean per unit time to

make sense the probabilitites of jumping upwards and downwards have to be "nearly" the same, the difference being a quantity depending on $(\Delta t)^{1/2}$ and therefore this difference decreases proportionally to $(\Delta t)^{1/2}$ when Δt tends to zero. This is reflected by the expressions (2-6) and (2-7).

Therefore let us assume that the probabilities a_{ij} and u_{ij}^{k} are of the form of (2-6) and (2-7) that is a constant term plus another depending on $(\Delta t)^{1/2}$, that is:

$$a_{ij} = A_{ij} + p_{ij}\sqrt{\Delta t} \qquad (4\text{-}28)$$

$$u_{ij}^{k} = U_{ij}^{k} + y_{ij}\sqrt{\Delta t} \qquad (4\text{-}29)$$

The channel capacities which are related with the service process will be considered of the same form too, that is a fixed term plus another varying with $(\Delta t)^{1/2}$ as shown:

$$c_{ij} = C_{ij} + q_{ij}\sqrt{\Delta t} \qquad (4\text{-}30)$$

which means that the capacity of channel i - j is variable about a fixed value C_{ij} in a quantity proportional to $(\Delta t)^{1/2}$

Then the conditions that have to be satisfied for the diffusion property, are obtained by plugging (4-28), (4-29) and (4-30) in (4-18). Thus we obtain:

$$A_{ij} + \sum_{\substack{m \in I(i) \\ m \neq j}} U_{mi}^{j} - \sum_{n \in D(i)} U_{in}^{j} = 0 \qquad (4\text{-}31)$$

$$p_{ij} + \sum_{\substack{m \in I(i) \\ m \neq j}} y^j_{mi} - \sum_{n \in D(i)} y^j_{in} = \sqrt{K_{ij}}\ \beta_{ij} \qquad (4\text{-}32)$$

for all pairs (i, j) = 1, 2, ... M ; $M \leqslant N(N-1)$

The constants K_{ij} are given by the buffer size. From (4-12) we see that $K_{ij}(\theta_{ij})^2 = K_{kl}(\theta_{kl})^2$ but $\theta_{ij} = 1/N_{ij}$ and thus $K_{ij}/K_{kl} = (N_{ij}/N_{kl})^2$.

The expression (4-31) is related with the deterministic flow at each queue and simply states the balance that has to exist at each node between inputs and outputs if the traffic were deterministic, whereas (4-32) gives an idea of the infinitesimal variation during $(t, t + \Delta t)$ since the drift β_{ij} indicates de tendency of queue (ij) to increase or decrease per unit time. Moreover from the capacity constraints (4-1):

$$c_{ij} \geqslant \sum_{i \neq k} U^k_{ij} \qquad ; \quad U^k_{ij} \geqslant 0 \qquad (4\text{-}33)$$

$$q_{ij} \geqslant \sum_{i \neq k} y^k_{ij} \qquad ; \quad y^k_{ij} \geqslant 0 \qquad (4\text{-}34)$$

4.5.- Optimization procedure

In the model we have established, we have a network whose channels c_{ij} are given and have fixed capacities. The input traffics a_{ij} depend on the users' demand so that are considered as fixed quantities too. The question is to find the best routing strategies within the system which are represented by the probabilities u^k_{ij} defined in section 4.1

subject to the capacity constraints (4-1). The quantities u_{ij}^{k} will be called routing variables and will have to be chosen to minimize the average queue length in the system according to the given input traffic and the fixed capacitites. This is an open-loop type control procedure.

Note(Fig 4.7) that the maximun number of routing variables we can have per queue is N - 1 so that in the whole system we can have at most M(N - 1) control variables.

The expression of the covariance elements follow from the expressions (4-20), (4-25), (4-26), (4-27) and (4-11), (4-28), (4-29). Thus we have

Variance elements (recall (4-20))

$$\alpha_{(ij)^2} = \frac{A_{ij}(1-A_{ij}) + \sum_{\substack{m \in I(i) \\ m \neq j}} U_{mi}^{j}(1-U_{mi}^{j}) + \sum_{n \in D(i)} U_{in}^{j}(1-U_{in}^{j})}{K_{ij}} \qquad (4\text{-}35)$$

Covariance elements: a) $i = k;\ i \neq j \neq l$ (recall(4-2))

$$\alpha_{(ij),(il)} = - \frac{\sum_{\substack{m \in I(i) \\ m \neq j}} U_{mi}^{j}\, U_{mi}^{l} + \sum_{n \in D(i)} U_{in}^{j}\, U_{in}^{l}}{\sqrt{K_{ij}\, K_{il}}} \qquad (4\text{-}36)$$

b) $i \neq k \neq j = l$ (recall (4-26))

$$\alpha_{(ij),(kj)} = - \frac{U_{ik}(1 - U_{ik}^{j}) + U_{ki}^{j}(1 - U_{ki}^{j})}{\sqrt{K_{ij}\, K_{kj}}} \qquad (4\text{-}37)$$

c) $i \neq k$; $j \neq l$ (recall(4-27))

$$\alpha_{(ij),(kl)} = \frac{U_{ik}^{j} \; U_{ik}^{l} + U_{ki}^{l} \; U_{ki}^{j}}{\sqrt{K_{ij} \; K_{kl}}} \qquad (4\text{-}38)$$

(Remember special cases of (4-27) when k = j or i = l)

Therefore the covariace per unit time is a M x M matrix $\underline{\Lambda}$ which depends on the M possible inputs A_{ij} assumed to be given and on the L control variablea U_{ij}^{k} where $L \leqslant M(N-1)$. We write $\underline{\Lambda} = \underline{\Lambda}(\underline{A},\underline{U})$ where $\underline{A}$ and $\underline{U}$ are respectively M and L dimensional vectors.

Notice than the L routing variables U_{ij}^{k} may not be chosen independently, because they have to satisfy the system of equations (4-31) so that only L - M variables are independent, provided that they satisfy the capacity constraints (4-33).

We want to minimize the overall mean (4-9) that is

$$\min F(\underline{\gamma}) = \min_{\text{all queues}} \sum_{i=1}^{M} \rho_{ij} \left(\frac{e^{\gamma_{ij}}}{e^{\gamma_{ij}} - 1} - \frac{1}{\gamma_{ij}} \right) \qquad (4\text{-}39)$$

where γ_{ij} are the elements of the vector $\underline{\gamma}$ defined as $\underline{\gamma} = 2\underline{\Lambda}^{-1} \underline{\beta}$ The drift vector $\underline{\beta}$ can be expressed in terms of the inputs p_{ij} and the control variables $y_{mi}^{j}(m \in I(i) \; ; \; m \neq j)$ and y_{in}^{j} $(n \in D(i))$. From (4-32) and using matrix notation:

$$\underline{\beta} = \underline{p} + \underline{H} \; \underline{y} \qquad (4\text{-}40)$$

where $\underline{\beta} = \left[\beta_{ij} \right]$

$$\underline{p} = \left[p_{ij} / \sqrt{K_{ij}} \right]$$

$$\underline{Y} = \begin{bmatrix} y^j_{mi} / \sqrt{K_{ij}} & m \in I(i);\ m \neq j \\ \hline y^j_{in} / \sqrt{K_{ij}} & n \in D(i) \end{bmatrix}$$

for all possible pairs (i j) and where $\underline{H}$ is a M x L matrix depending on the specific configuration of the system and whose elements can be only + 1 for incoming links, -1 for outgoing links and 0 when there is no conexion

Then we have:

$$\underline{\gamma} = 2\underline{\Lambda}^{-1}\ \underline{p} + 2\underline{\Lambda}^{-1}\ \underline{H}\ \underline{y} \qquad (4\text{-}41)$$

Call for convenience

$$2\underline{\Lambda}^{-1}\ \underline{p} = \underline{d} \quad \text{(M-dim. vector)} \qquad (4\text{-}42)$$

and

$$2\underline{\Lambda}^{-1}\ \underline{H} = \underline{D} \quad \text{(M x L matrix)} \qquad (4\text{-}43)$$

then

$$\underline{\gamma} = \underline{d} + \underline{D}\ \underline{y} \qquad (4\text{-}44)$$

and the problem is to find

$$\min_{\underline{u}} \quad F(\underline{d} + \underline{D}\ \underline{y}) \qquad (4\text{-}45)$$

where $\underline{u} = \underline{U} + \underline{y}\sqrt{\Delta t}$. The minimization (4-45) is to be carried over the vectors $\underline{U}$ and $\underline{y}$ with the constraints (4-31) and (4-33) on the vector $\underline{U}$ and with the constraints (4-34) on the vector $\underline{y}$.

The constraints the elements of $\underline{U}$ have to satisfy are those given by (4-33) (capacity constraints) and (4-31) (flow balance at

each queue). In general for given inputs A_{ij} and capacities C_{ij}, the vector $\underline{U}$ satisfying (4-31) and (4-33) will not be unique, and for different choices of $\underline{U}$ the covariance matrix $\underline{\Lambda}$ will be different and so will be the minimum of (4-45). In order to simplify the minimization (4-45) it will be assumed in this thesis that $\underline{\Lambda}$ is fixed, that is we have chosen a vector $\underline{U}$ satisfying the requirements of (4-31) and (4-33) which represents an equilibrium situation for the system and we shall be interested in how the system will behave for small alterations about the equilibrium situation. In particular we shall seek how the routing variables y_{ij}^{k} will vary so that the overall average queue length of the system is minimum.

With this asumption the vector $\underline{d}$ and the matrix $\underline{D}$ are constant and the minimization problem can be stated as :

$$\min_{\underline{y}} \quad F(\underline{d} + \underline{D}\,\underline{y}) = \min_{\underline{y}} \quad F_1(\underline{y}) \qquad (4\text{-}46a)$$

subject to:

$$\sum_{i \neq k} y_{ij}^{k} \leqslant q_{ij} \quad ; \quad y_{ij} \geqslant 0 \qquad (4\text{-}46b)$$

The aim is then to find the optimum vector $\underline{y}^*$ that satisfies the minimization (4-46).

Notice that the function we want to minimize es a sum of M functions like the one shown in Fig 4.4 which is convex for $\gamma_{ij} \leqslant 0$ the meaning of this being that the queue is loaded less than or equal to $0.5\,\rho_i$ on the average. Therefore if for all queues $\gamma_{ij} \leqslant 0$ then the function $F(\underline{\gamma})$ will be convex and will have a well defined minimum over the constrained region. The convexity property is convenient to

include it when the minimum is searched by numerical methods starting with some initial guess. Physically it reflects the fact that the messages arriving at the nodes do not stack up at the queues so that the system behaves "nicely" and does not become congested.

5.- ILLUSTRATIVE EXAMPLES

Let us now apply the theory developped in the preceding sections to some specific examples.

Our purpose is to minimize a function of L variables subject to some constraints over a convex region.

Because of the exponential nature of the function (4-8) even with a small number of variables it is not possible to find analytic solutions and one therefore has to use numerical procedures.

The minimization procedure that will be used is based upon the method of Zangwill [27] which is a modification of Powell's algorithm [18].

Basically an initial point and a set of L directions are chosen. Along each direction the minimum coordinate is found. In the next step the first L - 1 directions are taken as the L-1 last directions in the first step. The L-th direction is taken as the difference between the initial point and the minimum found in the preceding iteration and so on until convergence is reached.

5.1.- Example with two queues

In Fig 5.1 it is shown an example of three computers: Messages come to nodes 1 and 2 and have to be transmitted to node 3 either directly or via the indirect path. When messages get to computer 3 they leave the system. No messages enter at 3.

According to what was stated in section 4.1 the time is divided in

small intervals $(t, t + \Delta t)$ and during this time we call a_{13} snd a_{23} the

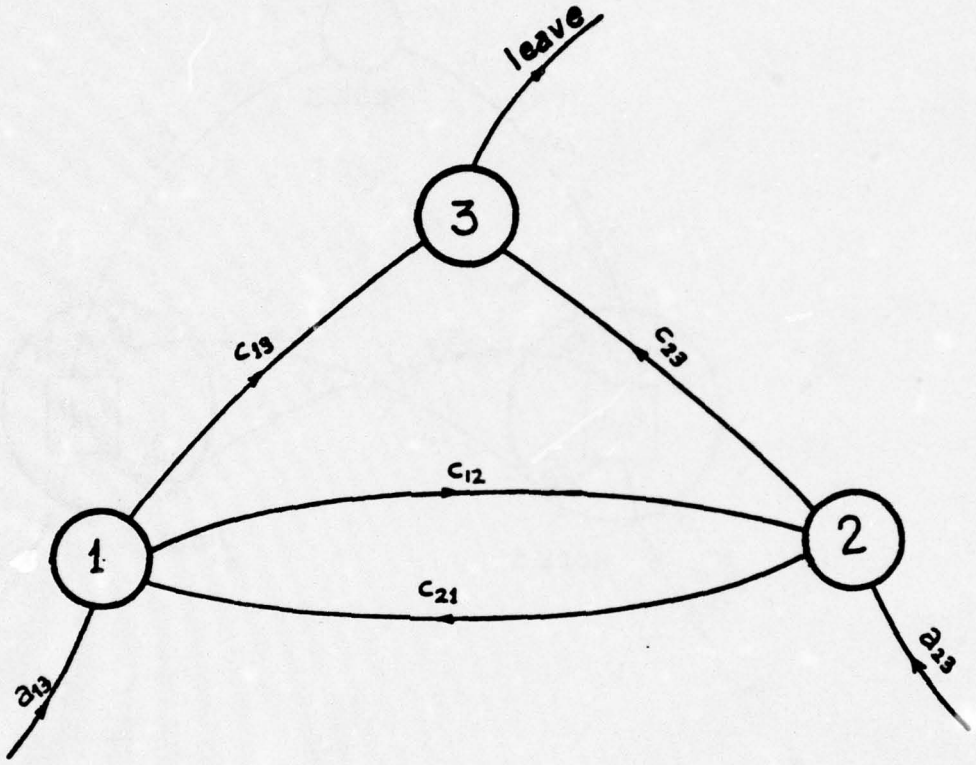


Fig 5.1.- Network of three nodes: two sources and one destination

probability that a message enter node 1 and node 2 respectively.

The traffic a_{13} can go directly through the channel c_{13} or through c_{12} and c_{23}. Similarly for the traffic a_{23}.

As it can be seen in Fig 5.2 there are only two queues in the system, one $x_{13}(t)$ corresponding to node 1 and the other $x_{23}(t)$ corresponding to node 2. There are no queues at node 3 because this is only destination node. Similarly there are no queues $x_{12}(t)$ at node 1 and $x_{21}(t)$ at node 2 because neither of those nodes are final destination but traffic source

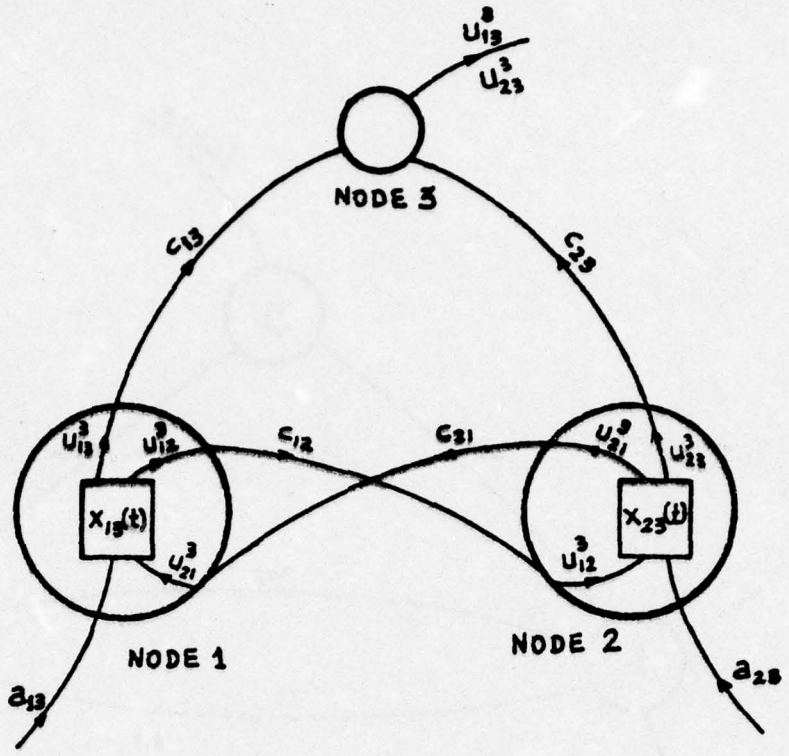


Fig. 5.2. Network of three nodes. Queues detail

or intermediate destination nodes.

The capacity constraints are

$$u_{13}^3 \leq c_{13} \; ; \quad u_{23}^3 \leq c_{23} \; ; \quad u_{12}^3 \leq c_{12}^3 ; \quad u_{21}^3 \leq c_{21} \qquad (5\text{-}1)$$

The capacities are assumed to be of the form (4-30)

$$\begin{aligned} c_{13} &= C_{13} + q_{13}\sqrt{\Delta t} \\ c_{23} &= C_{23} + q_{23}\sqrt{\Delta t} \end{aligned} \qquad (5\text{-}2)$$

$$c_{12} = C_{12} + q_{12}\sqrt{\Delta t}$$
$$c_{21} = C_{21} + q_{21}\sqrt{\Delta t} \qquad (5\text{-}2)$$

The external and internal inputs and outputs are of the form of (4-28) and (4-29)

$$a_{13} = A_{13} + p_{13}\sqrt{\Delta t}$$
$$a_{23} = A_{23} + p_{23}\sqrt{\Delta t}$$
$$u_{13} = U^3_{13} + y^3_{13}\sqrt{\Delta t}$$
$$u^3_{23} = U^3_{23} + y^3_{23}\sqrt{\Delta t} \qquad (5\text{-}3)$$
$$u^3_{12} = U^3_{12} + y^3_{12}\sqrt{\Delta t}$$
$$u^3_{21} = U^3_{21} + y^3_{21}\sqrt{\Delta t}$$

The proportionality constants K_{13} and K_{23} which relate the time interval Δt and the queue step sizes θ_{13} and θ_{23} will be taken as unity. Then we have $\Delta t = \theta^2_{13} = \theta^2_{23}$ that is the step size in both queues is the same and tends to zero as $(\Delta t)^{1/2}$. This makes sense in the case that both buffers have the same capacity and it will be assumed so.

The expressions for the means per unit time are from (4-32)

$$\beta_{13} = p_{13} + y^3_{21} - y^3_{13} - y^3_{12}$$
$$\beta_{23} = p_{23} + y^3_{12} - y^3_{23} - y^3_{21} \qquad (5\text{-}4)$$

and the flow equations from (4-31)

$$\left.\begin{aligned} A_{13} + U^3_{21} - U^3_{13} - U^3_{12} &= 0 \\ A_{23} + U^3_{12} - U^3_{23} - U^3_{21} &= 0 \end{aligned}\right\} \qquad (5\text{-}5)$$

The elements of the covariance matrix are: From (4-35)

$$\alpha_{(13),(13)} \equiv \alpha_{11} = A_{13}(1 - A_{13}) + U^3_{21}(1 - U^3_{21}) + U^3_{13}(1 - U^3_{13}) + U^3_{12}(1 - U^3_{12}) \qquad (5\text{-}6)$$

$$\alpha_{(23),(23)} \equiv \alpha_{22} = A_{23}(1 - A_{23}) + U^3_{12}(1 - U^3_{12}) + U^3_{23}(1 - U^3_{23}) + U^3_{21}(1 - U^3_{21}) \qquad (5\text{-}7)$$

From (4-37)

$$\alpha_{(13),(23)} \equiv \alpha_{12} = - U^3_{12}(1 - U^3_{12}) - U^3_{21}(1 - U^3_{21}) \qquad (5\text{-}8)$$

The expression we want to minimize is (4-39)

$$F(\underline{\gamma}) = \rho_{13}\left(\frac{e^{\gamma_{13}}}{e^{\gamma_{13}} - 1} - \frac{1}{\gamma_{13}}\right) + \rho_{23}\left(\frac{e^{\gamma_{23}}}{e^{\gamma_{23}} - 1} - \frac{1}{\gamma_{23}}\right) \qquad (5\text{-}9)$$

where $$\underline{\gamma} = (\gamma_{13}, \gamma_{23})^T \qquad (5\text{-}10)$$

and such that $$\underline{\gamma} = 2\,\underline{\Lambda}^{-1}\,\underline{\beta} \equiv \underline{V}\,\underline{\beta} \qquad (5\text{-}11)$$

and $$\underline{\beta} = (\beta_{13}, \beta_{23})^T \qquad (5\text{-}12)$$

From $$\underline{\Lambda} = \begin{bmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{12} & \alpha_{22} \end{bmatrix} \qquad (5\text{-}13)$$

We obtain

$$\underline{V} = \begin{bmatrix} V_{11} & V_{12} \\ V_{12} & V_{22} \end{bmatrix} \tag{5-14}$$

where

$$V_{11} = 2 \frac{\alpha_{22}}{\alpha_{11}\alpha_{22} - \alpha_{12}^2} \tag{5-15}$$

$$V_{22} = 2 \frac{\alpha_{11}}{\alpha_{11}\alpha_{22} - \alpha_{12}^2} \tag{5-16}$$

$$V_{12} = -2 \frac{\alpha_{12}}{\alpha_{11}\alpha_{22} - \alpha_{12}^2} \tag{5-17}$$

Observe from (5-6), (5-7) and (5-8) that V_{11}, V_{22} and V_{12} are non-negative and that $V_{11} \geq V_{12}$ and $V_{22} \geq V_{12}$.

Notice that (5-4) can be rearranged es:

$$\left.\begin{aligned} \beta_{13} &= (p_{13} - y_{13}^3) - (y_{12}^3 - y_{21}^3) \\ \beta_{23} &= (p_{23} - y_{23}^3) - (y_{12}^3 - y_{21}^3) \end{aligned}\right\} \tag{5-18}$$

so that $\underline{\gamma}$ as well as $\underline{\beta}$ are only dependent on three variables rather than four; they are y_{13}^3, y_{23}^3 and $y_{12}^3 - y_{21}^3$

Call for convenience

$$\delta_1 = p_{13} - y_{13} \tag{5-19}$$

$$\delta_2 = p_{23} - y^3_{23} \qquad (5\text{-}20)$$

$$z = y^3_{12} - y^3_{21} \qquad (5\text{-}21)$$

$$\beta_{13} = \delta_1 - z$$
$$\beta_{23} = \delta_2 + z \qquad (5\text{-}22)$$

Therefore from (5-11)

$$\gamma_{13} = V_{11}\,\beta_{13} + V_{12}\,\beta_{23} = V_{11}\,\delta_1 + V_{12}\,\delta_2 - (V_{11} - V_{12})\,z \qquad (5\text{-}23)$$

$$\gamma_{23} = V_{12}\,\beta_{13} + V_{22}\,\beta_{23} = V_{12}\,\delta_1 + V_{22}\,\delta_2 + (V_{22} - V_{12})\,z \qquad (5\text{-}24)$$

The utilization factors are:

$$\rho_{13} = \frac{A_{13} + U^3_{21}}{C_{13} + C_{12}} \qquad (5\text{-}25)$$

$$\rho_{23} = \frac{A_{23} + U^3_{12}}{C_{23} + C_{21}} \qquad (5\text{-}26)$$

Minimizing (5-9) requires γ_{13} and γ_{23} as negative as possible (Fig 4.4.) Therefore from (5-23) and (5-24) since the coeficients of δ_1 and δ_2 are non negative, δ_1 and δ_2 must be as small as possible or from (5-19) and (5-20) y^3_{13} and y^3_{23} must have the maximun value which is the corresponding capacity, that is: $y^3_{13} = q_{13}$ (5-27)

$$y^3_{23} = q_{23} \qquad (5\text{-}28)$$

as we could have expected from Fig 5.2. because there is no reason for not using the channels q_{13} and q_{23} at full capacity.

Then the expression (5-9) is only a function of the variable z = $= y_{12}^3 - y_{21}^3$ which is bounded by the capacitites q_{12} and q_{21}:

$$- q_{21} \leqslant z \leqslant q_{12} \qquad (5\text{-}29)$$

Let us take some values to illustrate the example. Assume:

$A_{13} = 0.8$	$U_{13}^3 = 0.6$	$C_{13} = 0.75$
$A_{23} = 0.4$	$U_{23}^3 = 0.6$	$C_{23} = 0.75$
	$U_{12}^3 = 0.2$	$C_{12} = 0.75$
	$U_{21}^3 = 0$	$C_{21} = 0.75$

which satisfy the flow equations (5-14) and (5-15) and the capacity constraints $0 \leqslant U_{ij} \leqslant C_{ij}$, $\forall(ij)$.

The covariance elements have the value

$\alpha_{11} = 0.56$; $\alpha_{22} = 0.64$; $\alpha_{12} = -0.16$ and

$V_{11} = 3.8462$; $V_{22} = 3.3654$; $V_{12} = 0.9616$

the utilization factors $\rho_{13} = 0.53$, $\rho_{23} = 0.40$

then $\gamma_{13} = 3.8462\,\delta_1 + 0.9616\,\delta_2 - 2.8846\,z \qquad (5\text{-}30)$

$$\gamma_{23} = 0.9616\,\delta_1 + 3.3654\,\delta_2 + 2.4038\,z \qquad (5\text{-}31)$$

bearing in mind that $\delta_1 = p_{13} - q_{13}$ and $\delta_2 = p_{23} - q_{23}$.

Several cases are shown in Fig 5.3 to Fig 5.8.

Consider first Fig 5.3. for which we have chosen:

$$p_{13} = p_{23} = 0 \quad , \qquad q_{23} = q_{12} = q_{21} = 1$$

and q_{13} varies between 0 and 1, For $q_{13} = 1$ the optimum value of

$z = y^3_{12} - y^3_{21} = 0.2380$ and the value of $F_{min} = 0.1918$.

As q_{13} decreases (δ_1 increases), both γ_{13} and γ_{23} increase but the effect is more remarkable on γ_{13} because $V_{11} > V_{12}$ (See expressions (5-37) (5-38) and (5-44), (5-45). In order to have this increase as small as possible, z will increase because its coeficient in the expression for γ_{13} is negative. This is what can be seen in Fig 5.3 as q_{13} decreases. This physically means that when the capacity of the link connecting nodes 1 and 3 decreases, more messages tend to be sent via node 2 to partially compensate the capacity loss. As a consequence of the overall capacity reduction the overall mean value increases.

Fig 5.4 :

$$p_{13} = p_{23} = 0 \quad ; \quad q_{13} = q_{12} = q_{21} = 1$$

and now it is q_{23} what varies. By a similar argument we can see that when q_{23} decreases δ_2 increases and γ_{13} and γ_{23} increase too although the latter more. Therefore z has to decrease to compensate for, that is less messages travel from node 1 to node 2.

We can observe that in both cases (Figs. 5.3 and 5.4) the overall mean has the same value. The reason for this can be drawn from equations (5-23) and (5-24). In the case of Fig 5.3 $\delta_2 = -1$ and $\delta_1 = \delta$ varying between -1 and 0. In the case of Fig 5.4 $\delta_1 = -1$ and $\delta_2 = \delta$ varying between -1 and 0. Then

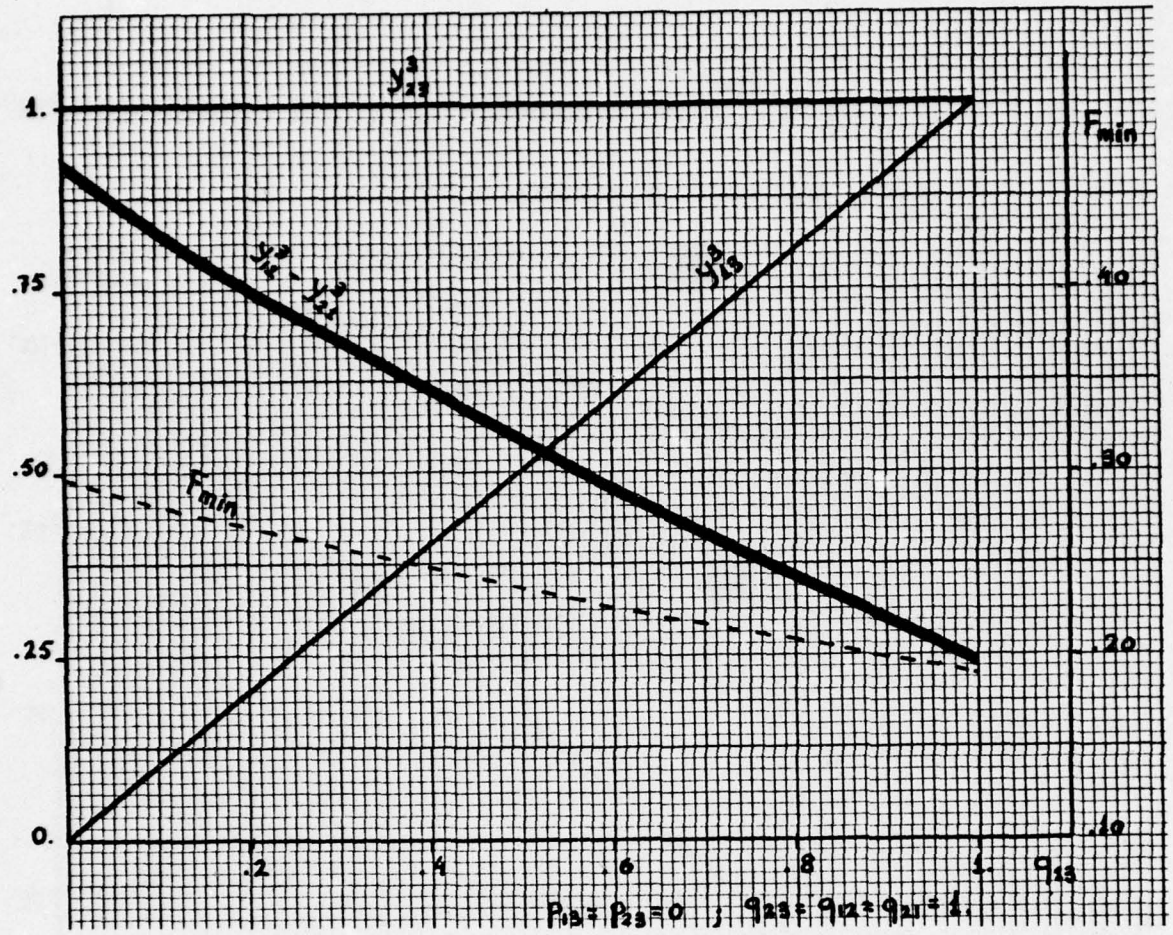


Fig. 5.3. Example 5.1. Routing variables and F_{min} in terms of q_{13}.

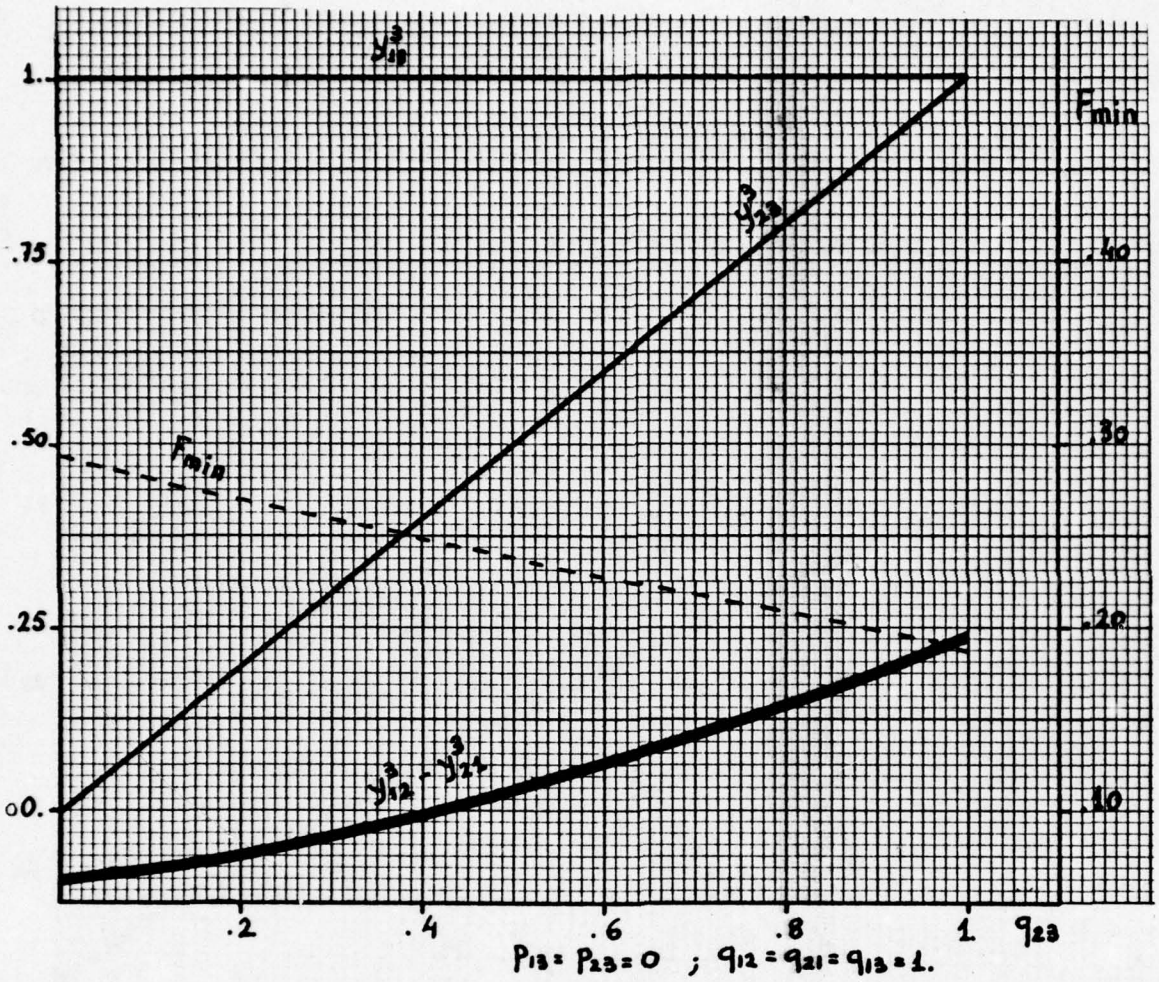


Fig. 5.4. Example 5.1. Routing variables and F_{min} in terms of q_{23}.

$$\left.\begin{aligned} \gamma'_{13} &= V_{11}\delta - V_{12} - (V_{11} - V_{12})\, z' \\ \gamma'_{23} &= V_{12} - V_{22} + (V_{22} - V_{12})\, z' \end{aligned}\right\} \quad \text{Fig 5.3}$$

$$\left.\begin{aligned} \gamma''_{13} &= -V_{11} + V_{12}\delta - (V_{11} - V_{12})\, z'' \\ \gamma''_{23} &= -V_{12} + V_{22}\delta + (V_{22} - V_{12})\, z'' \end{aligned}\right\} \quad \text{Fig 5.4}$$

From this equations we can see that whenever $z' - z'' = 1 + \delta$ then $\gamma'_{13} = \gamma''_{13}$ and $\gamma'_{23} = \gamma''_{23}$ so that the overall mean value is the same.

Fig 5.5: Now

$$p_{13} = p_{23} = 0 \quad ; \quad q_{13} = q_{23} = q_{21} = 1$$

and q_{12} varies between 1 and 0. For $q_{12} = 1$, $z = 0.2380$ so that there is no effect ehen decreasing q_{12} until it reaches the value 0.238. From this point on the value of $z = q_{12}$.

Therefore the effect over the overall mean takes place only when $z \leq 0.2380$. The reason for this is that when z decreases in (5-23) γ_{13} increases and in (5-24) γ_{23} decreases. Therefore the change in the overall mean is less.

Fig 5.6.: Decreasing q_{21} has no effect because this channel was not used. The overall mean does not change either.

Fig 5.7: Now all $q_{ij} = 1$, $p_{23} = 0$ and p_{13} increases. The effect of p_{13} increasing from 0 to 1 is the same as q_{13} decreasing from 1 to 0 in Fig 5.3 because in both cases δ_1 increases from 1 to 0. At some point z reaches its maximum value 1 and cannot increase any more. The overall

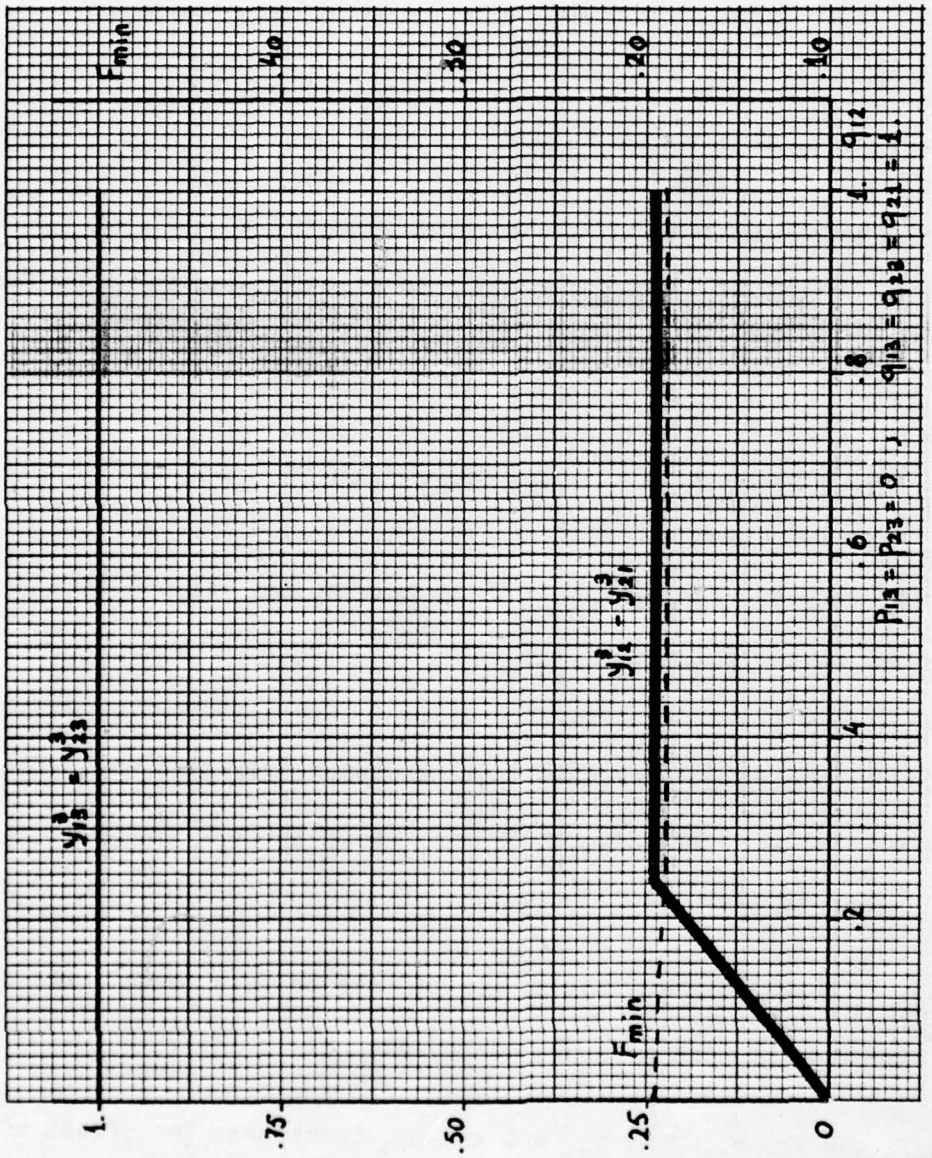


Fig. 5.5. Example 5.1. Routing variables and F_{min} in terms of q_{12}.

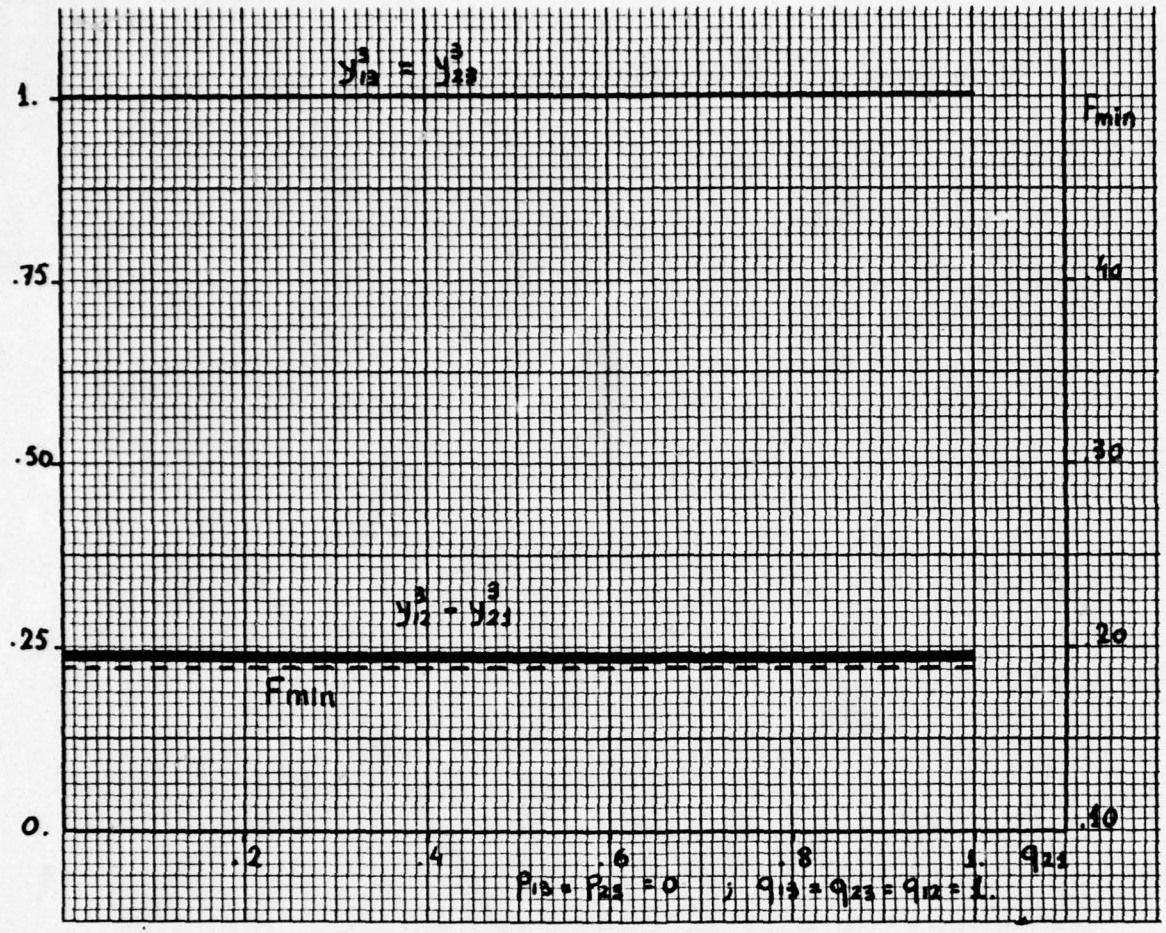


Fig. 5.6. Example 5.1. Routing variables and F_{min} in terms of q_{21}.

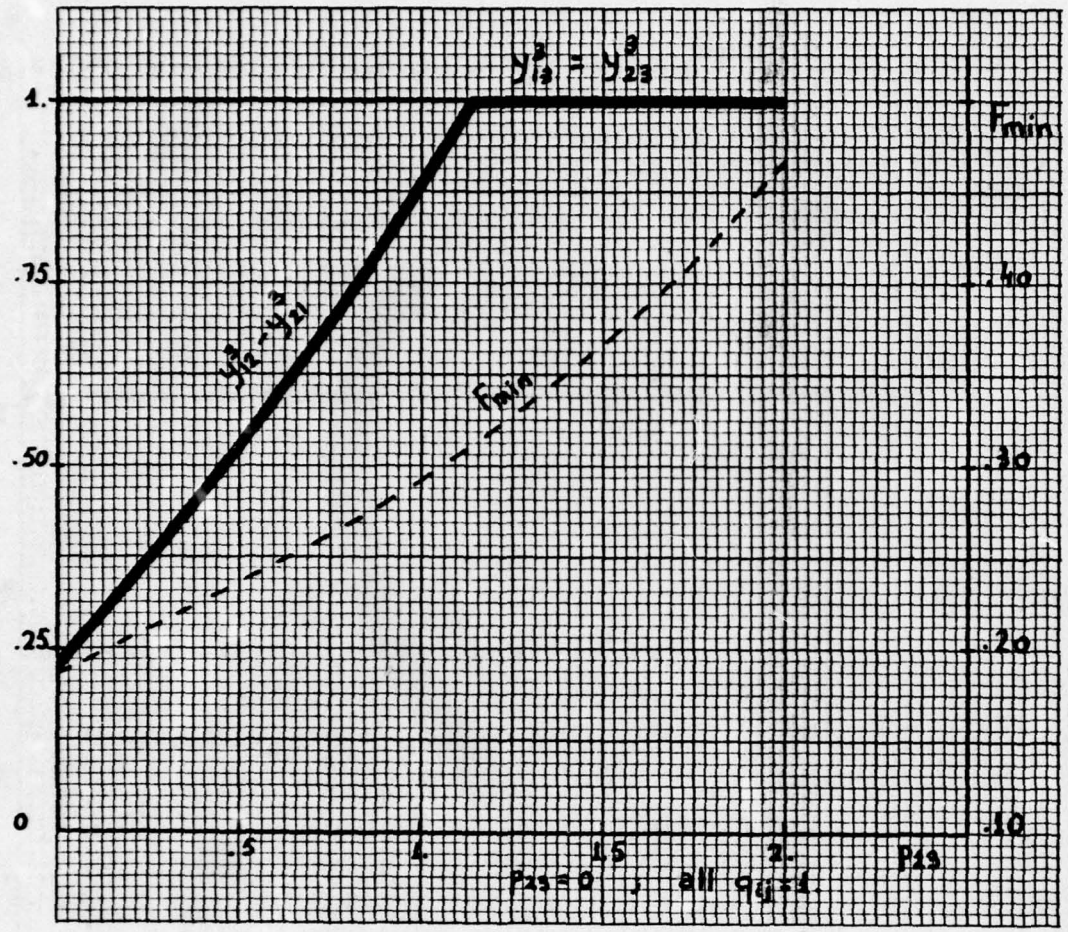


Fig. 5.7. Example 5.1. Routing variables and F_{min} in terms of p_{13}.

value increases then faster.

Fig. 5.8.- All $q_{ij} = 1$, $p_{13} = 0$ and p_{23} increases.

The efeffect of p_{23} increassing from 0 to 1 is the same as q_{23} decreasing from 1 to 0 in Fig. 5.54. For $p_{23} > 1$ we can see there is a minimum point in the overall mean value and then z increases again until it reaches 1. This can be explained from the expression of the derivative of $F(\underline{\gamma})$ with respect to z, From (5-9) it is obtained

$$\frac{dF}{dz} = \rho_{13} \frac{\partial F}{\partial \gamma_{13}} \frac{\partial \gamma_{13}}{\partial z} + \rho_{23} \frac{\partial F}{\partial \gamma_{23}} \frac{\partial \gamma_{23}}{\partial z}$$

The utilization factors were calculated: $\rho_{13} = 0.53$ and $\rho_{23} = 0.40$ and the derivatives of γ_{13} and γ_{23} with respect to z are straightforward from (5-30) and (5-31) yielding:

$$\frac{dF}{dz} = 0.53\ (-2.8846)\ \frac{\partial F}{\partial \gamma_{13}} + 0.40\ (2.4038)\ \frac{\partial F}{\partial \gamma_{23}} =$$

$$= -1.5384\ \frac{\partial F}{\partial \gamma_{13}} + 0.9615\ \frac{\partial F}{\partial \gamma_{23}} \qquad (5\text{-}31a)$$

where

$$\frac{\partial F}{\partial \gamma} = \frac{1}{\gamma^2} - \frac{e^{\gamma}}{(e^{\gamma} - 1)^2} \qquad (5\text{-}31b)$$

The expression (5-32b) is represented in Fig 5.8a.

From the observation of equations (5-30), (5-31) and (5-32) we can see how the variation of the optimum z is going to be when p_{23} (or equivalently σ_2) increases.

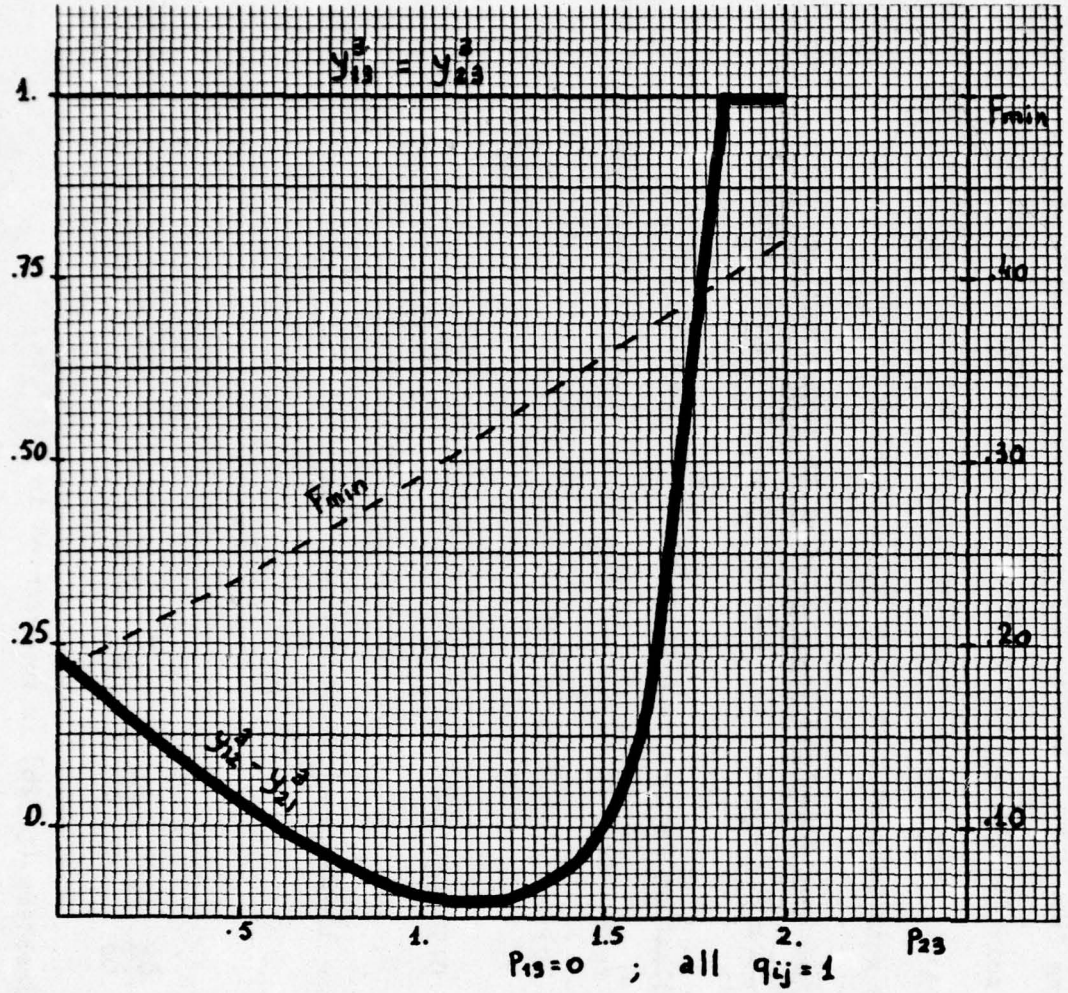


Fig. 5.8. Example 5.1. Routing variables and F_{min} in terms of p_{23}.

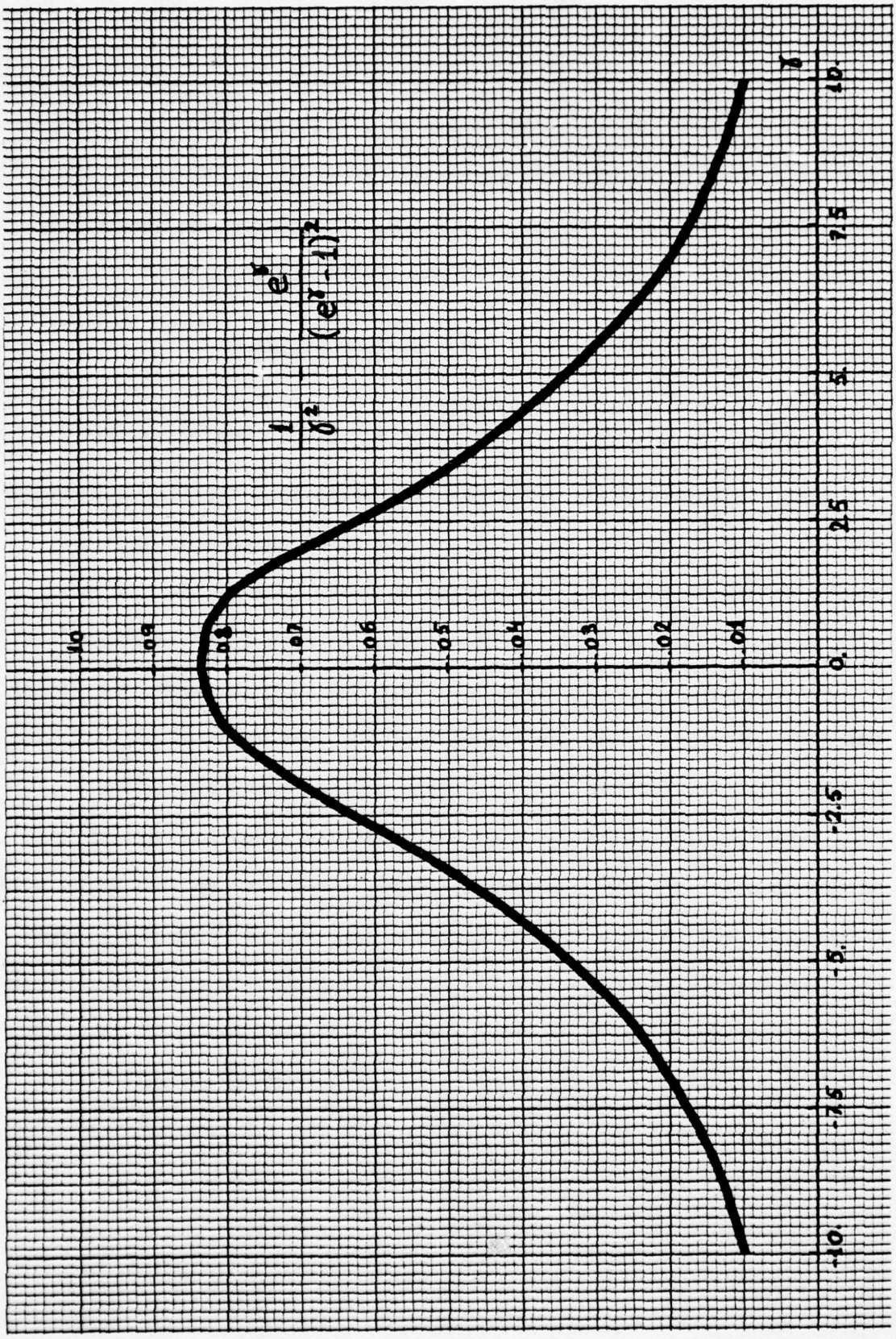


Fig. 5.8a. Variation of the first derivative of the average queue length in terms of γ

For $p_{23} = 0$ ($\delta_2 = -1$) we obtain $z^* = 0.24$ and thus $\gamma_{13} = -5.49$ and $\gamma_{23} = -3.75$ so that the expression of dF/dz in (5-32a) is null.

If p_{23} increases, say $p_{23} = 0.2$ ($\delta_2 = -0.8$) for the same value of $z = 0.24$ the values of γ_{13} are from (5-30) and (5-31):

$$\gamma_{13} = -5.30 \quad ; \quad \gamma_{23} = -3.08 \text{ yielding}$$

$$\frac{dF}{dz} = 0.0066 > 0 \quad \text{for } z = 0.24 \quad \text{and } \delta_2 = -0.8$$

The derivative dF/dz being positive means that the length increase rate in queue $x_{23}(t)$ due to the input increase, is greater than the corresponding to queue $x_{13}(t)$ so that in order to minimize the total effect, more messages are sent from node 2 to node 1 (z^* decreases)

As p_{23} keeps on increasing both γ_{13} and γ_{23} increase, the latter one more (see (5-30) and (5-31)) so that for some value of δ_2, γ_{23} becomes positive whereas γ_{13} increases too but remains negative.

From Fig 5.8a it can be seen that for this situation $\partial F/\partial\gamma_{13}$ increases and $\partial F/\partial\gamma_{23}$ decreases for δ_2 increasing so that dF/dz will be negative and z^* will have to increase in order to reach the minimum.

For example for $p_{23} = 1.2$ ($\delta_2 = 0.2$) the optimum z^* is $z^* = -0.10$ and the corresponding γ_{13} and γ_{23} are from (5-30) and (5-31).

$\gamma_{13} = -3.36$ $\gamma_{23} = -0.54$ so that dF/dz is null

If p_{23} increases: $p_{23} = 1.4$ ($\delta_2 = 0.4$) for the same value of $z = -0.10$ tha values of γ_{13} and γ_{23} are

$\gamma_{13} = -3.17$, $\gamma_{23} = 0.13$ yielding

$\frac{dF}{dz} = -0.0080 < 0$ for $z = -0.10$ and $\delta_2 = 0.4$

The derivative dF/dz being negative means that the length increase rate due to the increase of p_{23} is greater in queue $x_{13}(t)$ than in queue $x_{23}(t)$ and therefore more messages have to be sent from node 1 to node 2 (z^* increases). This change of behavior of z^* occurs when queue $x_{23}(t)$ becomes too loaded (above half the full-capacity). Ehen this is the situation and the input traffic at node 2 increases more, the increase of the queue length is more reflected on queue $x_{13}(t)$ (which is below half the full-capacity) than in $x_{23}(t)$ precisely because of the upper barrier for the queues which causes the corresponding queue length to be of the form shown by Fig 4.4 and therefore for $\gamma_i > 0$ the rate of length increases is slower.

The next figures 5-9 to 5-14 are shown for the same example of Fig. 5.2. but for other values of U_{12}^3 and U_{21}^3, that is for $A_{13} = 0.8$; $A_{23} = 0.4$, $U_{13}^3 = 0.6$, $U_{23}^3 = 0.6$, $U_{12}^3 = 0.5$ and $U_{21}^3 = 0.3$ and the same value for all $C_{ij} = 0.75$

The covariance matrix per unit time is now

$$\underline{\Lambda} = \begin{bmatrix} 0.86 & -0.46 \\ -0.46 & 0.94 \end{bmatrix}$$

and

$$\underline{V} = \begin{bmatrix} -3.1501 & 1.5416 \\ 1.5416 & -2.8820 \end{bmatrix}$$

And the utilization factors

$\rho_{13} = 0.73$, $\rho_{23} = 0.60$, greater than the former ones. For this reason the overall mean value is greater for this example

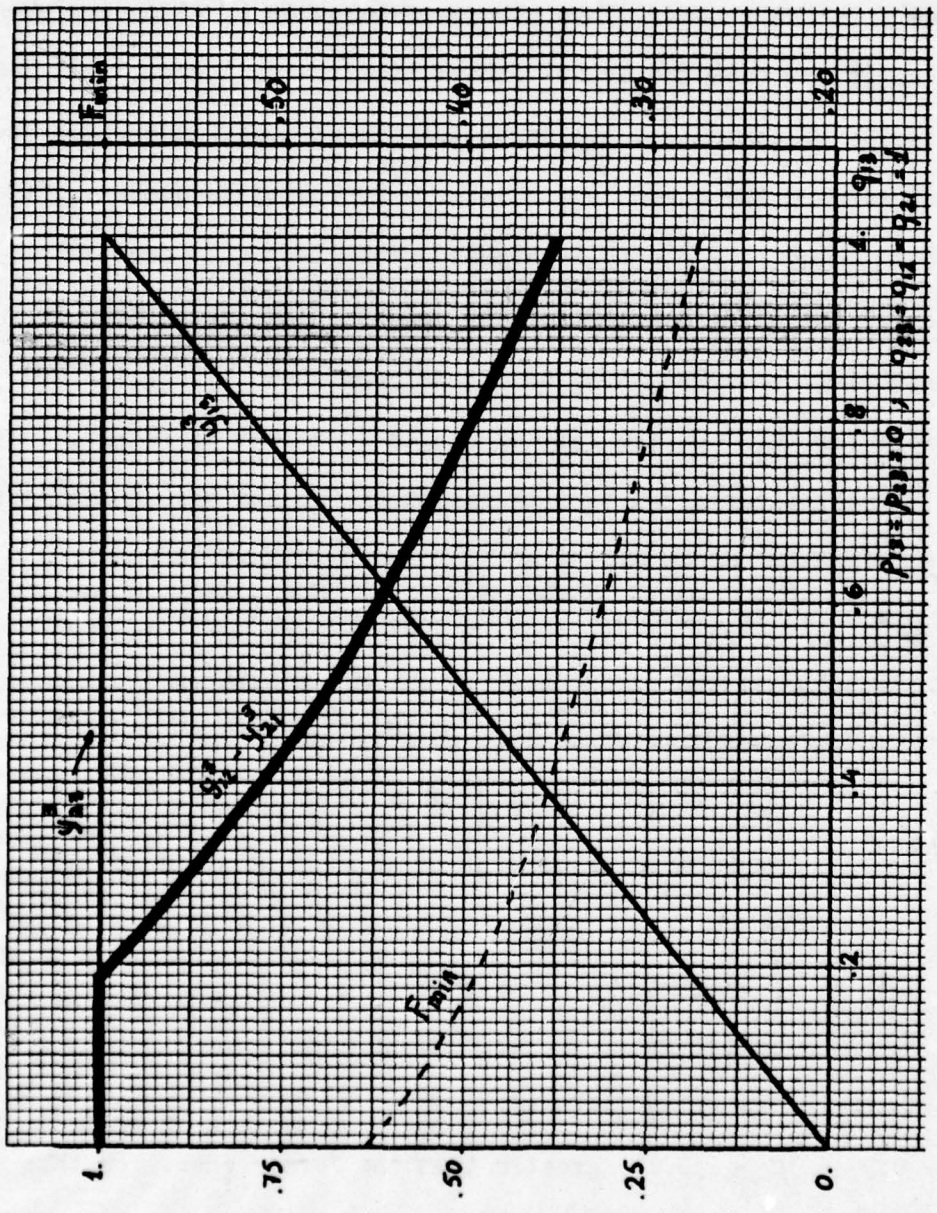


Fig. 5.9. Example 5.1. Routing variables and F_{min} in terms of q_{13}.

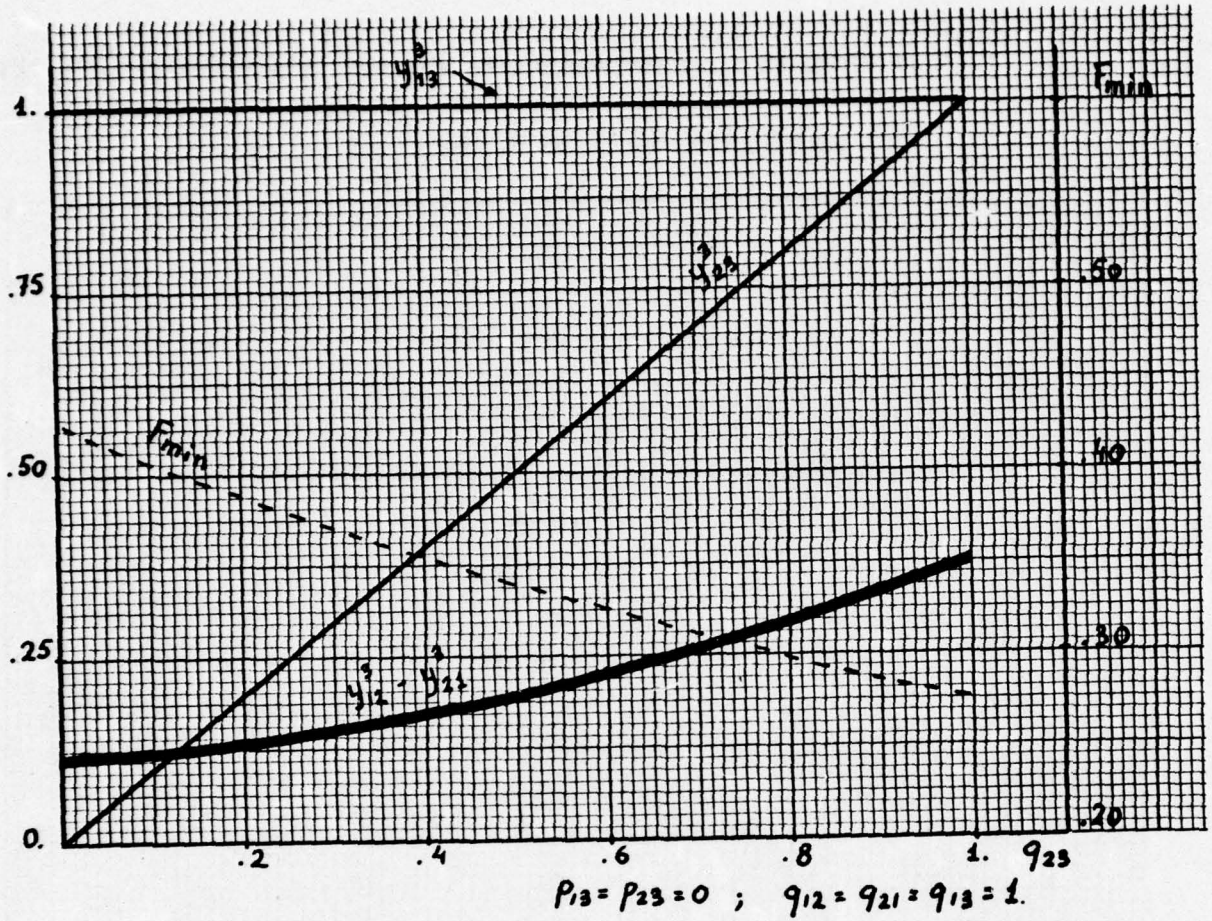


Fig. 5.10. Example 5.1. Routing variables and F_{min} in terms of q_{23}

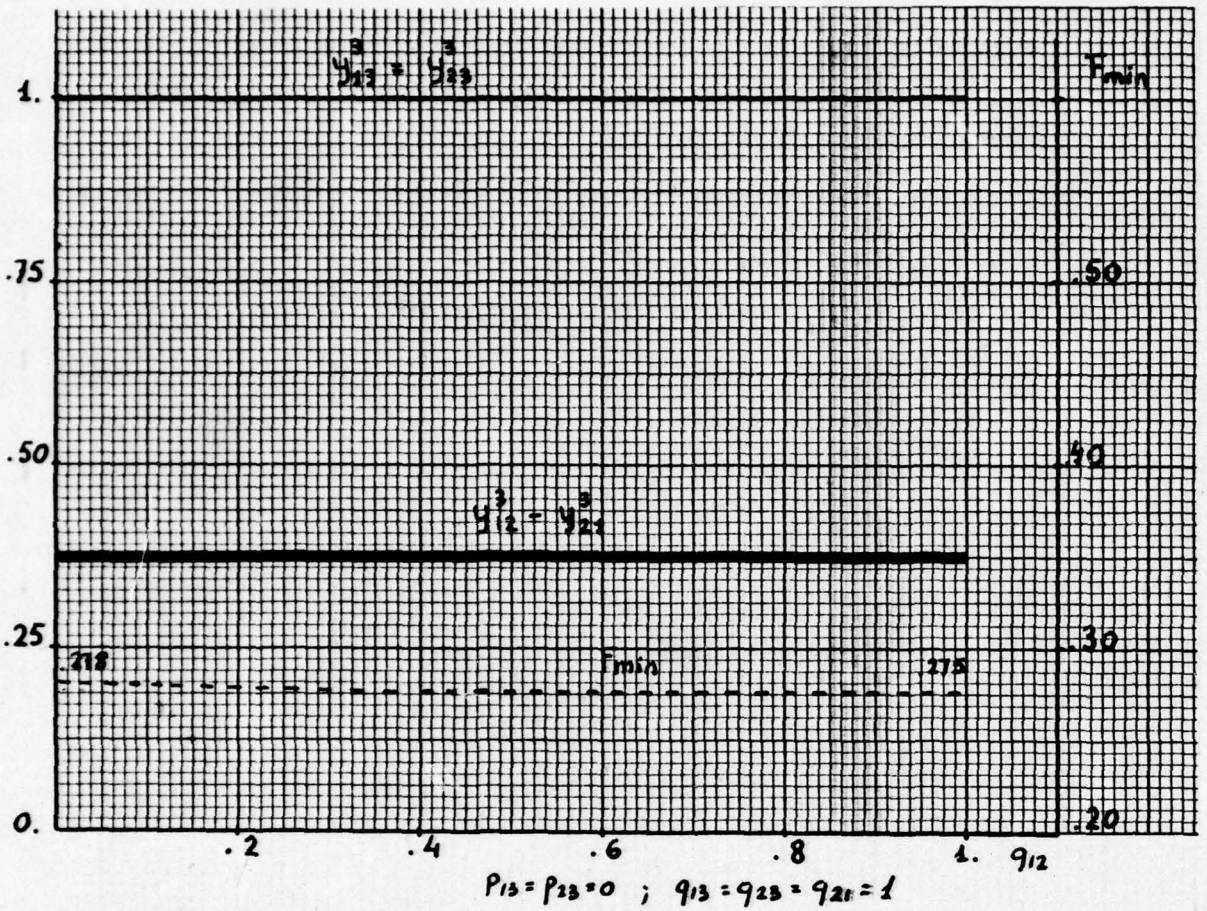


Fig. 5.11. Example 5.1. Routing variables and F_{min} in terms of q_{12}.

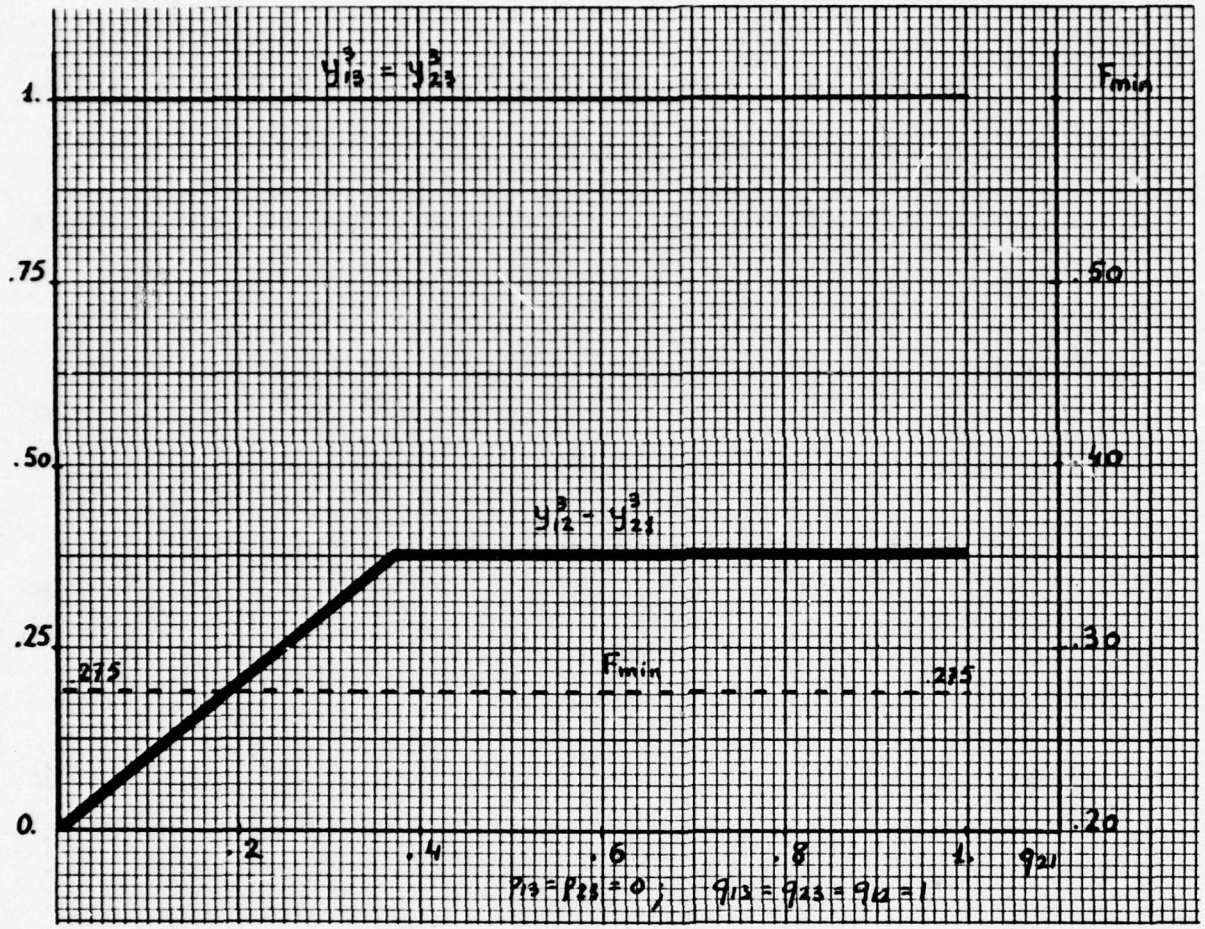


Fig. 5.12. Example 5.1. Routing variables and F_{min} in terms of q_{21}.

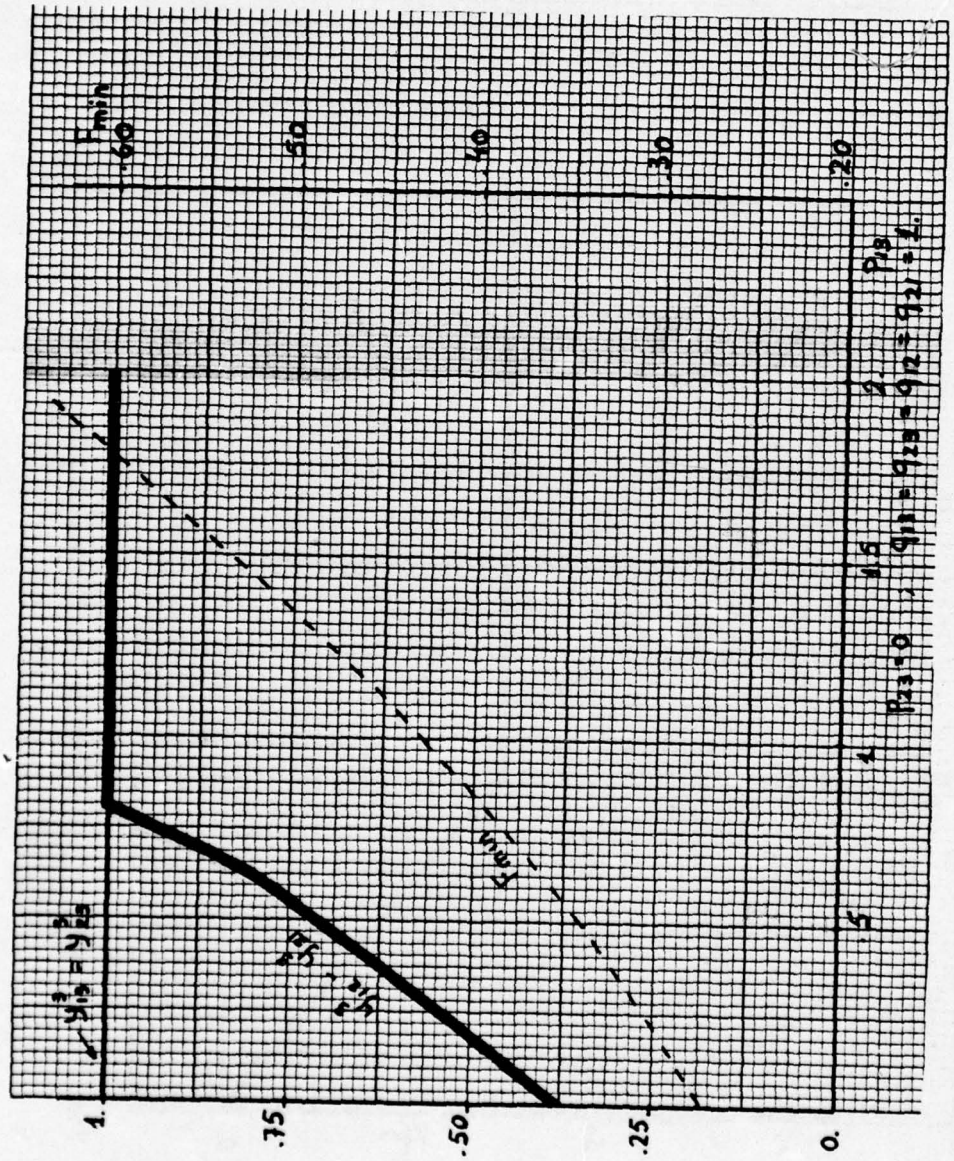


Fig. 5.13. Example 5.1. Routing variables and F_{min} in terms of P_{13}.

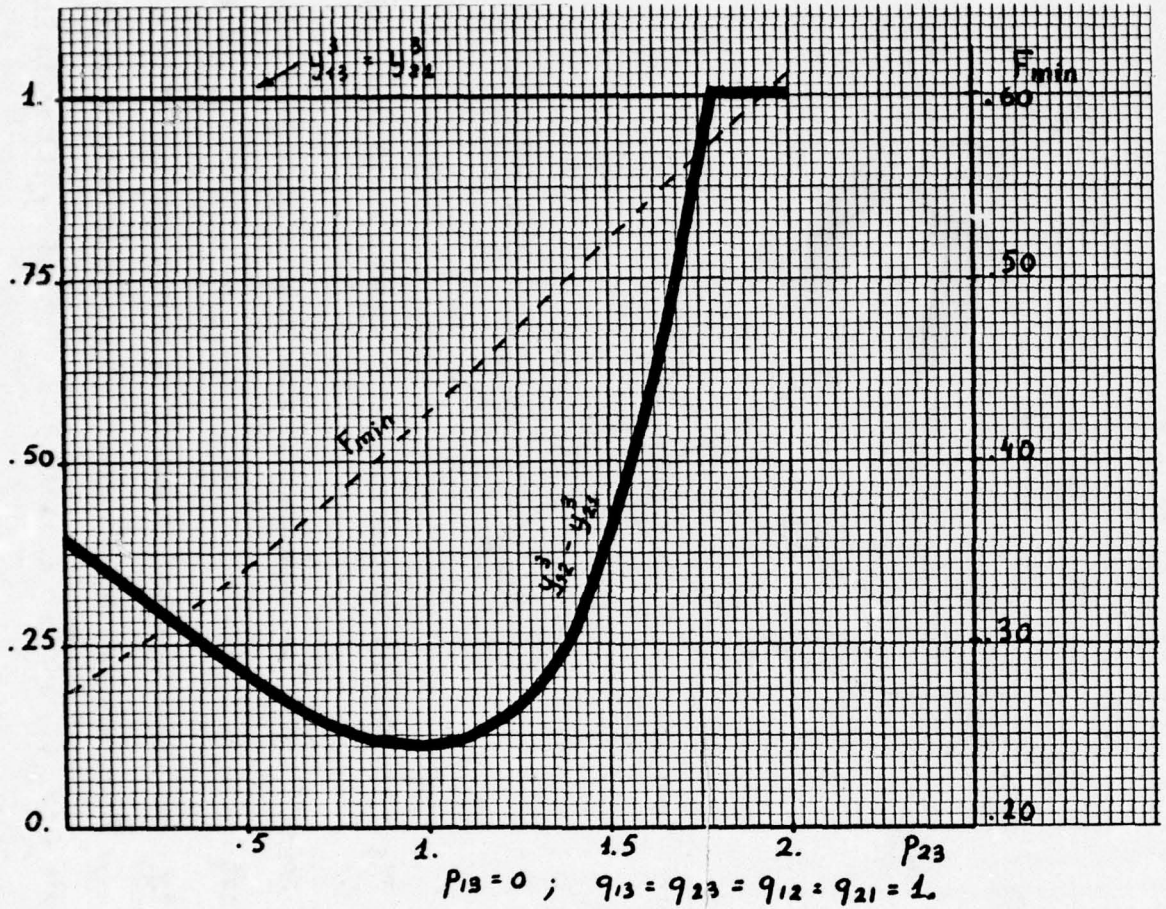


Fig. 5.14. Example 5.1. Routing variables and F_{min} in terms of p_{23}.

5.2.- Example with four queues

Assume a 3-node network as shown in Fig. 5.15

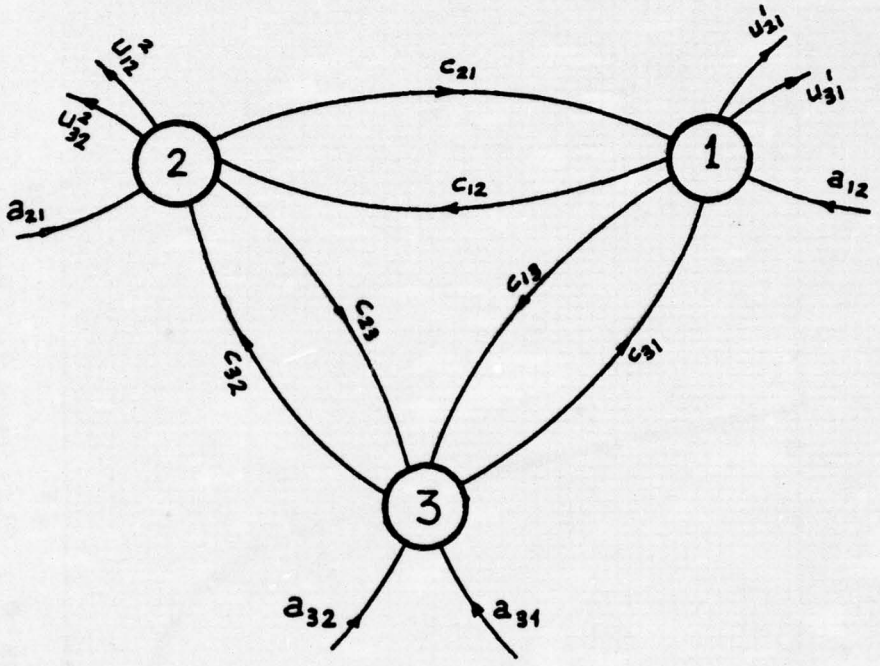


Fig 5.15: Network of three nodes, 3 sources and 2 destinations

Nodes 1 and 2 are sources and receivers and node 3 is only source so that messages a_{12} go from 1 to 2 via c_{12} or c_{13} and c_{32} and similarly for a_{21}. At node 3 messages go either to node 1 via c_{31} or to node 2 via c_{32}. Therefore we have one queueing process at node 1: $x_{12}(t)$, one at node 2: $x_{21}(t)$ and two at nodes 3 $x_{31}(t)$ and $x_{32}(t)$ as in Fig 5.16.

According to section 4.4 the capacities and traffic rates are of the form specified in (4-28), (4-29) and (4-30). The system of equations (4-31) is now

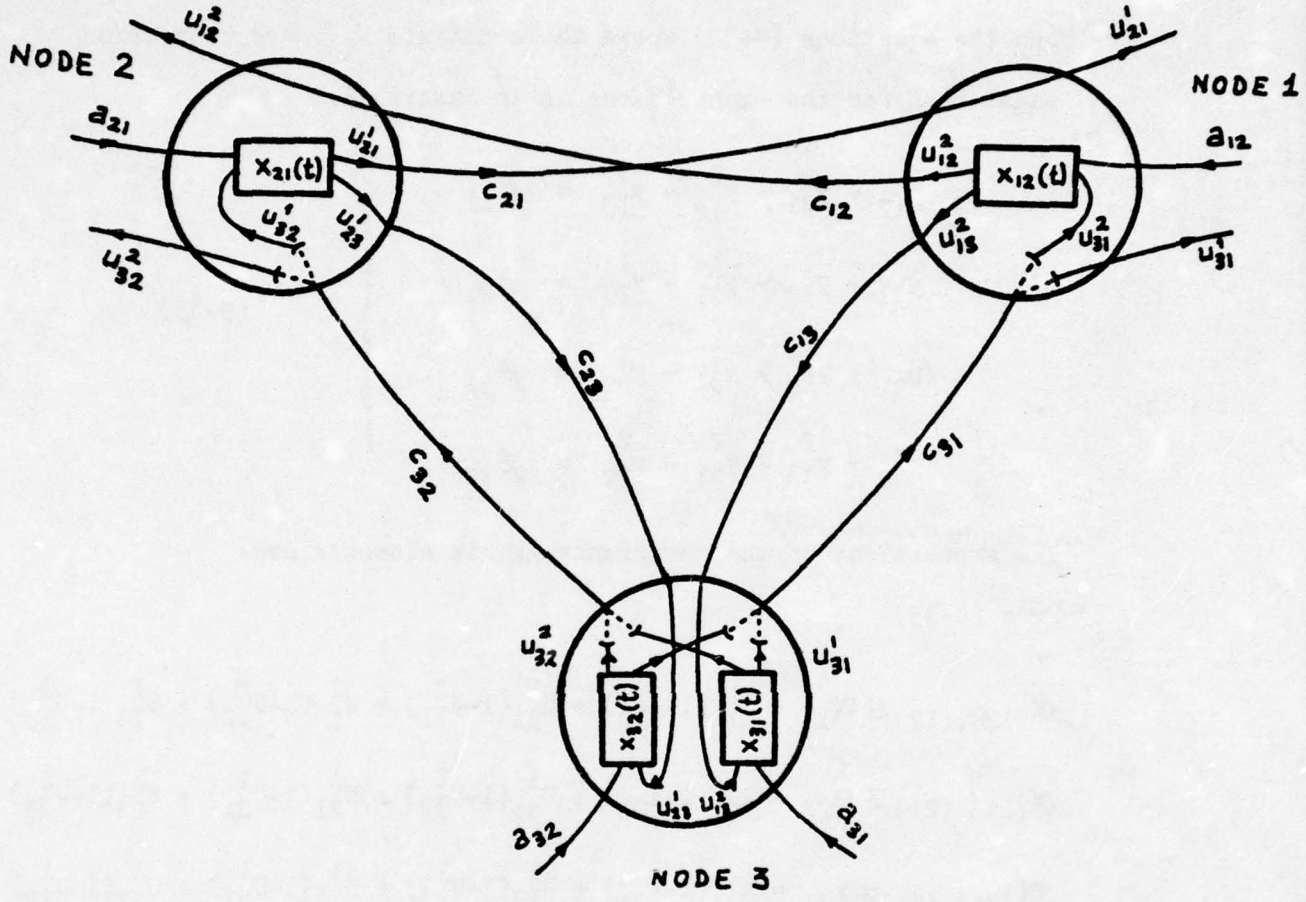


Fig 5.16: Network of three nodes. Detail of queues at each node.

$$\left.\begin{aligned} A_{12} + U_{31}^2 - U_{12}^2 - U_{13}^2 &= 0 \\ A_{21} + U_{32}^1 - U_{21}^1 - U_{23}^1 &= 0 \\ A_{31} + U_{23}^1 - U_{31}^1 - U_{32}^1 &= 0 \\ A_{32} + U_{13}^2 - U_{31}^2 - U_{32}^2 &= 0 \end{aligned}\right\} \qquad (5\text{-}32)$$

and the equations (4-32) where the constants K_{ij} have been taken equal to 1 for the same reasons as in Section 5.1

$$\left.\begin{aligned} P_{12} + y^2_{31} - y^2_{12} - y^2_{13} &= \beta_{12} \\ P_{21} + y^1_{32} - y^1_{21} - y^1_{23} &= \beta_{21} \\ P_{31} + y^1_{23} - y^1_{31} - y^1_{32} &= \beta_{31} \\ P_{32} + y^2_{13} - y^2_{31} - y^2_{32} &= \beta_{32} \end{aligned}\right\} \qquad (5\text{-}33)$$

The expressions of the covariance matrix elements are:

From (4-35)

$$\alpha_{(12),(12)} \equiv \alpha_{11} = A_{12}(1-A_{12}) + U^2_{31}(1-U^2_{31}) + U^2_{12}(1-U^2_{12}) + U^2_{13}(1-U^2_{13})$$

$$\alpha_{(21),(21)} \equiv \alpha_{22} = A_{21}(1-A_{21}) + U^1_{32}(1-U^1_{32}) + U^1_{21}(1-U^1_{21}) + U^1_{23}(1-U^1_{23})$$

$$\alpha_{(31),(31)} \equiv \alpha_{33} = A_{31}(1-A_{31}) + U^1_{23}(1-U^1_{23}) + U^1_{31}(1-U^1_{31}) + U^1_{32}(1-U^1_{32})$$

$$\alpha_{(32),(32)} \equiv \alpha_{44} = A_{32}(1-A_{32}) + U^2_{13}(1-U^2_{13}) + U^2_{31}(1-U^2_{31}) + U^2_{32}(1-U^2_{32})$$

From (4-36)

$$\alpha_{(31),(32)} \equiv \alpha_{34} = -U^1_{31}\,U^2_{31} - U^1_{32}\,U^2_{32}$$

From (4-37)

$$\alpha_{(21),(31)} \equiv \alpha_{23} = -U^1_{23}(1 - U^1_{23}) - U^1_{32}(1 - U^1_{32})$$

$$\alpha_{(12),(32)} \equiv \alpha_{14} = -U^2_{13}(1 - U^2_{13}) - U^2_{31}(1 - U^2_{31})$$

From (4-38)

$$\alpha_{(12),(21)} \equiv \alpha_{12} = 0$$

$$\alpha_{(12),(31)} \equiv \alpha_{13} = U_{31}^{1}\ U_{31}^{2}$$

$$\alpha_{(21),(32)} \equiv \alpha_{24} = U_{32}^{2}\ U_{32}^{1}$$

Then

$$\underline{\Lambda} = \begin{bmatrix} \alpha_{11} & 0 & \alpha_{13} & \alpha_{14} \\ 0 & \alpha_{22} & \alpha_{23} & \alpha_{24} \\ \alpha_{13} & \alpha_{23} & \alpha_{33} & \alpha_{34} \\ \alpha_{14} & \alpha_{24} & \alpha_{34} & \alpha_{44} \end{bmatrix} \qquad (5\text{-}34)$$

with the former expressions for each element.

To minimize (4-39) we have to take into account the relationship among the γ_{ij}. From (5-33):

$$\beta_{12} + \beta_{21} + \beta_{31} + \beta_{32} = (p_{12}-y_{12}^{2}) + (p_{21}-y_{21}^{1}) + (p_{31}-y_{31}^{1}) + (p_{32}-y_{32}^{2}) \qquad (5\text{-}35)$$

From (3-7) : $\underline{\beta} = \frac{1}{2}\ \underline{\Lambda}\ \underline{\gamma}$

Therefore (5-35) becomes

$$\left(\frac{1}{2}\sum_{i=1}^{4}\alpha_{i1}\right)\gamma_{12} + \left(\frac{1}{2}\sum_{i=1}^{4}\alpha_{i2}\right)\gamma_{21} + \frac{1}{2}\left(\sum_{i=1}^{4}\alpha_{i3}\right)\gamma_{31} +$$

$$+ \frac{1}{2}\left(\sum_{i=1}^{4} \alpha_{i4}\right)\gamma_{32} = \tag{5-36}$$

$$= (p_{12} - y_{12}^2) + (p_{21} - y_{21}^1) + (p_{31} - y_{31}^1)(p_{32} - y_{32}^2)$$

The expression (5-36) is a hyperplane in the γ-space and the optimum $\underline{\gamma}$ has to be on it. From (5-34) the coefficients of γ_{ij} in (5-36) are non-negative so that the constraint (5-36) will have a smaller minimum when the right hand side is more negative, that is when y_{12}^1, y_{21}^1, y_{31}^1 and y_{32}^2 take on their maximum value. From the capacity constraints (4-34)

$$\begin{aligned} y_{12}^2 &\leqslant q_{12} \; ; \qquad y_{21}^1 \leqslant q_{21} \\ y_{31}^1 + y_{31}^2 &\leqslant q_{31} \; ; \; y_{32}^2 + y_{32}^1 \leqslant q_{32} \end{aligned} \tag{5-38}$$

Therefore the maximum values are:

$$\begin{aligned} (y_{12}^2)_{\max} &= q_{12} \; ; \quad (y_{21}^1)_{\max} = q_{21} \\ (y_{31}^1)_{\max} &= q_{31} \; ; \quad (y_{32}^2)_{\max} = q_{32} \end{aligned} \tag{5-39}$$

and from (5-38) and (5-39)

$$y_{31}^2 = y_{32}^1 = 0 \tag{5-40}$$

Thus we are left in the minimization of (4-39) over the eight components of $\underline{y}$ with only two of them y_{13}^2 and y_{23}^1, the other six being given by (5-39) and (5-40).

The expression of the elements of $\underline{\beta}$ (5-33) becomes then

$$\left.\begin{aligned} \beta_{12} &= (p_{12} - q_{12}) - y_{13}^{2} \\ \beta_{21} &= (p_{21} - q_{21}) - y_{23}^{1} \\ \beta_{31} &= (p_{31} - q_{31}) + y_{23}^{1} \\ \beta_{32} &= (p_{32} - q_{32}) + y_{13}^{2} \end{aligned}\right\} \qquad (5\text{-}41)$$

expressions similar to (5-22) of the former example.

In this case the expression of the elements of $\underline{V} = 2\,\underline{\Lambda}^{-1}$ is not so straightforward because it is a 4 x 4 matrix.

Let us take some values for the traffic rates and capacities. Assume:

$A_{12} = 0.85$; $U_{12}^{2} = 0.75$; $U_{13}^{2} = 0.40$

$A_{21} = 0.60$; $U_{21}^{1} = 0.70$; $U_{23}^{1} = 0.25$

$A_{31} = 0.45$; $U_{31}^{1} = 0.35$; $U_{31}^{2} = 0.30$

$A_{32} = 0.25$; $U_{32}^{2} = 0.35$; $U_{32}^{1} = 0.35$

and $C_{ij} = 0.75$ for all (ij) which satisfy the flow equations (5-32) and the capacity constraints.

The covariance matrix is then:

$$\underline{\Lambda} = \begin{bmatrix} 0.765 & 0 & 0.105 & -0.450 \\ 0 & 0.865 & -0.415 & 0.2275 \\ 0.105 & -0.415 & 0.890 & -0.3325 \\ -0.450 & 0.2275 & -0.3325 & 0.9250 \end{bmatrix} \qquad (5\text{-}42)$$

and the corresponding $\underline{V} = 2\underline{\Lambda}^{-1}$:

$$\underline{V} = \begin{bmatrix} 3.7683 & -0.4881 & 0.0665 & 1.9772 \\ -0.4881 & 3.0710 & 1.2922 & -0.5283 \\ 0.0665 & 1.2922 & 3.1596 & 0.8503 \\ 1.9772 & -0.5283 & 0.8503 & 3.5596 \end{bmatrix} \qquad (5\text{-}43)$$

In (4-41) call for convenience

$$\begin{aligned} p_{12} - q_{12} &= \delta_1 & \qquad p_{21} - q_{21} &= \delta_2 \\ p_{31} - q_{31} &= \delta_3 & \qquad p_{32} - q_{32} &= \delta_4 \end{aligned} \qquad (5\text{-}44)$$

from (3-7) and (5-41) we have

$$\left.\begin{aligned} \gamma_{12} &= 3.7683\ \delta_1 - 0.4881\ \delta_2 + 0.0665\ \delta_3 + 1.9772\ \delta_4 - \\ &\quad - 1.7911\ y^2_{13} + 0.5546\ y^1_{23} \\ \gamma_{21} &= -0.4881\ \delta_1 + 3.0710\ \delta_2 + 1.2922\ \delta_3 - 0.5283\ \delta_4 - \\ &\quad - 0.0402\ y^2_{13} - 1.7788\ y^1_{23} \\ \gamma_{31} &= 0.0665\ \delta_1 + 1.2922\ \delta_2 + 3.1596\ \delta_3 + 0.8503\ \delta_4 + \\ &\quad + 0.7838\ y^2_{13} + 1.8674\ y^1_{23} \\ \gamma_{32} &= 1.9772\ \delta_1 - 0.5283\ \delta_2 + 0.8503\ \delta_3 + 3.5596\ \delta_4 + \\ &\quad + 1.5824\ y^2_{13} + 1.3786\ y^1_{23} \end{aligned}\right\} \qquad (5\text{-}45)$$

and the minimization of (4-39) is carried over.

In the next figures from 5.17 to 5.16 several results are shown when one δ_i varies, the others being held fixed:

Fig 5.17 : The effect of q_{12} varying between 0 and 1 is shown. It can be seen that when q_{12} decreases y_{12}^3 increases that is when the capacity of the channel 1 - 2 decreases more messages are sent from 1 to 2 via node 3.

Fig 5.18: Now q_{21} varies between 0 and 1. The effect when q_{21} decreases is to decrease the rate of messages that go from $1 \rightarrow 3 \rightarrow 2$, that is y_{13} and to increase the rate of messages from $2 \rightarrow 3 \rightarrow 1$ because the capacity q_{21} of the direct link decreases.

As earlier in section 5.3 the overall mean increases when the capacity decreases.

Fig 5.19: When q_{31} decreases y_{13}^2 decreases too in order to compensate for the increase of the drift in $x_{31}(t)$.

Fig 5.20: When q_{32} decreases y_{13}^2 has to decrease to compensate this effect.

Fig 5.21: The effect of q_{12} decreasing on y_{12}^3 is null until $q_{13} = 0.1701$. From this point on $y_{13}^2 = q_{13}$ and overall mean value F_{min} increases slightly. This effect is similar to that of Fig 5.5..

Fig 5.22: Similar to Fig 5.6. There is no effect on y_{13}^2 when q_{13} decreases.

Fig 5.23: Now the capacities q_{ij} are held fixed and the input p_{12}

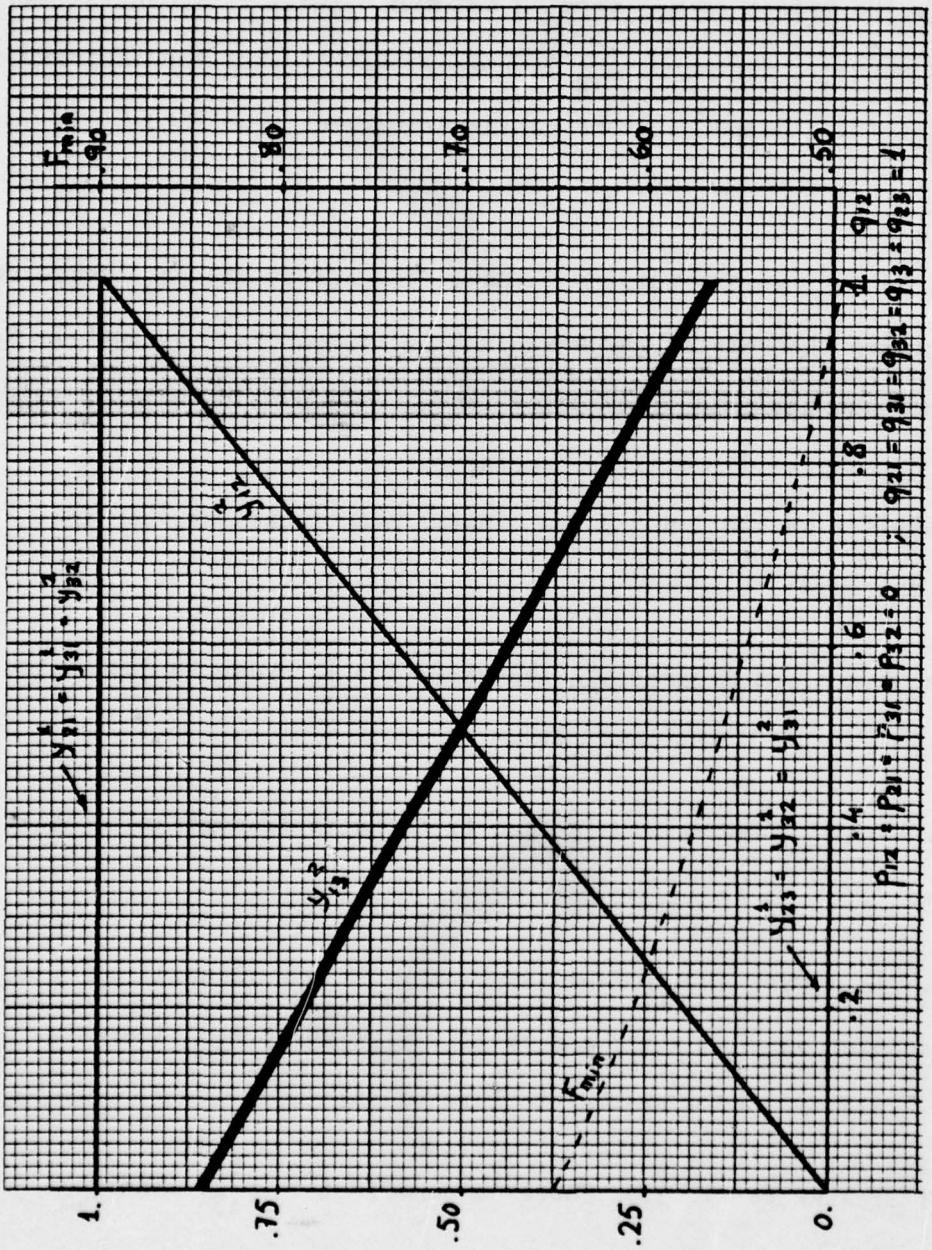


Fig. 5.17 Example 5.2. Routing variables and F_{min} in terms of q_{12}.

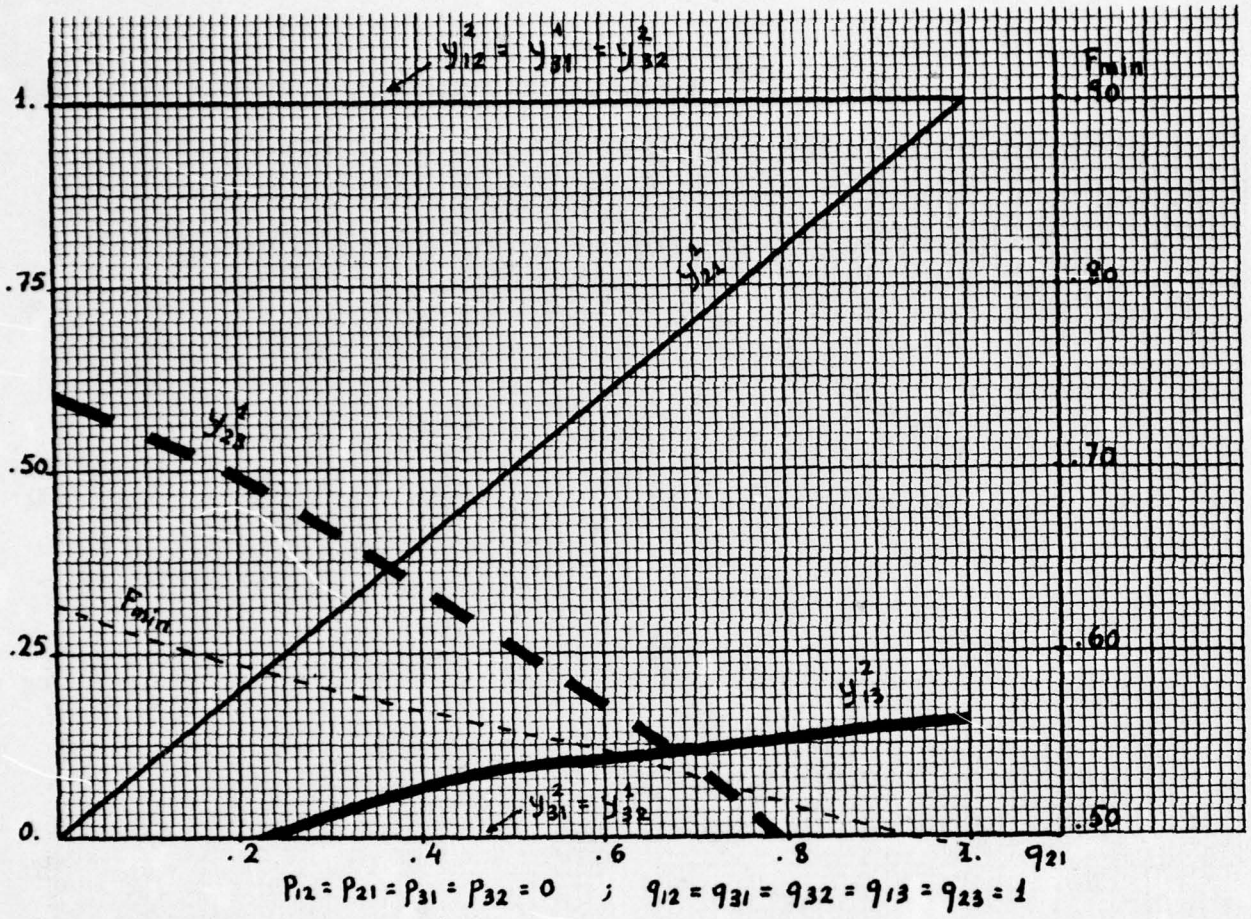


Fig. 5.18. Example 5.2. Routing variables and F_{min} in terms of q_{21}.

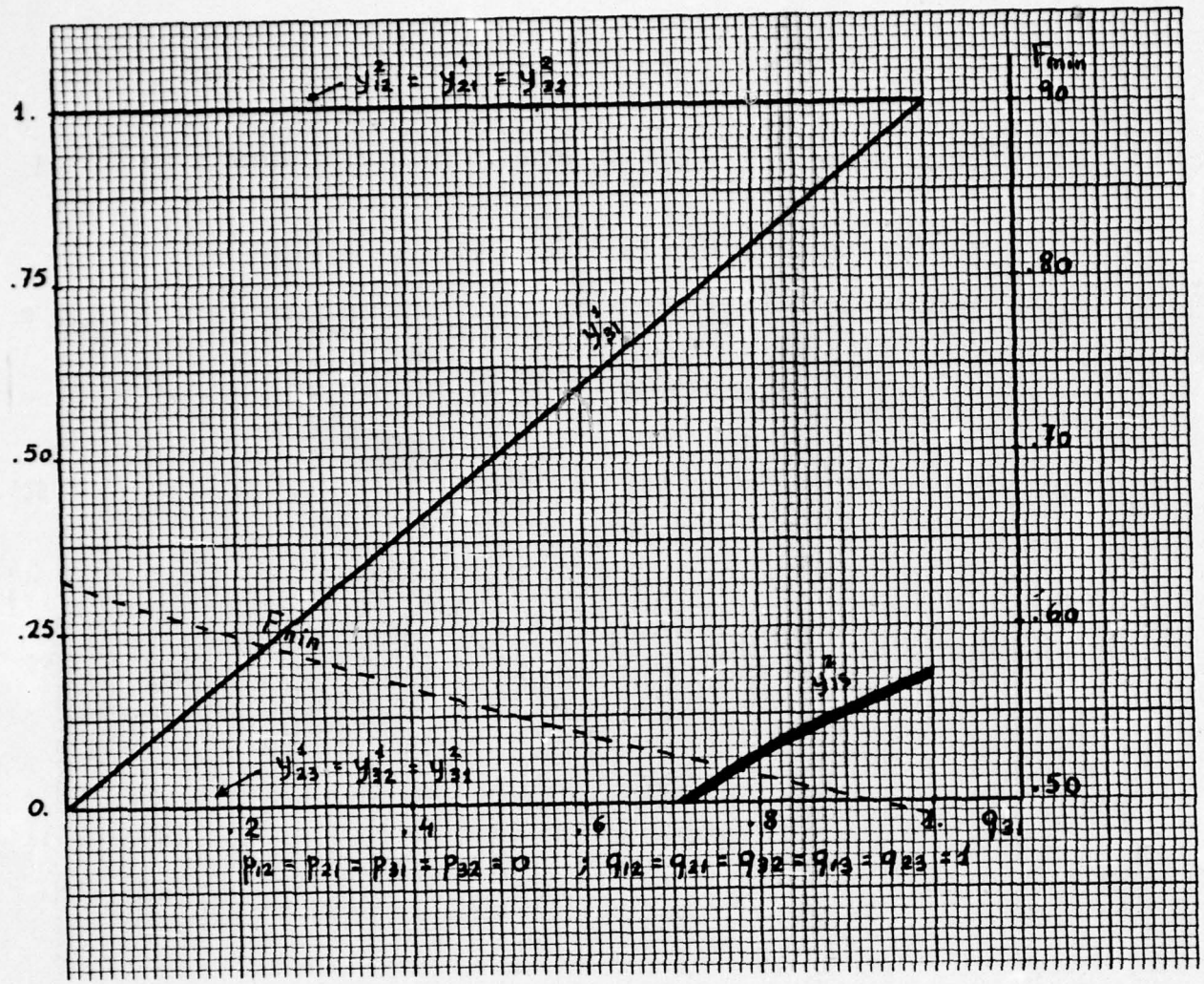


Fig. 5.19. Example 5.2. Routing variables and F_{min} in terms of q_{31}.

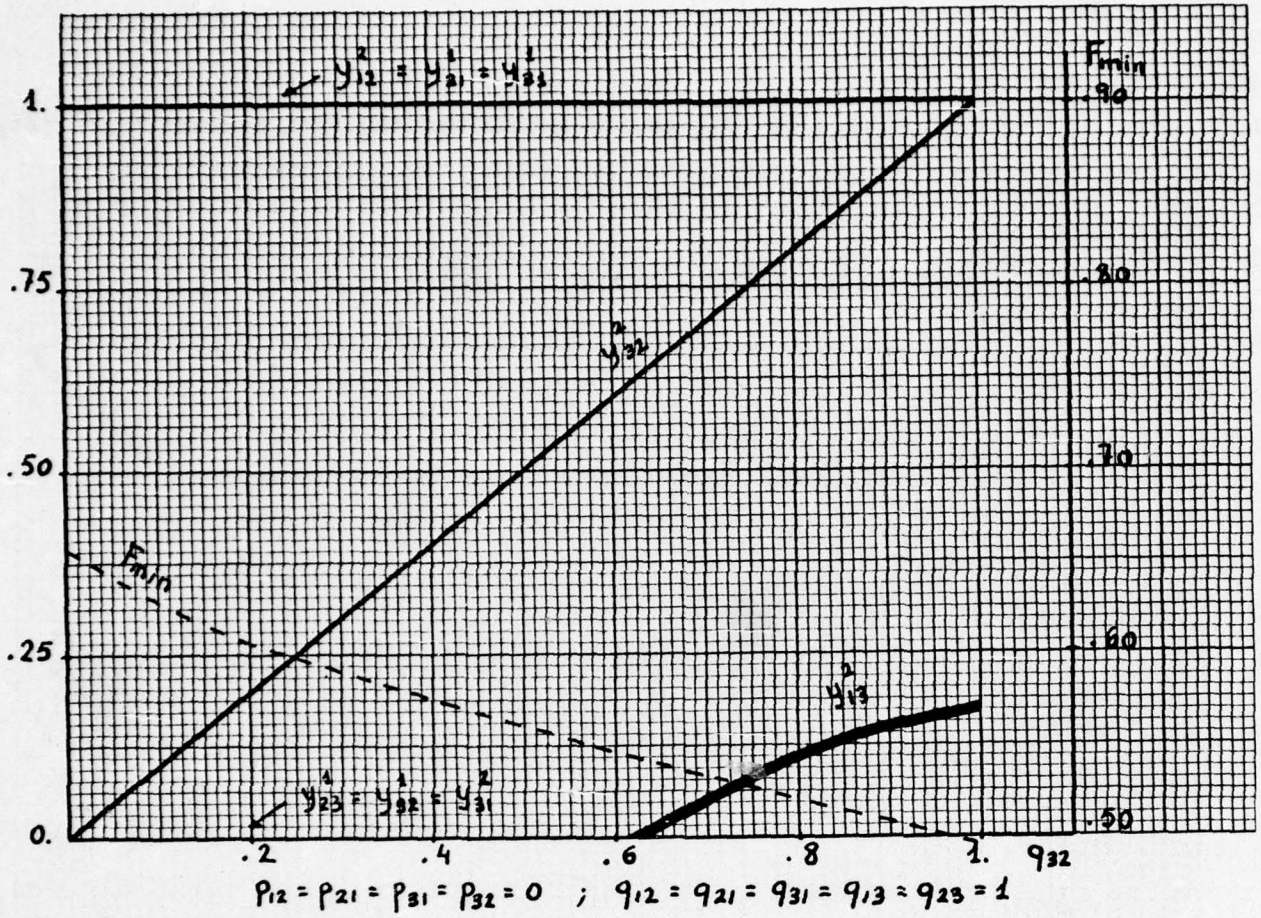


Fig. 5.20. Example 5.2. Routing variables and F_{min} in terms of q_{32}.

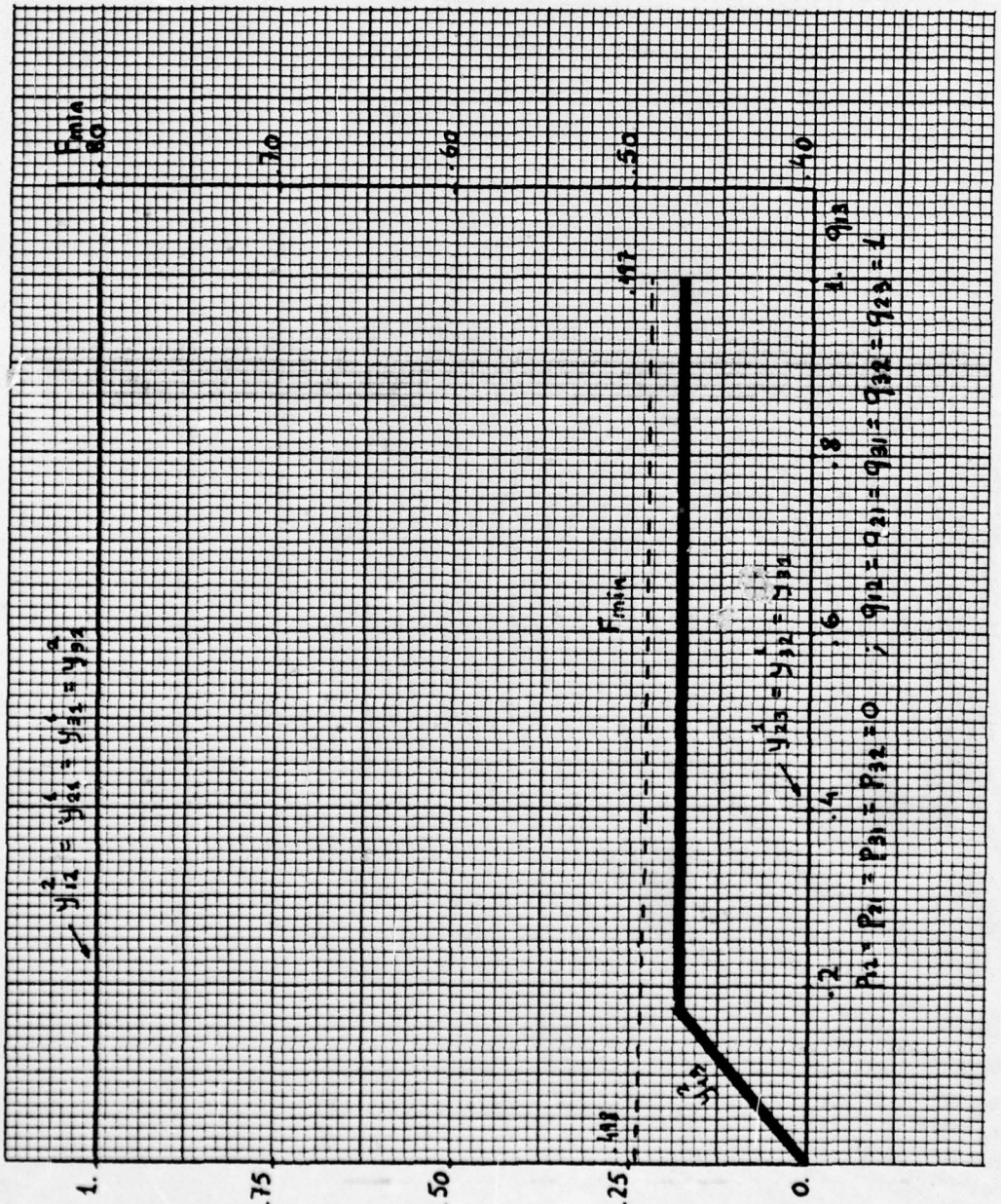


Fig. 5.21. Example 5.2. Routing variables and F_{min} in terms of q_{13}.

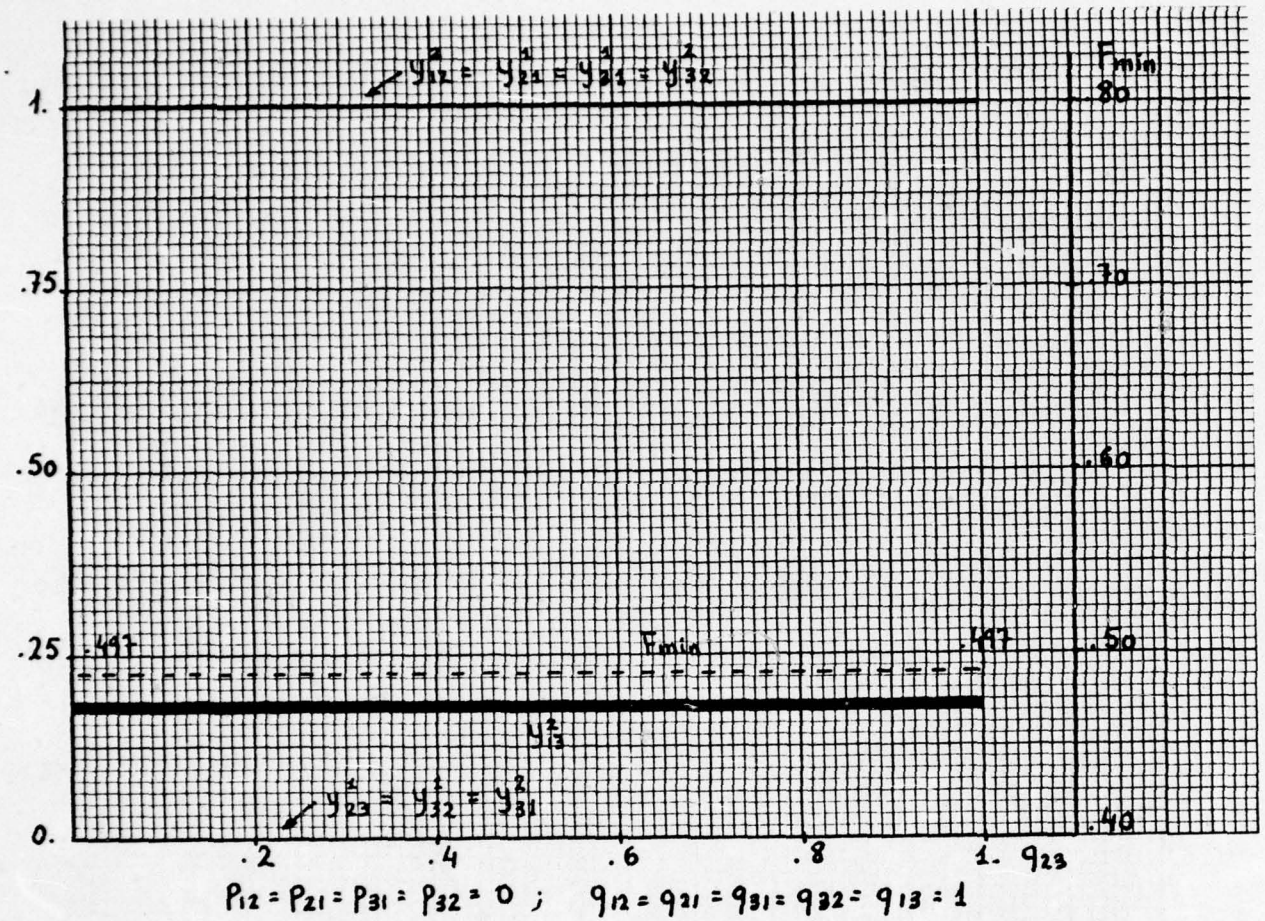


Fig. 5.22. Example 5.2. Routing variables and F_{min} in terms of q_{23}.

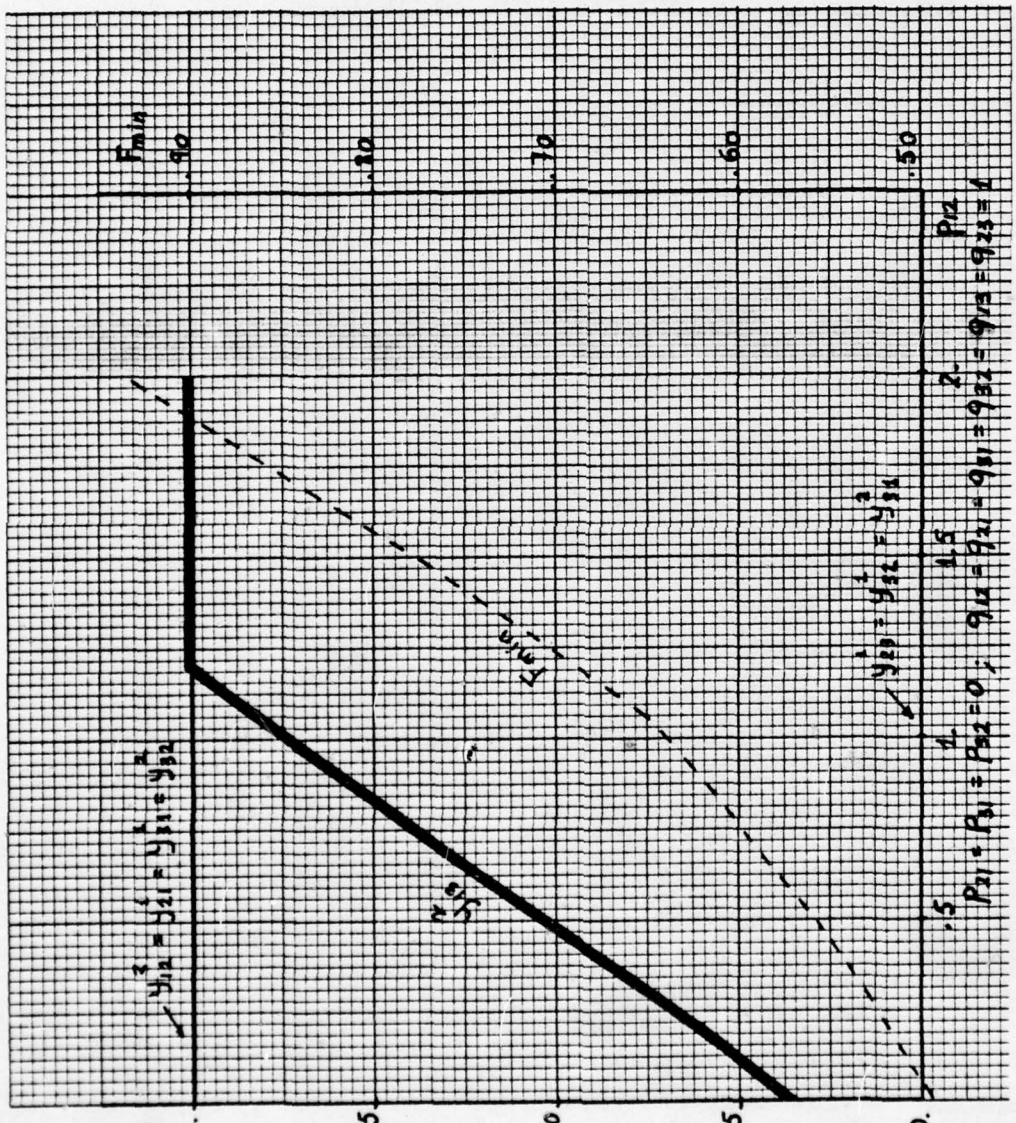


Fig. 5.23. Example 5.2. Routing variables and F_{min} in terms of p_{12}.

increases. For p_{12} between 0 and 1 the effect is the same as in Fig 5.17 when q_{12} varies between 1 and 0 that is y^2_{13} increases up to its maximum value $q_{13} = 1$. At this saturation point F_{min} increases faster.

Fig 5.24: The input p_{21} increases. The same effect as in Fig 5.18 for p_{21} between 0 and 1. Beyond $p_{21} = 1, y^1_{23}$ keeps on increasing up to some value and then decreases very fast. This effect is similar to the one was shown in Fig 5.8. for the former case

Fig 5.25 and 5.26: When p_{31} and p_{32} increase respectively. Similar to Figs. 5.19 and 5.20 for q_{31} and q_{32} decreasing. In both cases y^2_{13} decreases until it reaches 0 and $y^1_{23} = 0$.

5.3.- Diffusion approximation for computer load sharing.

The general model described in Section 4.1 for message routing in a computer network can be also applied to the load sharing problem among a system of computers. References for this topic are given in [26] and [21].

The goal in a loaded sharing system is to increase the processing capabilities of a single computer, so that if transmission channels interconnecting several computers are provided, the overall system performance can be increased by taking advantage of the computer different loads. For instance if one computer in the system becomes more loaded than the others, it makes sense to process some of its incoming messages at other distant computers that are less loaded. Therefore some messages arriving at computer A will be sent to some other computers, processed

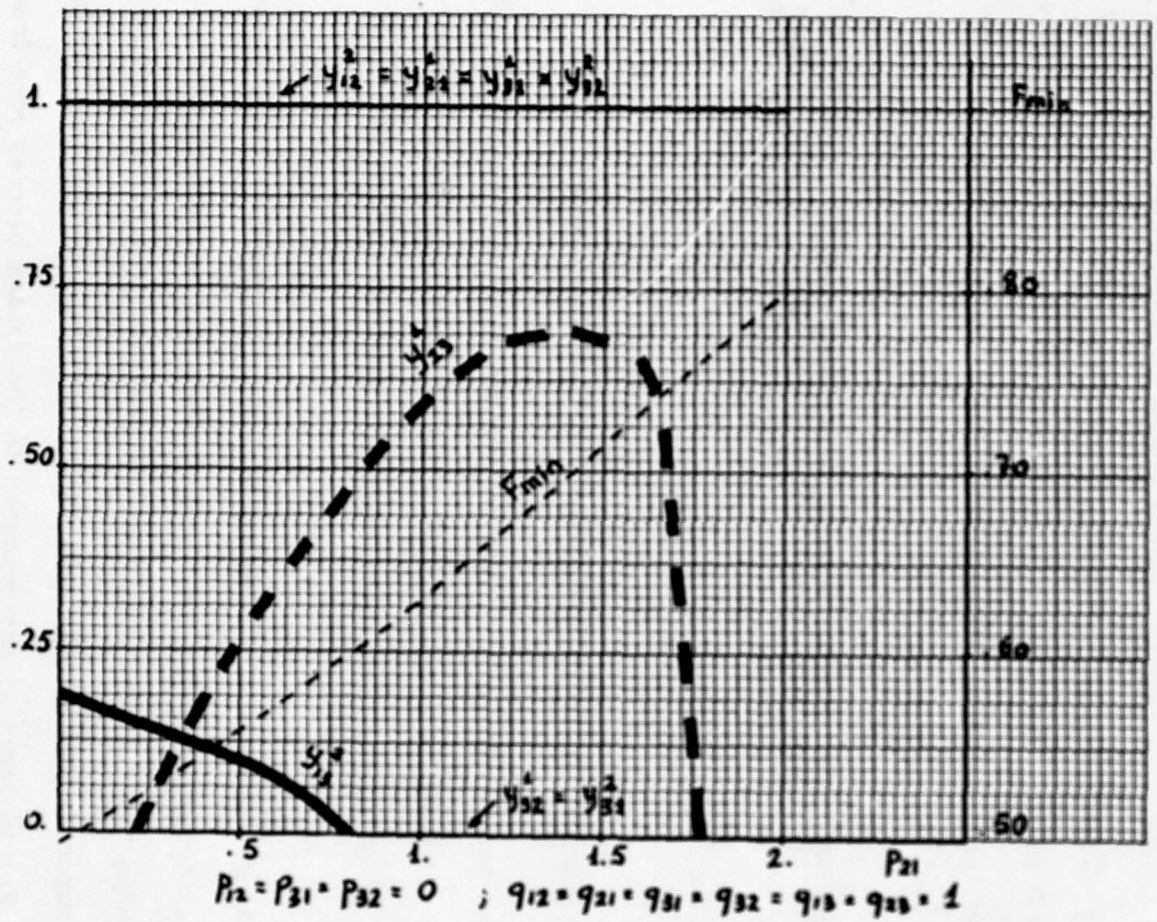


Fig. 5.24. Example 5.2. Routing variables and F_{min} in terms of p_{21}.

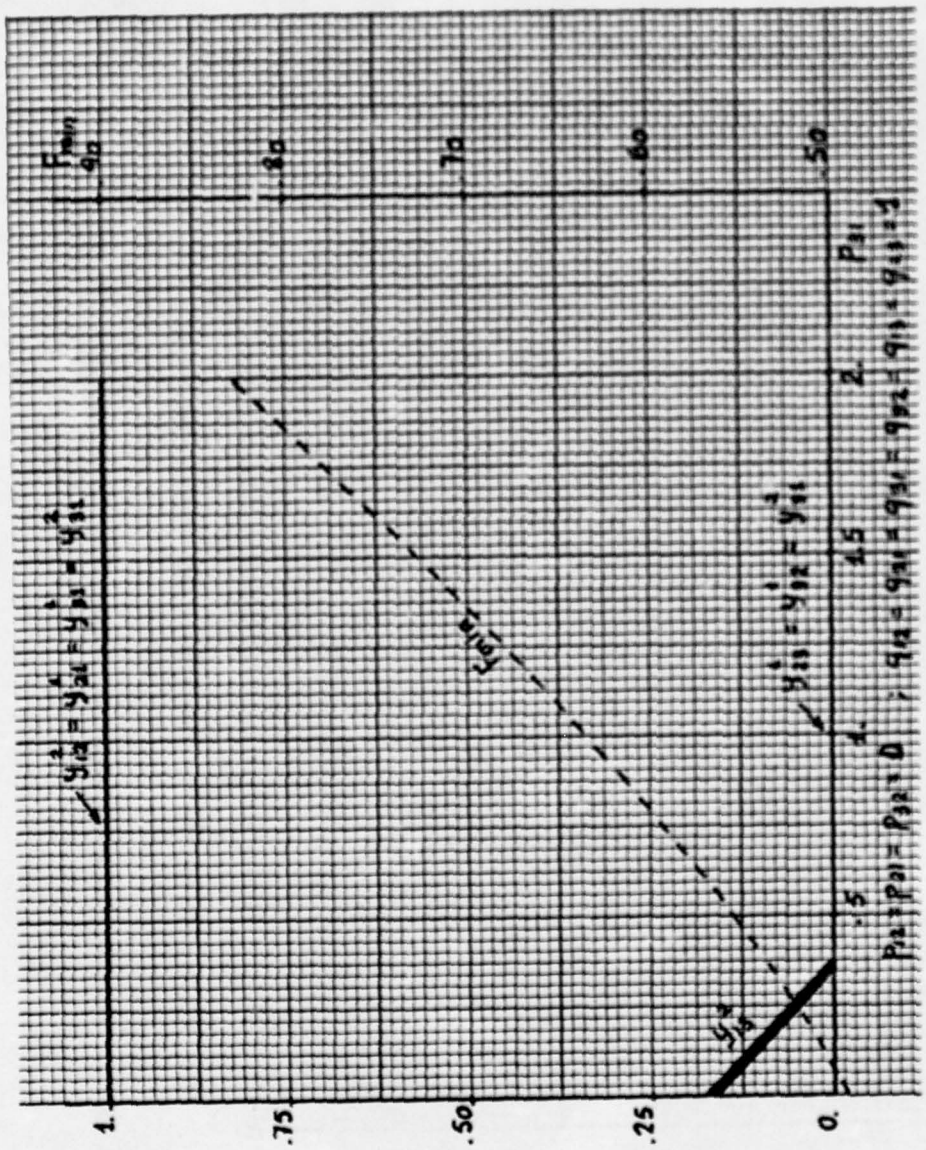


Fig. 5.25. Example 5.2. Routing variables and F_{min} in terms of P_{31}.

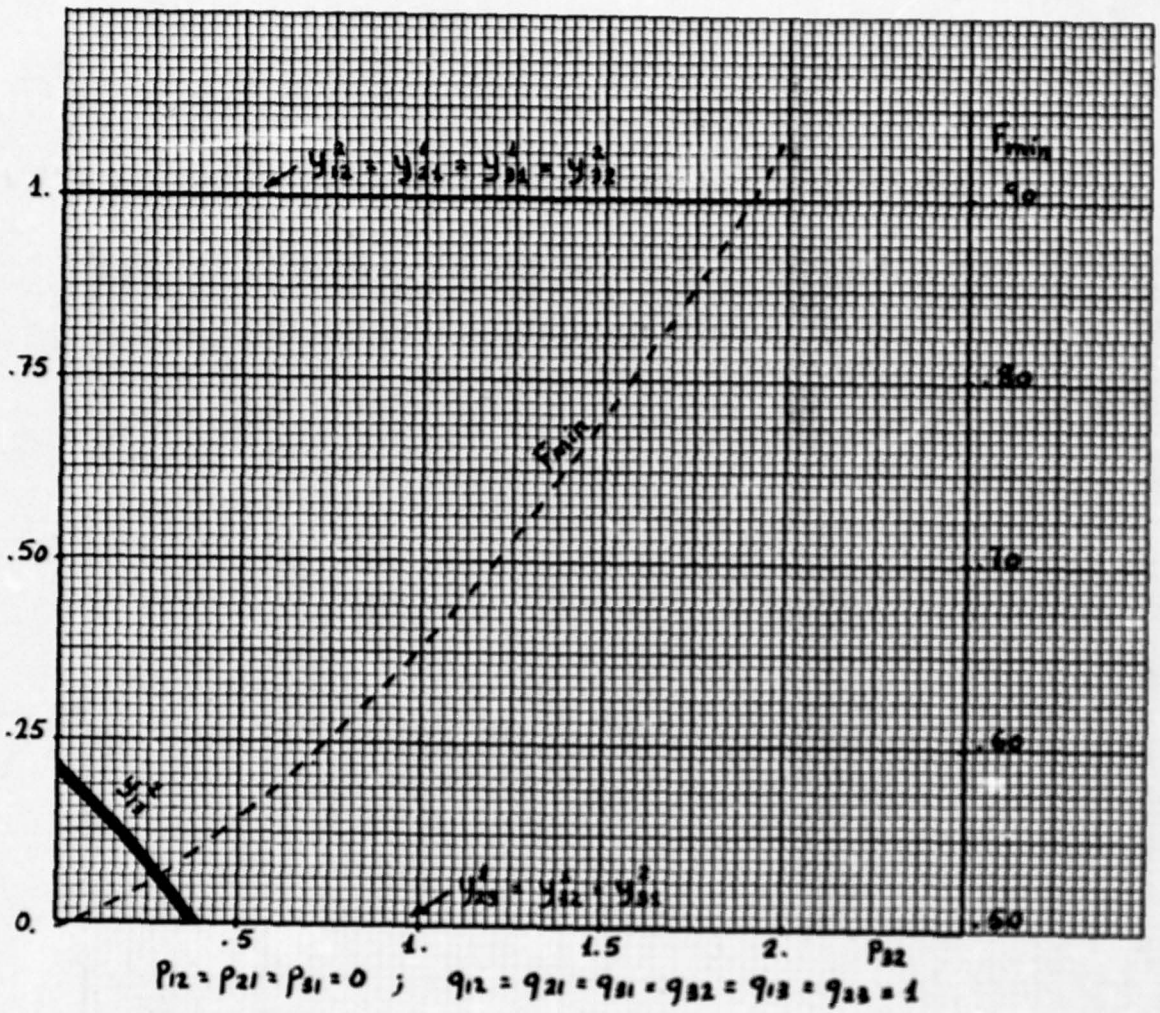


Fig. 5.26. Example 5.2. Routing variables and F_{min} in terms of p_{32}.

there and sent back to their origination point.

The problem is similar to the routing case we considered only the transmission channels as the servers in the different queues. For the load-sharing case we have two different types of servers, the transmission channels and the processors themselves. Therefore there will be a distinction between the transmission queues (in both ways: forwards for the programms that have to be processed in the distant computer and backwards for the results that come out from the computers)

The model that was established in section 4.1 can be applied here by considering queues and nodes associated with them so that they are connected by "links" which can either represent transmission channels or processors. This is the general formulation for the Statistical Load Sharing Problem" that Wunderlich suggests [26, chapter 4.3] .

The load sharing is formulated as a multicommodity flow problem in which the optimal flow through the network has to be found for fixed capacities and given inputs.

To see how the model of section 4.1 can be applied to a load sharing computer system consider the following example: (Fig 5.27)

There is a system of two computers called 1 and 2 which are connected by two transmission channels of capacities l_1 and l_2. The jobs entering computer 1 for example are queued up forming the queueing process $x_1(t)$. They can be served either by the computer 1 (and upon servicing leave the system) or can be transmitted by channel l_1 to the other distant center. There they wait at queue $x_5(t)$ until its turn comes and are processed by computer 2. After being processed they enter queue $x_6(t)$ and

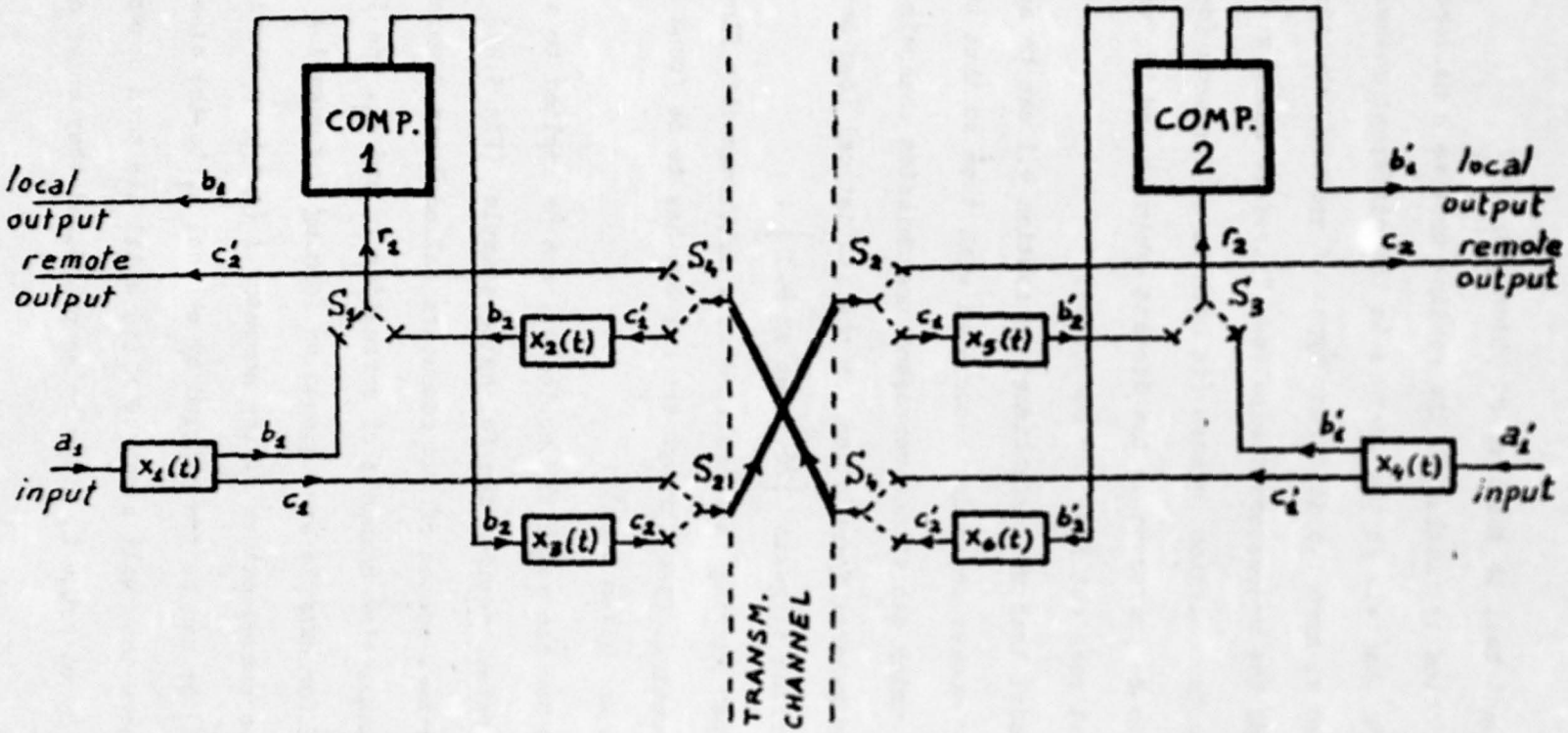


Fig 5.27. Load sharing example between two computers

wait their turn for transmission to their origination point via channel l_2. It can be similarly said about the messages entering computer 2. As we said earlier there are two types of of queues processing queues and communication queues.

In order to approximate this case by a diffusion process, let us consider the time divided in small intervals $(t, t + \Delta t)$ and the messages composed of small blocks of duration Δt. During this interval we call

a_1 = Prob. that a message of length Δt enters at computer 1.

a_1' = Prob. that a message of length Δt enters at computer 2.

b_1 = Prob. that a message of length Δt waiting at $x_1(t)$ is processed by computer 1.

c_1 = Prob. that a message of length Δt waiting at $x_1(t)$ is transmitted through channel l_1 to computer 2 and enters queue $x_5(t)$ to be processed.

b_1' = Prob. that a message of length Δt waiting at $x_4(t)$ is processed by computer 2.

c_1' = Prob. that a message of length Δt waiting at $x_4(t)$ is transmitted through channel l_2 to computer 1 and enters queue $x_2(t)$ to be processed

b_2 = Prob. that a message of length Δt waiting at $x_2(t)$ is processed by computer 1 and enters queue $x_3(t)$ to be retransmitted

c_2 = Prob. that a message of length Δt waiting at $x_3(t)$ is transmitted to computer 2 and leaves the system.

b_2' = Prob. that a message of length Δt waiting at $x_5(t)$ is processed by computer 2 and enters queue $x_6(t)$ to be retransmitted.

c_2' = Prob. that a message of length Δt waiting at $x_6(t)$ is transmitted to computerl and leaves the system

The computer service processed are represented by r_1 and r_2, the computer capacities and defined as the maximum processing probabilities of

a message of length Δt during $(t, t+\Delta t)$ at computers 1 and 2 respectively.

The channel capacities are l_1 and l_2 defined as the maximum probability that a message of length Δt is transmitted through the respective channel during $(t, t+\Delta t)$.

According to this the computer processing constraints are

$$b_1 + b_2 \leqslant r_1 \quad \text{and} \quad b_1' + b_2' \leqslant r_2 \tag{5-46}$$

and the channel capacities constraints

$$c_1 + c_2 \leqslant l_1 \quad \text{and} \quad c_1' + c_2' \leqslant l_2 \tag{5-47}$$

As it can be seen computer 1 processes its own jobs with probability b_1 and jobs from the other center with probability b_2. As soon as a job gets through the computer it comes out and waits in case it has to go to the other end. Similarly for computer 2.

The transmission channels are either used to transmit jobs to be processed at the other computer (probabilities c_1 and c_1') or jobs that have been processed to their origin (probabilities c_2 and c_2')

In order to use for this case the general expressions that were developed in section 4 this model will be made equivalent to that of Fig. 4.1 by adding two dummy nodes and considering general capacities for the links between nodes without specifying whether they are refered to the computer processing capacities. This is shown in Fig 5.28.

As it can be seen, computer 1 is unfolded in two nodes 1 and 2. Messages enter at node 1. The "channel" $1 \rightarrow 2$ represemts the computer processing so that $c_{12} = l_1$. Once a message is served is goes out if it was local or waits at node 2 to be transmitted to its origin. The capacity of the "channel" $2 \rightarrow 1$ $c_{21} = 1$ because the only queueing process for

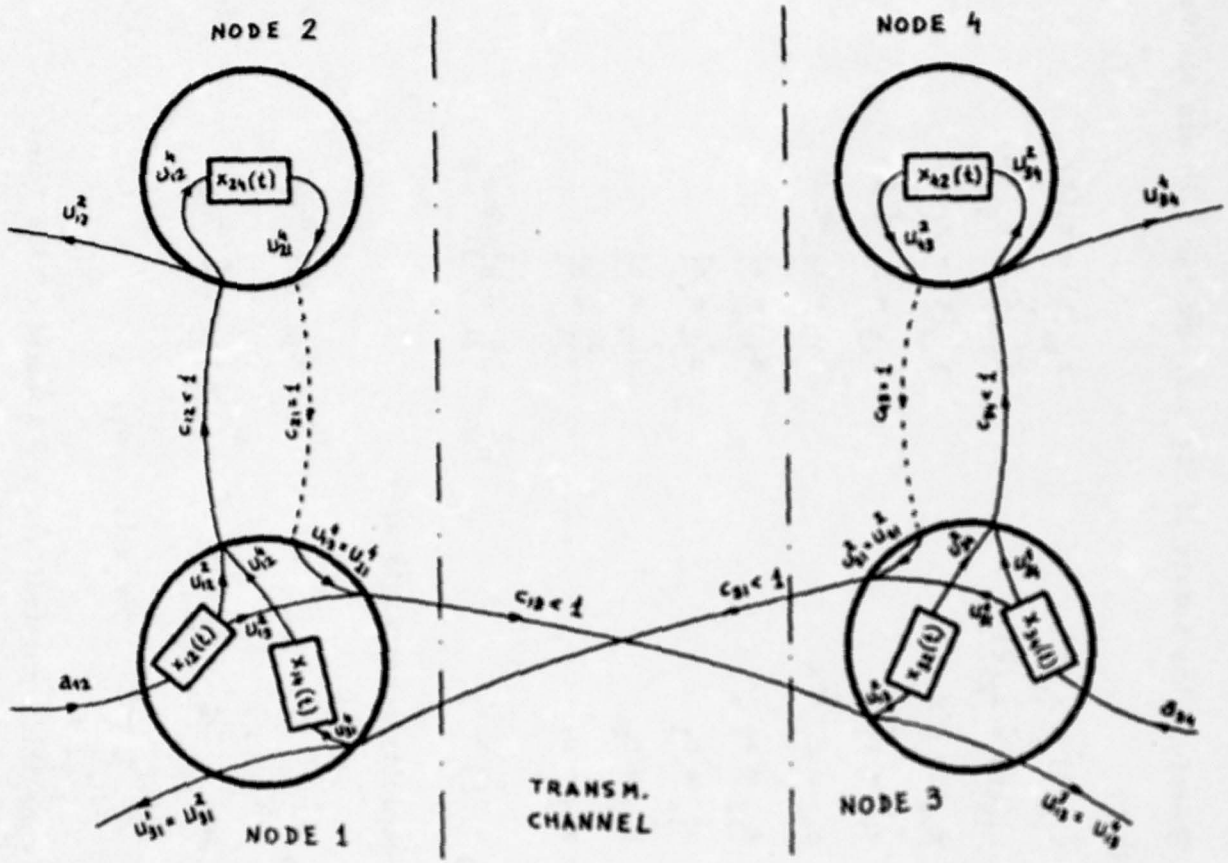


Fig. 5.28. Equivalent representation of the load sharing example of Fig. 5.27.

retransmission is due to the transmission channel whose capacity is $c_{13} = l_1$. Therefore the models of Fig 5.27 and Fig 5.28 are equivalent provided that:

$$x_{12}(t) = x_1(t) \qquad x_{34}(t) = x_5(t)$$
$$x_{14}(t) = x_2(t) \qquad x_{32}(t) = x_5(t)$$
$$x_{24}(t) = x_3(t) \qquad x_{42}(t) = x_6(t)$$
$$c_{12} = r_1 < 1 \qquad c_{13} = l_1 < 1$$
$$c_{34} = r_2 < 1 \qquad c_{31} = l_2 < 1$$
$$a_{12} = a_1 \qquad a_{34} = a_1'$$
$$u_{12}^2 = b_1 \qquad u_{34}^4 = b_1'$$
$$u_{12}^4 = b_2 \qquad u_{34}^2 = b_2'$$
$$u_{13}^2 = c_1 \qquad u_{31}^4 = c_1'$$
$$u_{21}^2 = u_{13}^4 = u_{13}^3 = c_2 \qquad u_{43}^2 = u_{31}^2 = u_{31}^1 = c_2'$$

The capacity constraints are:

$$u_{12}^2 + u_{12}^4 \leqslant c_{12} \quad ; \quad u_{34}^4 + u_{34}^2 \leqslant c_{34} \tag{5-48}$$

$$u_{13}^2 + u_{13}^4 \leqslant c_{13} \quad ; \quad u_{31}^4 + u_{31}^2 \leqslant c_{31} \tag{5-49}$$

and $$u_{ij}^k \geqslant 0 \quad \text{for all} \quad u_{ij}^k \tag{5-50}$$

The capacities and traffics are assumed of the form:

$$\left.\begin{aligned} c_{ij} &= C_{ij} + q_{ij}\sqrt{\Delta t} \\ a_{ij} &= A_{ij} + p_{ij}\sqrt{\Delta t} \\ u_{ij}^k &= U_{ij}^k + y_{ij}^k\sqrt{\Delta t} \end{aligned}\right\} \quad \forall(ij) \tag{5-51}$$

We can obtain the expression of the vector drift components $\underline{\beta}$

From (4-32) (assume the constants $K_{ij} = 1$ as in the former sections)

$$\begin{aligned}
\beta_{12} &= p_{12} - y_{12}^2 - y_{13}^2 \\
\beta_{14} &= y_{31}^4 - y_{12}^4 \\
\beta_{24} &= y_{12}^4 - y_{13}^4 \\
\beta_{34} &= p_{34} - y_{34}^4 - y_{31}^4 \\
\beta_{32} &= y_{13}^2 - y_{34}^2 \\
\beta_{42} &= y_{34}^2 - y_{31}^2
\end{aligned} \qquad (5\text{-}52)$$

Expression of the covariance matrix elements

From (4-35)

$$\alpha_{(12)(12)} \equiv \alpha_{11} = A_{12}(1-A_{12}) + U_{12}^2(1-U_{12}^2) + U_{13}^2(1-U_{13}^2)$$

$$\alpha_{(14)(14)} \equiv \alpha_{22} = U_{31}^4(1-U_{31}^4) + U_{12}^4(1-U_{12}^4)$$

$$\alpha_{(24)(24)} \equiv \alpha_{33} = U_{12}^4(1-U_{12}^4) + U_{13}^4(1-U_{13}^4)$$

$$\alpha_{(34)(34)} \equiv \alpha_{44} = A_{34}(1-A_{34}) + U_{34}^4(1-U_{34}^4) + U_{31}^4(1-U_{31}^4)$$

$$\alpha_{(32)(32)} \equiv \alpha_{55} = U_{13}^2(1-U_{13}^2) + U_{34}^2(1-U_{34}^2)$$

$$\alpha_{(42)(42)} \equiv \alpha_{66} = U_{34}^2(1-U_{34}^2) + U_{31}^2(1-U_{31}^2)$$

From (4-36): $\alpha_{(12)(14)} \equiv \alpha_{12} = -U_{12}^2\, U_{12}^4$

From (4-38) and (4-36) (Remember $u_{21}^4 = u_{13}^4$)

$$\alpha_{(12)(24)} \equiv \alpha_{13} = U_{12}^2 U_{12}^4 - U_{13}^2 U_{13}^4$$

From (4-38)

$$\alpha_{(12)(34)} \equiv \alpha_{14} = 0$$

From (4-37)

$$\alpha_{(12)(32)} \equiv \alpha_{15} = -U_{13}^2(1 - U_{13}^2)$$

$$\alpha_{(12)(42)} \equiv \alpha_{16} = 0$$

$$\alpha_{(14)(24)} \equiv \alpha_{23} = U_{12}^4(1 - U_{12}^4)$$

$$\alpha_{(14)(34)} \equiv \alpha_{24} = -U_{31}^4(1 - U_{31}^4)$$

From (4-38):

$$\alpha_{(14)(32)} \equiv \alpha_{25} = 0$$

$$\alpha_{(14)(42)} \equiv \alpha_{26} = U_{31}^2 U_{31}^4$$

From (4-37)

$$\alpha_{(24)(34)} \equiv \alpha_{34} = 0$$

From (4-38)

$$\alpha_{(24)(32)} \equiv \alpha_{35} = U_{13}^4 U_{13}^2$$

$$\alpha_{(24)(42)} \equiv \alpha_{36} = 0$$

From (4-36)

$$\alpha_{(34)(32)} \equiv \alpha_{45} = -U_{34}^4 U_{34}^2$$

From (4-38) and (4-36) (Remember $U^2_{43} = U^2_{31}$) :

$$\alpha_{(34)(42)} \equiv \alpha_{46} = U^4_{34}\, U^2_{34} - U^4_{31}\, U^2_{31}$$

From (4-37)

$$\alpha_{(32)(42)} \equiv \alpha_{56} = -U^2_{34}(1 - U^2_{34})$$

Then

$$\underline{\Delta} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} & 0 & \alpha_{15} & 0 \\ \alpha_{12} & \alpha_{22} & \alpha_{23} & \alpha_{24} & 0 & \alpha_{26} \\ \alpha_{13} & \alpha_{23} & \alpha_{33} & 0 & \alpha_{35} & 0 \\ 0 & \alpha_{24} & 0 & \alpha_{44} & \alpha_{45} & \alpha_{46} \\ \alpha_{15} & 0 & \alpha_{35} & \alpha_{45} & \alpha_{55} & \alpha_{56} \\ 0 & \alpha_{26} & 0 & \alpha_{46} & \alpha_{56} & \alpha_{66} \end{bmatrix} \qquad (5\text{-}53)$$

The flow equations are (from (4-31)

$$\left.\begin{aligned} A_{12} - U^2_{12} - U^2_{13} &= 0 \\ U^4_{31} - U^4_{12} &= 0 \\ U^4_{12} - U^4_{13} &= 0 \\ A_{34} - U^4_{34} - U^4_{31} &= 0 \\ U^2_{13} - U^2_{34} &= 0 \\ U^2_{34} - U^2_{31} &= 0 \end{aligned}\right\} \qquad (5\text{-}34)$$

Subject to:

$$\left.\begin{array}{lll} U^2_{12} + U^4_{12} \leqslant C_{12} & ; & U^4_{34} + U^2_{34} \leqslant C_{34} \\ U^2_{13} + U^4_{13} \leqslant C_{13} & ; & U^4_{31} + U^2_{31} \leqslant C_{31} \end{array}\right\} \quad (5\text{-}55)$$

for all $U^k_{ij} \geqslant 0$

As we did in the preceding examples, we assume values for C_{ij}, A_{ij} and U^k_{ij} that satisfy (5-54) and (5-55). For given inputs p_{ij} and capacities q_{ij} we want to find the set of y^k_{ij} that minimize the overall mean constrained by:

$$\left.\begin{array}{lll} y^2_{12} + y^2_{12} \leqslant q_{12} & ; & y^4_{34} + y^2_{34} \leqslant q_{34} \\ y^2_{13} + y^4_{13} \leqslant q_{13} & ; & y^4_{31} + y^2_{31} \leqslant q_{31} \end{array}\right\} \quad (5\text{-}56)$$

for all $y^k_{ij} \geqslant 0$

summing up both sides of (5-52):

$$\sum_{\substack{\text{all} \\ \text{pairs } (ij)}} \beta_{ij} = p_{12} + p_{34} - (y^2_{12} + y^2_{34} + y^4_{13} + y^2_{31}) \quad (5\text{-}57)$$

From (3-7) $\underline{\beta} = \frac{1}{2}\ \underline{\Lambda}\ \underline{\gamma}$:

$$\left(\tfrac{1}{2} \sum_{i=1}^{6} \alpha_{i1}\right) \gamma_{12} + \left(\tfrac{1}{2} \sum_{i=1}^{6} \alpha_{i2}\right) \gamma_{14} + \left(\tfrac{1}{2} \sum_{i=1}^{6} \alpha_{i3}\right) \gamma_{24} +$$

$$+ \left(\tfrac{1}{2} \sum_{i=1}^{6} \alpha_{i4}\right) \gamma_{34} + \left(\tfrac{1}{2} \sum_{i=1}^{6} \alpha_{i5}\right) \gamma_{32} + \left(\tfrac{1}{2} \sum_{i=1}^{6} \alpha_{i6}\right) \gamma_{42} =$$

$$= p_{12} + p_{34} - (y^2_{12} + y^2_{34} + y^4_{13} + y^2_{31}) \quad (5\text{-}58)$$

From the expressions (5-53) it can be seen that not all the coefficients of the γ_{ij} in (5-58) have to be positive. Therefore we cannot assume that the optimum values for y_{12}^2, y_{34}^2, y_{13}^4, y_{31}^2 are the maximum ones that is q_{12}, q_{34}, q_{13}, q_{31} respectively.

The minimization of (4-39) has to be carried over all the components of the vector $\underline{y}$:

$$\underline{y} = (y_{12}^2, y_{34}^4, y_{13}^2, y_{31}^4, y_{12}^4, y_{34}^2, y_{13}^4, y_{31}^2)^T$$

Let us assume an example. Suppose

$A_{12} = 0.8$ $\quad C_{12} = 0.75$ $\quad C_{13} = 0.50$

$A_{34} = 0.50$ $\quad C_{34} = 0.75$ $\quad C_{31} = 0.50$

that is, the computer capacities are greater than the transmission channel capacities and computer 1 is more loaded than computer 2.

The set of values

$$U_{12}^2 = 0.50 \quad ; \quad U_{13}^2 = U_{34}^2 = U_{31}^2 = 0.30$$

$$U_{34}^4 = 0.40 \quad ; \quad U_{31}^4 = U_{12}^4 = U_{13}^4 = 0.20$$

satisfy (5-54) constrained by (5-55). Thus the matrix $\underline{\Lambda}$ is determined

$$\underline{\Lambda} = \begin{bmatrix} 0.62 & -0.05 & 0.02 & 0 & -0.21 & 0 \\ -0.05 & 0.18 & -0.09 & -0.09 & 0 & 0.03 \\ 0.02 & -0.09 & 0.18 & 0 & 0.03 & 0 \\ 0 & -0.09 & 0 & 0.58 & -0.12 & 0.09 \\ -0.21 & 0 & 0.03 & -0.12 & 0.42 & -0.21 \\ 0 & 0.03 & 0 & 0.09 & -0.21 & 0.42 \end{bmatrix} \qquad (5\text{-}59)$$

Several cases are shown in the next figures:

Consider Fig. 5.29: The inputs $p_{12} = p_{34} = 0$ and the capacities $q_{12} = q_{34} = q_{31} = 1$ whereas q_{13} varies.

For $q_{13} = 1$ we obtain the optimum y_{ij}^k

$$y_{12}^2 = y_{34}^4 = y_{31}^2 = 1 \quad ; \quad y_{13}^2 = y_{31}^4 = y_{12}^4 = y_{34}^2 = 0$$

and $y_{13}^4 = 0.8141$. As q_{13} decreases nothing happens until $q_{13} = 0.8141$. From this point on $y_{13}^4 = q_{13}$ and because of this decreasing in the capacity the overall mean $F_{min}(\underline{y})$ starts to increase considerably.

Fig 5.30: q_{31} varies and the other parameters are held fixed. Now we obtain

$$y_{12}^2 = y_{34}^4 = 1 \quad ; \quad y_{31}^2 = q_{31}$$

$y_{13}^2 = y_{31}^4 = y_{12}^4 = y_{34}^2 = 0$ and y_{13}^4 decreases with q_{31} decreasing. This can be explained from the fact that if the capacity of the link between computer 2 and 1 decreases, less messages have to be sent from computer 2 to computer 1 and therefore less messages have to come back. The consequence of q_{31} decreasing is an increase of the overall mean $F_{min}(\underline{y})$

Fig. 5.31: All $q_{ij} = 1$, $p_{12} = 0$ and p_{34} increases that is computer 2 starts being more loaded. It is obtained $y_{12}^2 = y_{34}^4 = y_{31}^2 = 1$; $y_{13}^2 = y_{31}^4 = y_{12}^4 = y_{34}^2 = 0$ and y_{13}^4 increases from 0.8141 up to 1: messages coming back to computer 2 have to do it faster.

Fig. 5.32: All $q_{ij} = 0$, $p_{34} = 0$ and p_{12} increases. The effect is the

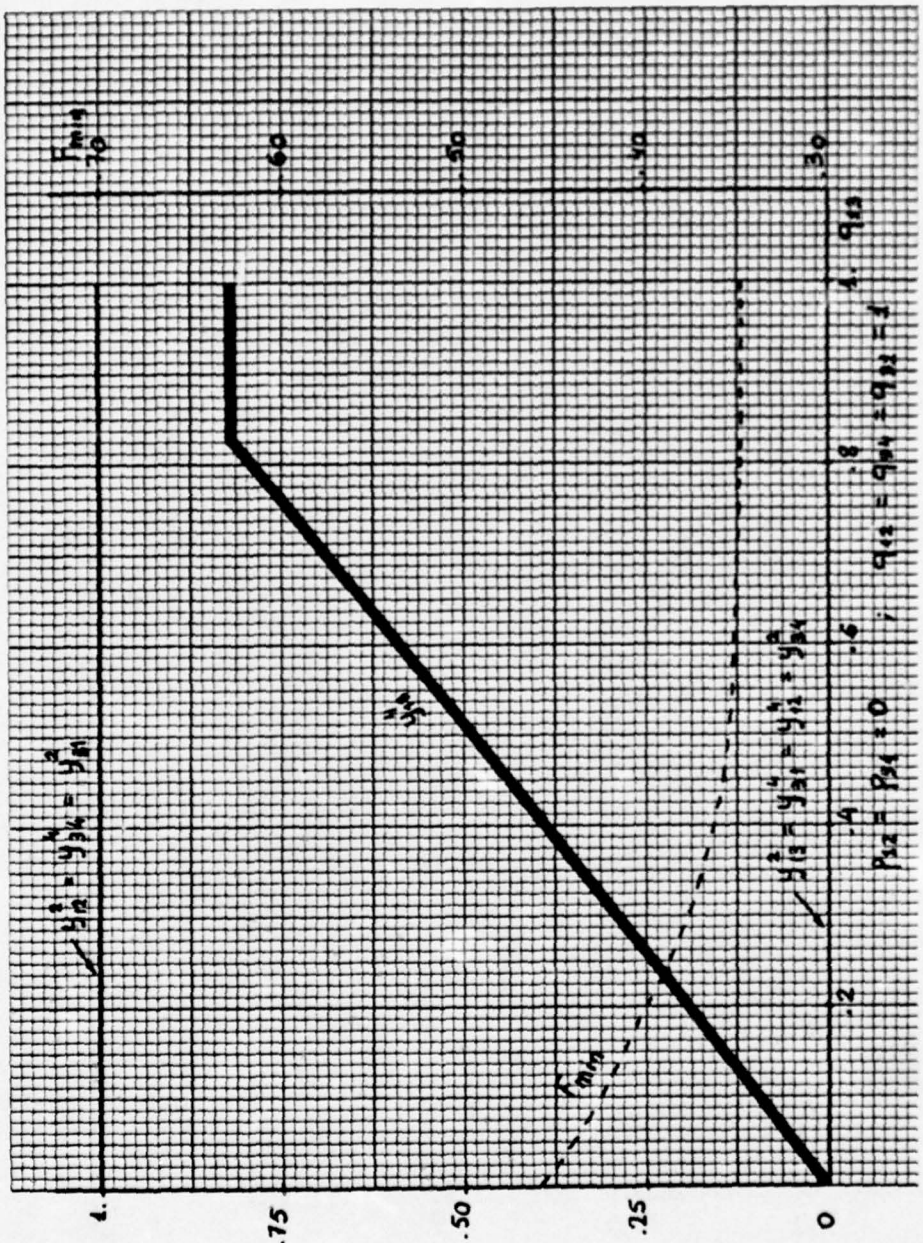


Fig. 5.29. Example 5.3. Control variables and F_{min} in terms of q_{13}.

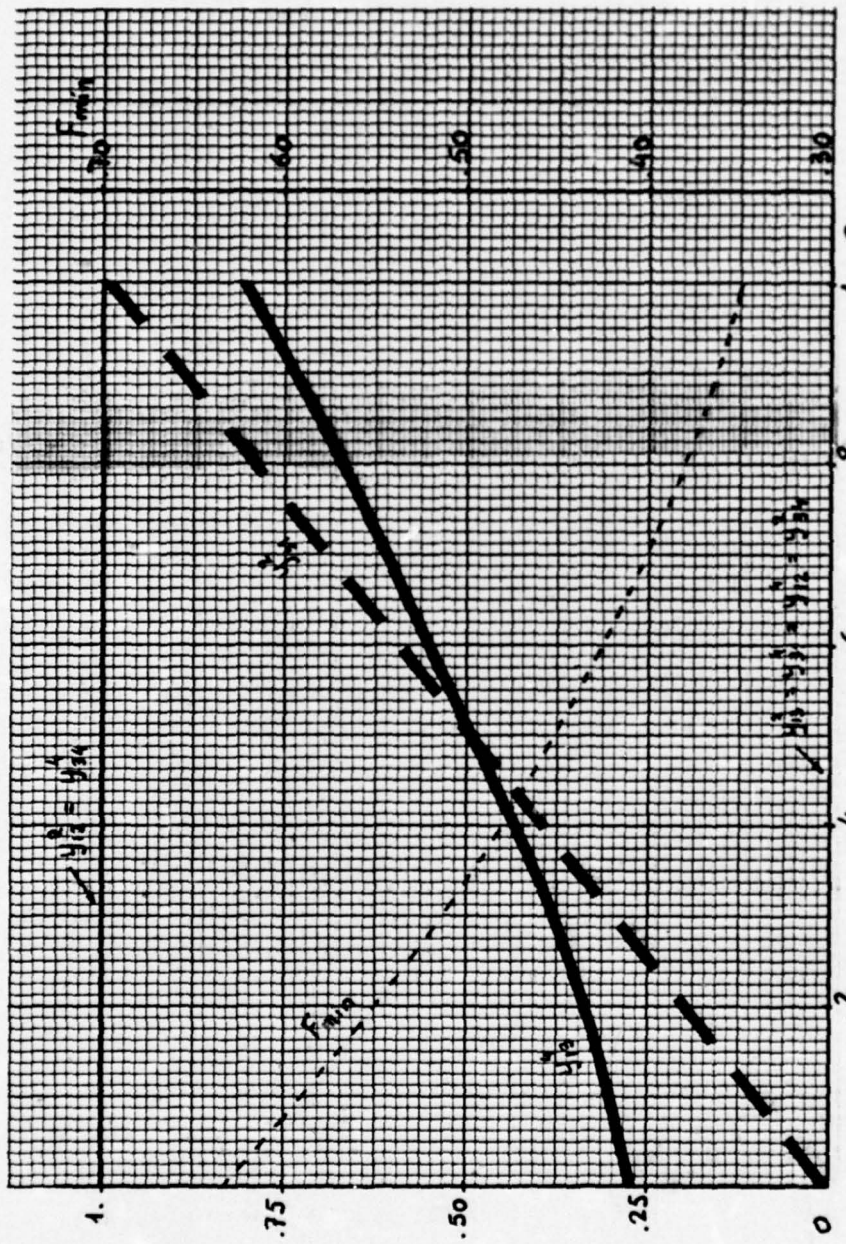


Fig. 5.30. Example 5.3. Control variables and F_{min} in terms of q_{31}.

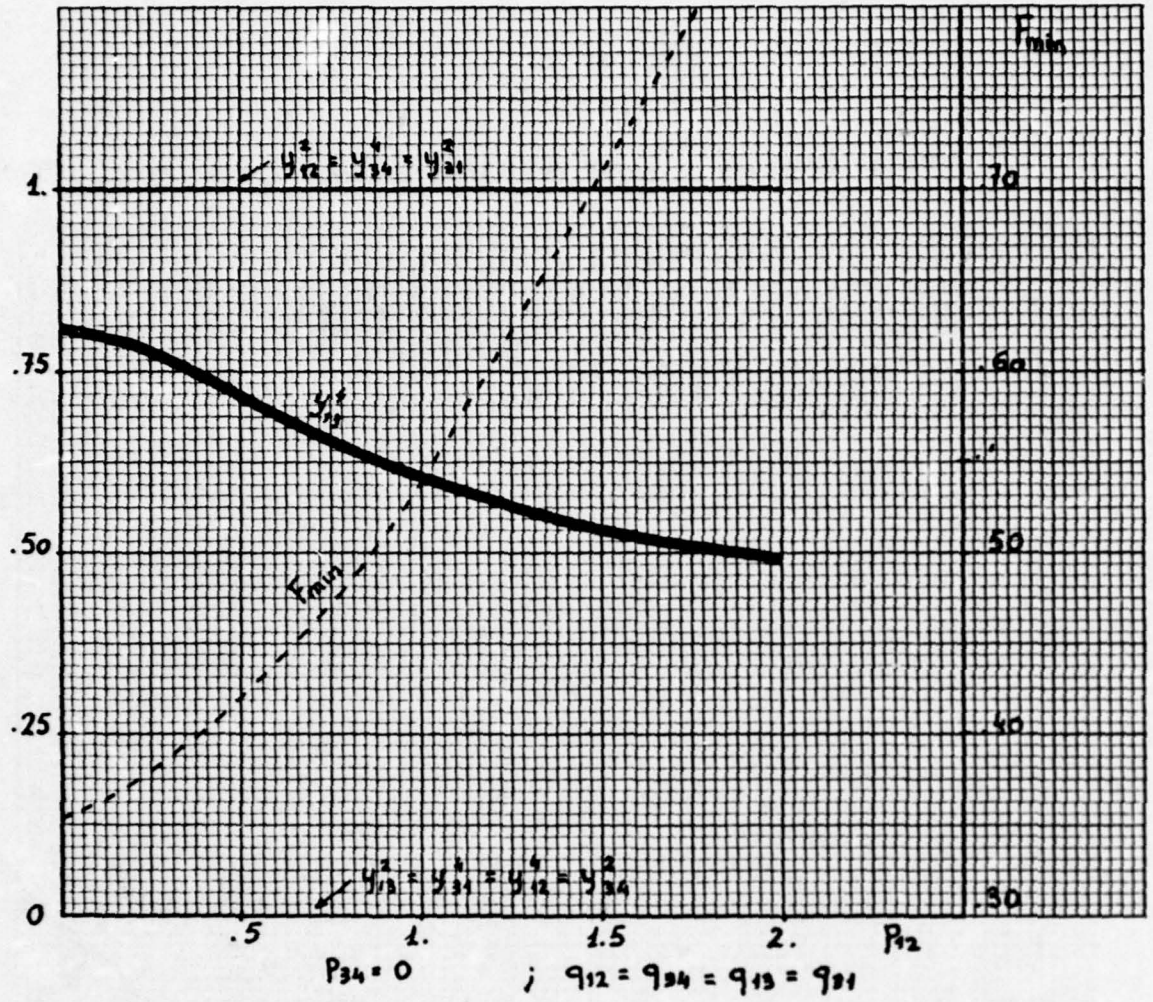


Fig. 5.31 Example 5.3. Control variables and F_{min} in terms of p_{12}.

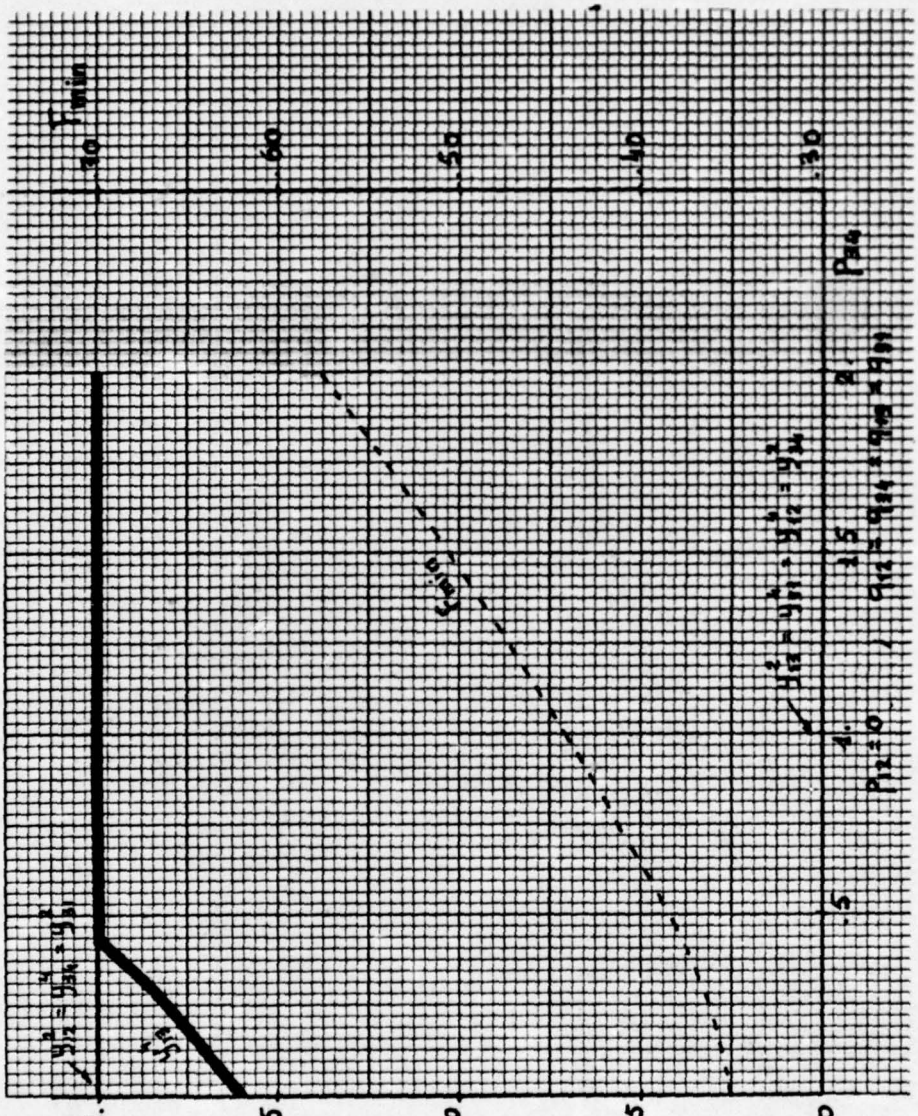


Fig. 5.32. Example 5.3. Control variables and F_{min} in terms of p_{34}.

opposite to that of the previous figure that is y_{13}^4 decreases. The overall mean increases more than in the former case.

Let us assume now that the capacities of the transmission channels are greater than those of the computers. For example:

$$A_{12} = 0.60 \quad ; \quad C_{12} = 0.50 \quad ; \quad C_{13} = 0.75$$

$$A_{34} = 0.50 \quad ; \quad C_{34}\ 0.50 \quad ; \quad C_{31} = 0.75$$

The set of values:

$$U_{12}^2 = 0.40 \quad ; \quad U_{13}^2 = U_{34}^2 = U_{31}^2 = 0.20$$

$$U_{34}^4 = 0.20 \quad ; \quad U_{31}^4 = U_{12}^4 = U_{13}^4 = 0.10$$

satisfy (5-54) and (5-55). The matrix $\underline{\Lambda}$ is

$$\underline{\Lambda} = \begin{bmatrix} 0.64 & -0.04 & 0.02 & 0 & -0.16 & 0 \\ -0.04 & 0.18 & -0.09 & -0.09 & 0. & 0.02 \\ 0.02 & -0.09 & 0.18 & 0 & 0.02 & 0 \\ 0 & -0.09 & 0 & 0.46 & -0.04 & 0.02 \\ -0.16 & 0 & 0.02 & -0.04 & 0.32 & -0.16 \\ 0 & 0.02 & 0 & 0.02 & -0.16 & 0.32 \end{bmatrix} \quad (5\text{-}60)$$

The same cases are plotted in the next figures.

Fig. 5.33: For $p_{12} = p_{34} = 0$, $q_{12} = q_{34} = q_{31} = 1$ and q_{13} varying it is obtained:

$$y_{12}^2 = y_{34}^4 = y_{31}^2 = 1 \quad ; \quad y_{13}^2 = y_{31}^4 = y_{12}^4 = y_{34}^2 = 0$$

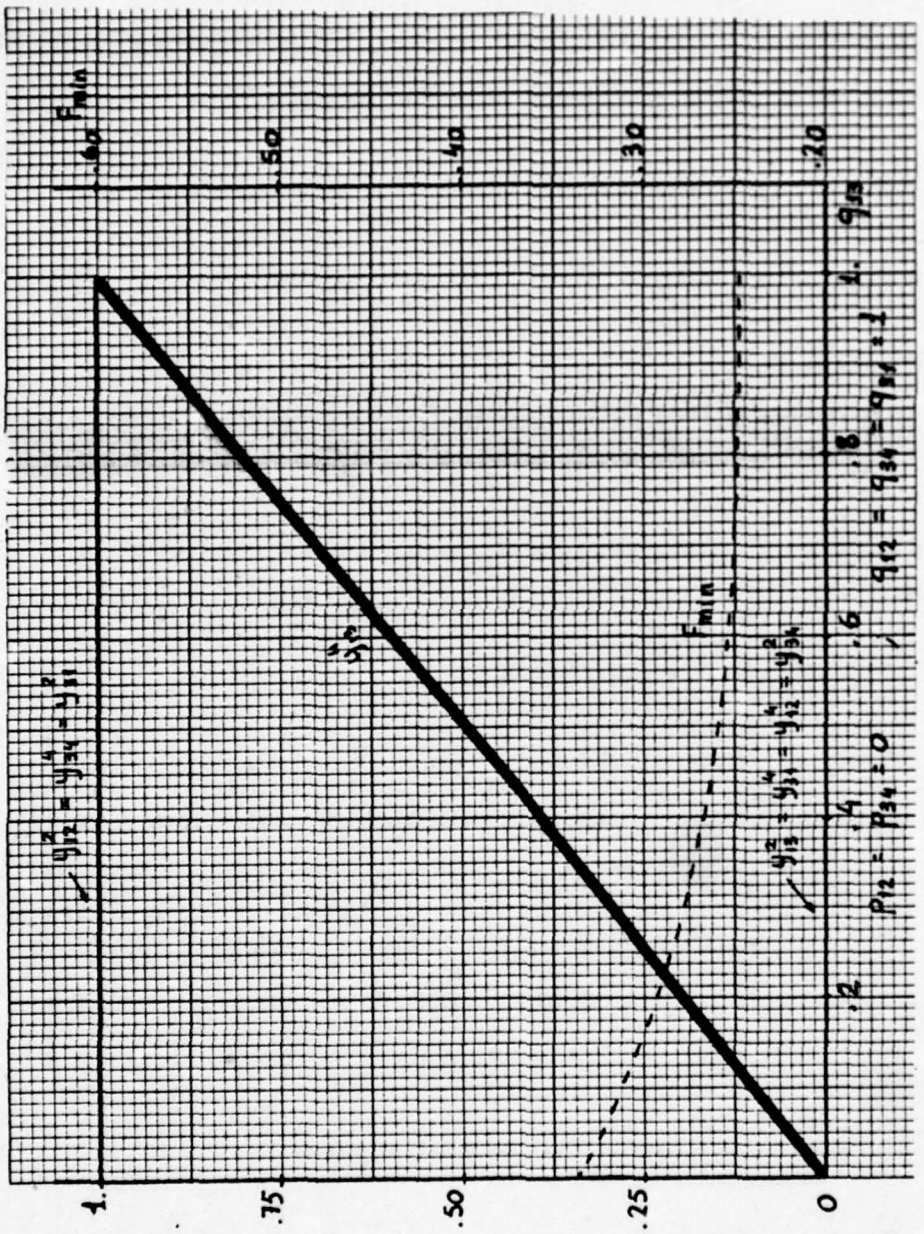


Fig. 5.33. Example 5.3. Control variables and F_{min} in terms of q_{13}

and $y_{13}^4 = q_{13}$. The overall mean increases for q_{13} decreasing.

Fig. 5.34: For $p_{12} = p_{34} = 0$; $q_{12} = q_{34} = q_{13} = 1$ and q_{31} varying, we obtain

$$y_{12}^2 = y_{34}^4 = 1 \quad ; \quad y_{13}^2 = y_{31}^4 = y_{12}^4 = y_{34}^2 = 0$$

$y_{31}^2 = q_{31}$ and $y_{13}^4 = 1$ for $q_{31} = 1$. As q_{31} decreases y_{13}^4 remains the same up to some value of $q_{31} \approx 0.8$ at which y_{13}^4 decreases.

Compare Figs. 5.29 and 5.33 and Figs. 5.30 and 5.34. Similar behavior is observed. In these cases the overall mean is less than in the former ones because both computers are less loaded.

Fig. 5.35: $q_{12} = q_{34} = q_{13} = q_{31} = 1$; $p_{12} = 0$ and p_{34} increases. It is obtained:

$$y_{12}^2 = y_{34}^4 = y_{13}^4 = y_{31}^2 = 1 \quad \text{and} \quad y_{13}^2 = y_{31}^4 = y_{12}^4 = y_{34}^2$$

Fig. 5.36: $q_{12} = q_{34} = q_{13} = q_{31} = 1$; $p_{34} = 0$ and p_{12} increases. It is obtained:

$$y_{12}^2 = y_{13}^4 = y_{31}^2 = 1 \quad ; \quad y_{12}^4 = y_{13}^2 = y_{31}^4 = 0$$

and at some value of $p_{12} \approx 0.2$ y_{34}^4 starts decreasing from 1 and y_{34}^2 starts increasing from 0 such that $y_{34}^4 + y_{34}^2 = q_{34} = 1$. That is computer 2 processes more jobs from computer 1 until $y_{34}^4 = y_{34}^2 = q_{34}/2 = 0.5$. This load sharing in this case can be explained from the fact that now the transmission channels are faster than the computers. In Fig. 5.32 there was not such effect because the computer were faster. One can also

observe that due to that fact the increase on the overall mean is considerably less than in Fig. 5.32

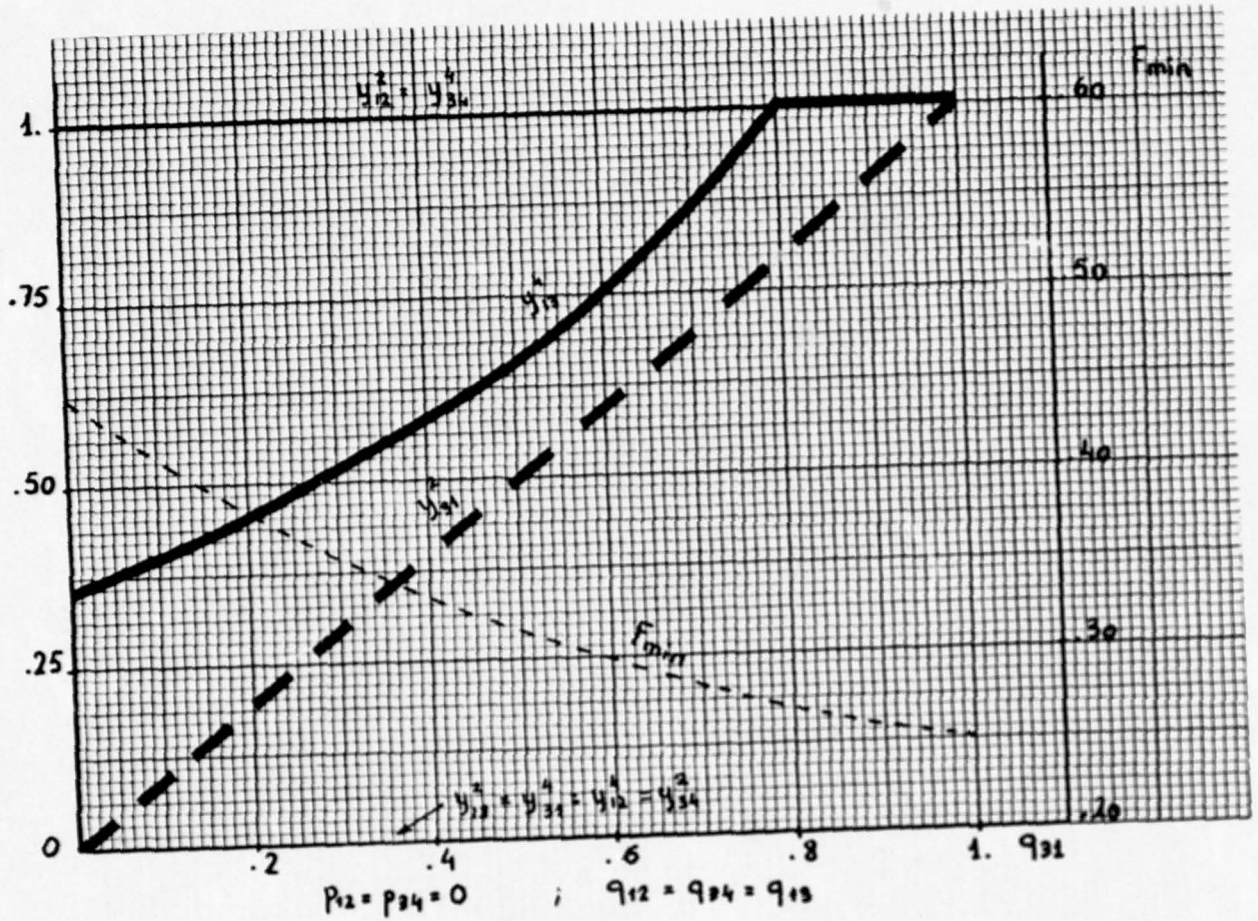


Fig. 5.34. Example 5.3. Control variables and F_{min} in terms of q_{31}.

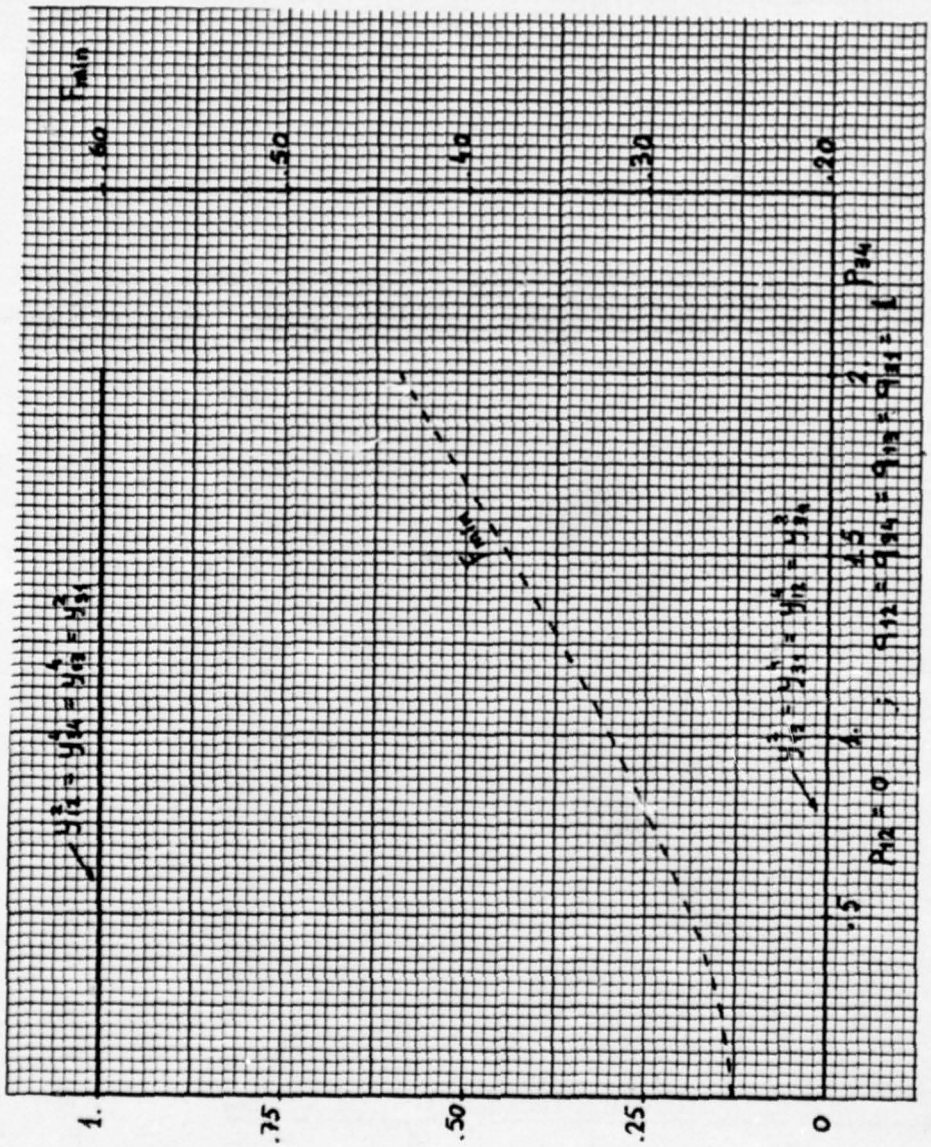


Fig. 5.35. Example 5.3. Control variables and F_{min} in terms of P_{34}.

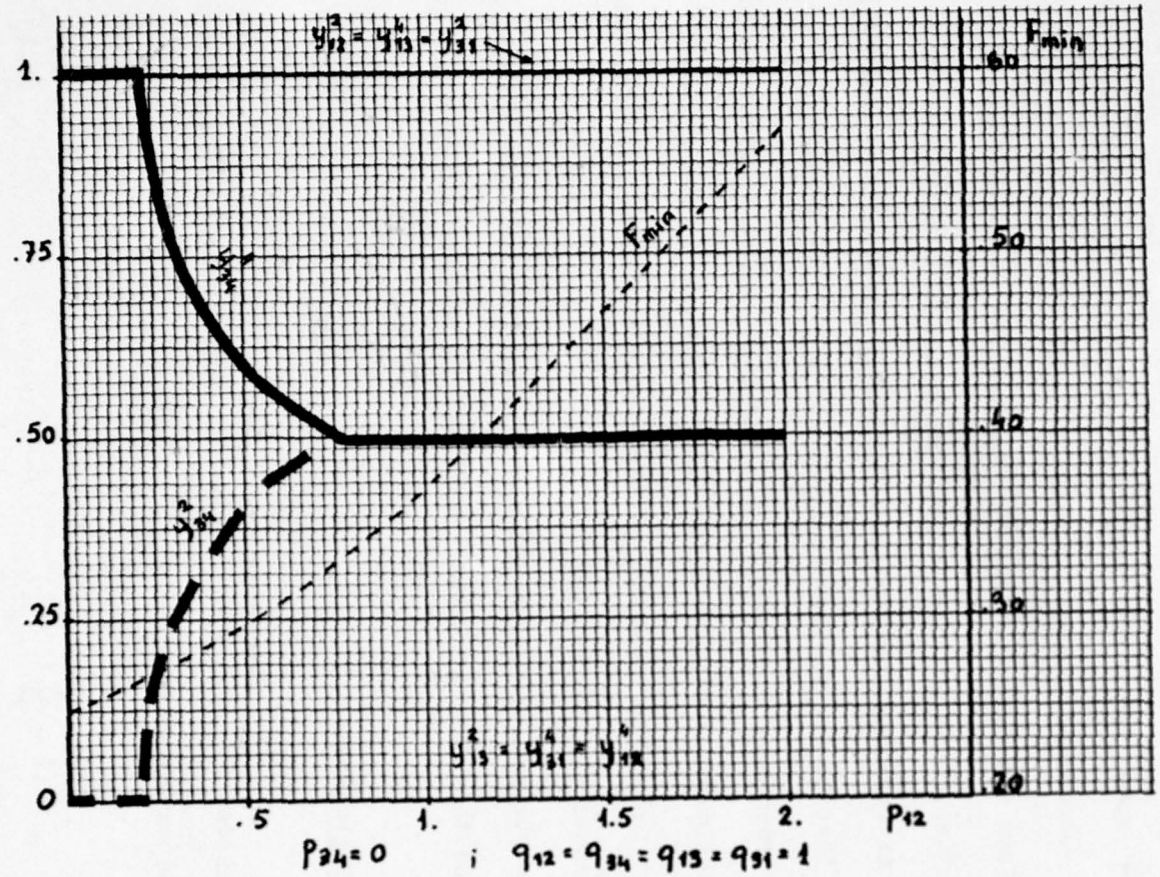


Fig. 5.36. Example 5.3. Control variables and F_{min} in terms of P_{12}.

6.- COMPARISON OF THE DIFFUSION MODEL WITH THE M/M/1

The model M/M/1 has been extensively used to study networks of queues because of its mathematical tractability [11] . Its attractivness stems from the fact that the interarrival and service times are exponential. This and the assumption of infinite buffer capacity lead to simple expressions for the distribution in steady-state. For M/M/1/N queues that is for finite buffer capacity N, the mathematical analysis becomes much more complicated [7, 12] . Therefore in order to compare the constinuous-state diffusion model developed in the preceeding sections for computer-communication networks with a discrete-state model, M/M/1 queues will be assumed.

The comparison between both models will be made first for a single queue and then for the network analyzed in Section 5.1 with two queues.

6.1.- Single queue

Consider a single M/M/1 queue with input rate λ and service rate μ The steady-state distribution is : [7]

$$P_n = (1 - \frac{\lambda}{\mu}) \ (\frac{\lambda}{\mu})^n \quad ; \ n = 0,1,2,... \quad \lambda < \mu \tag{6-1}$$

and the mean value

$$\bar{n}_e = \frac{\lambda}{\mu - \lambda} \tag{6-2}$$

Consider now the diffusion queue x(t) with capacity N messages,

and where β and α are respectively the mean and variance per unit time of the process $x(t)$.

As we saw in Section 4.1 the number of messages in the queue is $n(t) = N\,x(t)$ so that its mean and variance per unit time respectively

$$\frac{E[n(t)]}{t} = N\,\frac{E[x(t)]}{t} = N\beta \qquad (6\text{-}3a)$$

$$\frac{\mathrm{Var}[n(t)]}{t} = N^2\,\frac{\mathrm{Var}[x(t)]}{t} = N^2\alpha \qquad (6\text{-}3b)$$

For the M/M/1 queue the mean and variance per unit time are

$$\frac{E[n(t)]}{t} = \lambda - \mu \qquad (6\text{-}4a)$$

$$\frac{\mathrm{Var}[n(t)]}{t} = \lambda + \mu \qquad (6\text{-}4b)$$

Therefore the diffusion queue has the same mean and variance per unit time as an M/M/1 queue whose arrival and service rate are respectively

$$\lambda = \frac{N}{2}(N\alpha + \beta) \qquad (6\text{-}5a)$$

$$\mu = \frac{N}{2}(N\alpha - \beta) \qquad (6\text{-}5b)$$

Provided $\beta < 0$ and $|\beta| < N\alpha$

Assuming the input and service rates given by (6-5) let us calculate the expression of the mean value.

Substituting (6-5) into (6-2) we obtain:

$$\bar{n}_e = \frac{N\alpha + \beta}{-2\beta} = -N\,\frac{\alpha}{2\beta} - \frac{1}{2} \qquad (6\text{-}6)$$

$\beta < 0$ and $|\beta| < N\alpha$

Let us call $\bar{n}_d$ the mean value corresponding to the diffusion queue. Clearly $\bar{n}_d = N\,\bar{x}$ where

$$\bar{x} = \left(\frac{e^{\gamma}}{e^{\gamma} - 1} - \frac{1}{\gamma} \right) \qquad (6\text{-}7)$$

and is represented in Fig. 4.4 and where $\gamma = 2\beta/\alpha$ (Remember Sec.3) Therefore

$$\bar{n}_d = N \left(- \frac{\alpha}{2\beta} + \frac{e^{(2\beta/\alpha)}}{e^{(2\beta/\alpha)} - 1} \right) \qquad (6\text{-}8)$$

For $(2\beta/\alpha) < -3$ the expression (6-8) can be approximated by

$$\bar{n}_d \approx -N \frac{\alpha}{2\beta} \qquad (6\text{-}9)$$

Looking at (6-6) and (6-9) we can see that for $(2\beta/\alpha) < -3$ both expressions are very close except for the term -(1/2) in (6-6), but this difference becomes paltry as N increases.

For $-3 < (2\beta/\alpha) < 0$ the exponential term in (6-8) becomes significant and cancels the pole of $\alpha/2\beta$ at $\beta = 0$ whereas the expression (6-6) becomes unbounded for $\beta = 0$. This is due to the fact that the M/M/1 queue has infinite capacity. For $\beta > 0$ it does not make sense to compare $\bar{n}_e$ and $\bar{n}_d$ because for $\beta > 0$ there is no steady-state for the M/M/1 queue since this would imply $\lambda > \mu$ (See expressions (6-5)).

In Fig 6.1 $\bar{n}_e$ and $\bar{n}_d$ are compared in terms of β for a fixed value of $\alpha = 0.5$ and for different values of N.

The solid curves, corresponding to $\bar{n}_d$ do not become unbounded

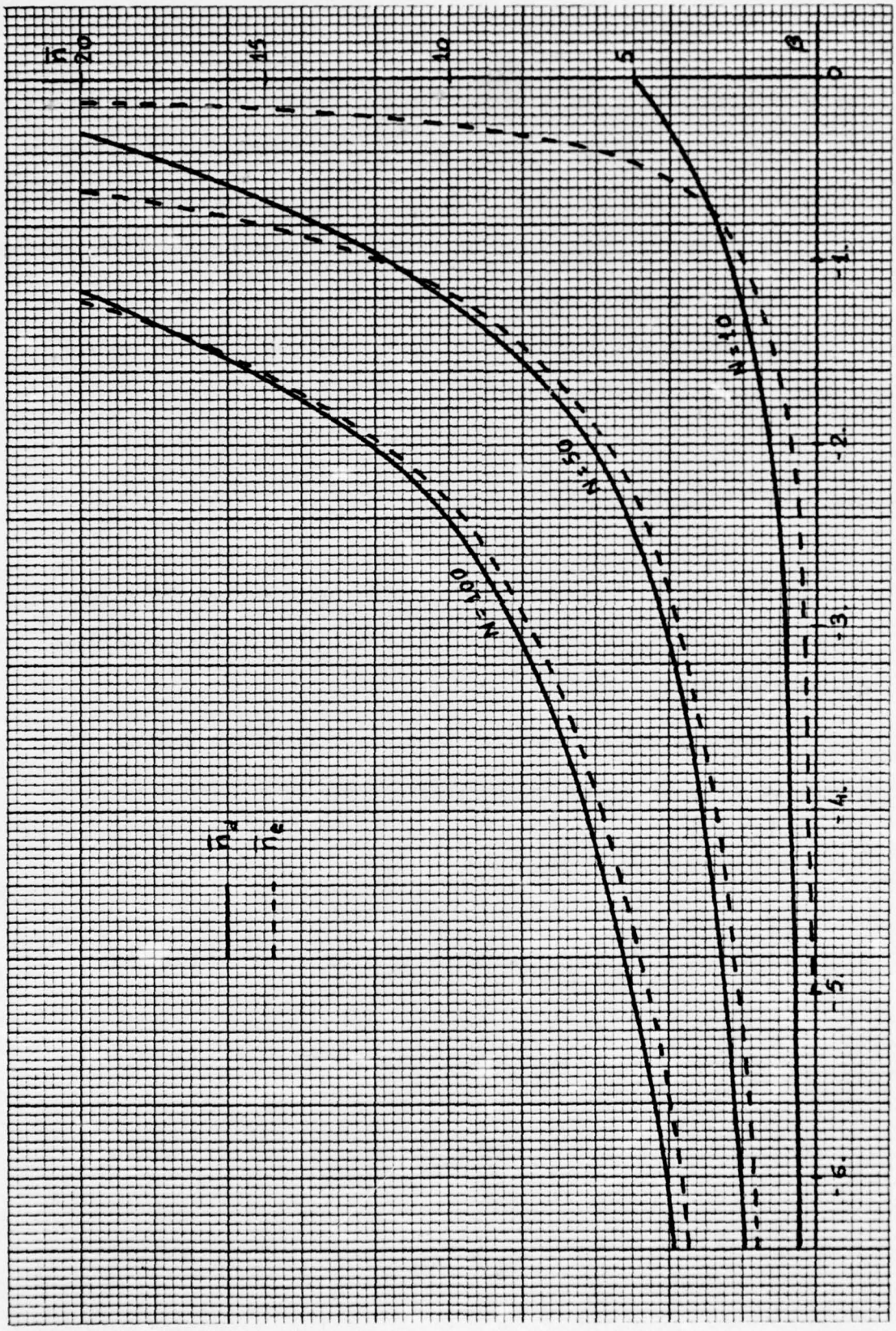


Fig. 6.1. Average length of a diffusion queue (solid lines) and an exponential M/M/1 queue (dashed lines) with the same drift β for different size of buffer N.

for $\beta = 0$ but their value at this point is the corresponding to N/2.

At some value of $\beta = \beta_0$, $\bar{n}_d$ and $\bar{n}_e$ intersect. The value β_0 decreases as N increases. For $\beta < \beta_0$ $\bar{n}_d$ and $\bar{n}_e$ differ approximately in 1/2 as we saw until $|\beta| = N\alpha$ in which case $\bar{n}_e = 0$ since this implies $\lambda = 0$. For N = 10 this happens for $\beta = 5$.

Therefore we can conclude that both models are very close when the M/M/1 queue is far from saturation ($\lambda = \mu$)

6.2.- System of two queues

In Fig. 6.2 it is reproduced the model of Fig. 5.2 where the queues $n_1(t)$ and $n_2(t)$ are of M/M/1 type with the indicated arrival and service rates.

The arrival and service processes are assumed independent. The expressions for the means and covariances per unit times are

$$\frac{E\left[n_1(t)\right]}{t} = \lambda_1 + \mu_{21} - \mu_1 - \mu_{12} \qquad (610a)$$

$$\frac{E\left[n_2(t)\right]}{t} = \lambda_2 + \mu_{12} - \mu_2 - \mu_{21} \qquad (6\text{-}10b)$$

$$\frac{Var\left[n_1(t)\right]}{t} = \lambda_1 + \mu_{21} + \mu_1 + \mu_2 \qquad (6\text{-}10c)$$

$$\frac{Var\left[n_2(t)\right]}{t} = \lambda_2 + \mu_{12} + \mu_2 + \mu_{21} \qquad (6\text{-}10d)$$

$$\frac{Cov\left[n_1(t)\ n_2(t)\right]}{t} = -\mu_{12} - \mu_{21} \qquad (6\text{-}10e)$$

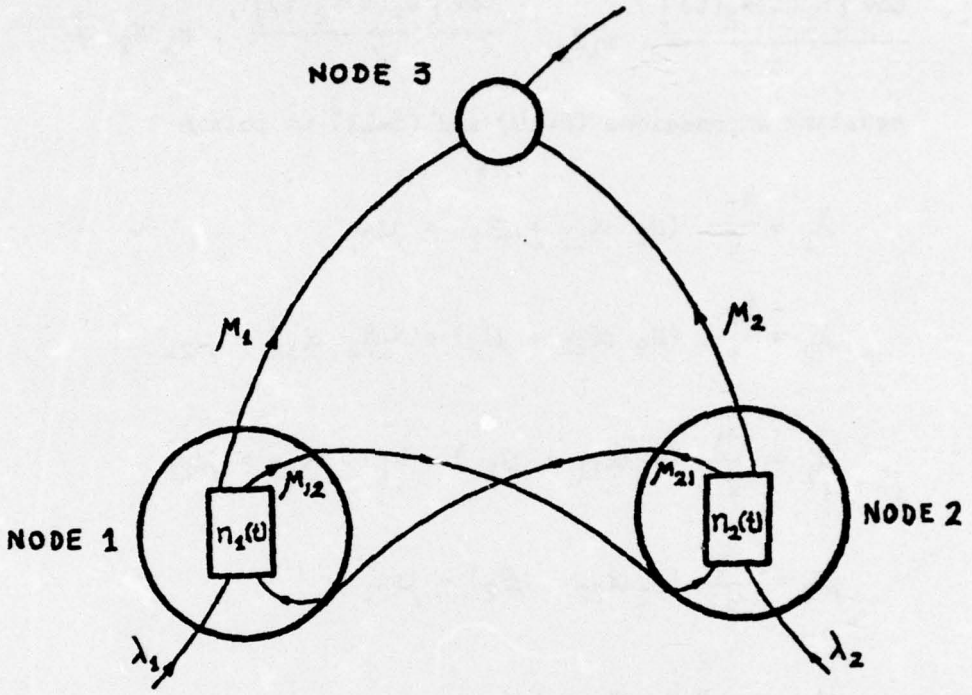


Fig. 6.2: Network of three nodes and two queues

For the diffusion model of Section 5.1, if the maximum queue lengths is N_1, N_2, respectively we have.

$$\frac{E\left[n_1(t)\right]}{t} = N_1 \quad \frac{E\left[x_1(t)\right]}{t} = N_1 \beta_1 \tag{6-11a}$$

$$\frac{E\left[n_2(t)\right]}{t} = N_2 \quad \frac{E\left[x_2(t)\right]}{t} = N_2 \beta_2 \tag{6-11b}$$

$$\frac{Var\left[n_2(t)\right]}{t} = N_1^2 \quad \frac{Var\left[x_1(t)\right]}{t} = N_1^2 \alpha_{11} \tag{6-11c}$$

$$\frac{Var\left[n_2(t)\right]}{t} = N_2^2 \quad \frac{Var\left[x_2(t)\right]}{t} = N_2^2 \alpha_{22} \tag{6-11d}$$

$$\frac{\mathrm{Cov}\left[n_1(t)n_2(t)\right]}{t} \underset{N_1N_2}{} \frac{\mathrm{Cov}\left[x_1(t)x_2(t)\right]}{t} = N_1 N_2 \alpha_{12} \qquad (6\text{-}11c)$$

equating expressions (6-10) and (6-11) we obtain

$$\lambda_1 = \frac{N_1}{2} (N_1 \alpha_{11} + \beta_1) - \mu_{21} \qquad (6\text{-}12a)$$

$$\lambda_2 = \frac{N_2}{2} (N_2 \alpha_{22} + \beta_2) + N_1N_2 \alpha_{12} + \mu_{21} \qquad (6\text{-}11b)$$

$$\mu_1 = \frac{N_1}{2} (N_1\alpha_{11} - \beta_1) + N_1N_2 \alpha_{12} + \mu_{21} \qquad (6\text{-}12c)$$

$$\mu_2 = \frac{N_2}{2} (N_2\alpha_{22} - \beta_2) - \mu_{21} \qquad (6\text{-}12d)$$

$$\mu_{12} = - N_1N_2 \alpha_{12} - \mu_{21} \qquad (6\text{-}12e)$$

where μ_{21} can be chosen freely provided the others are not negative.

The equivalent arrival and service rates of queues $n_1(t)$ and $n_2(t)$ are respectively

$$\lambda_1' = \lambda_1 + \mu_{21} \quad ; \quad \mu_1' = \mu_1 + \mu_{12} \qquad (6\text{-}13a)$$

$$\lambda_2' = \lambda_2 + \mu_{12} \quad ; \quad \mu_2' = \mu_2 + \mu_{21} \qquad (6\text{-}13b)$$

and therefore the corresponding mean values are:

$$\bar{n}_{1e} = \frac{\lambda_1'}{\mu_1' - \lambda_1'} = \frac{\lambda_1 + \mu_{21}}{\mu_1 + \mu_{12} - \lambda_1 - \mu_{21}} = \frac{N_1 \alpha_{11} + \beta_1}{-2\beta_1} = \qquad (6\text{-}14a)$$

$$= - N_1 \frac{\alpha_{11}}{2\beta_1} - \frac{1}{2}$$

$$\bar{n}_{2e} = \frac{\lambda_2'}{\mu_2' - \lambda_2'} = \frac{\lambda_2 + \mu_{12}}{\mu_2 + \mu_{21} - \lambda_2 - \mu_{12}} = \frac{N_2 \alpha_{22} + \beta_2}{-2\beta_2} =$$

$$= -N_2 \frac{\alpha_{22}}{-2\beta_2} - \frac{1}{2} \qquad (6\text{-}14b)$$

provided that

$$\beta_1 < 0 \quad ; \quad |\beta_1| < N_1 \alpha_{11} \qquad (6\text{-}15a)$$

$$\beta_2 < 0 \quad ; \quad |\beta_2| < N_2 \alpha_{22} \qquad (6\text{-}15o)$$

For the diffusion queues the corresponding mean values are

$$\bar{n}_{1d} = N_1 \bar{x}_1 = N_1 \left(\frac{e^{\gamma_1}}{e^{\gamma_1} - 1} - \frac{1}{\gamma_1} \right) \qquad (6\text{-}16\text{-}a)$$

$$\bar{n}_{2d} = N_2 \bar{x}_2 = N_2 \left(\frac{e^{\gamma_2}}{e^{\gamma_2} - 1} - \frac{1}{\gamma_2} \right) \qquad (6\text{-}16b)$$

where γ_1 and γ_2 are the components of the vector $\underline{\gamma} = 2\underline{\Lambda}^{-1}\underline{\beta}$ that is:

$$\gamma_1 = 2 \frac{\alpha_{22}\beta_1 - \alpha_{12}\beta_2}{\alpha_{11}\alpha_{22} - \alpha_{12}^2} \qquad (6\text{-}17a)$$

$$\gamma_2 = 2 \frac{\alpha_{11}\beta_2 - \alpha_{12}\beta_1}{\alpha_{11}\alpha_{22} - \alpha_{12}^2} \qquad (6\text{-}17b)$$

When γ_1 and γ_2 are less than -3, the expressions (6-16) can be approximated by

$$\bar{n}_{1d} \approx -N_1 \frac{1}{\gamma_1} = -N_1 \frac{\alpha_{11} - \frac{\alpha_{12}^2}{\alpha_{22}}}{2\left(\beta_1 - \frac{\alpha_{12}}{\alpha_{22}}\beta_2\right)} \qquad (6\text{-}18a)$$

$$\bar{n}_{2d} \approx -N_2 \frac{1}{\gamma_2} = -N_2 \frac{\alpha_{22} - \frac{\alpha_{12}^2}{\alpha_{11}}}{2\left(\beta_2 - \frac{\alpha_{12}}{\alpha_{11}}\beta_1\right)} \qquad (6\text{-}18b)$$

Compare expressions (6-15) and (6-18). The difference now is not only the term 1/2 in (6-15) but the terms affected by the element α_{12} of the covariance which did not show up in (6-15).

Let us take for instance the values of the first example of Section 5.1, that is $\alpha_{11} = 0.56$, $\alpha_{22} = 0.64$, $\alpha_{12} = -0.16$ and assume $N_1 = N_2 = 10$.

Take $\delta_2 = p_{23} - q_{23} = -1$ and $\delta_1 = p_{13} - q_{13}$ variable. The value of z^* that minimizes the overall mean value is represented in Fig. 6.3 as a function of δ_1. The corresponding values of β_1 and β_2 are given by expression (5-22) that is

$$\beta_1 = \delta_1 - z^* \qquad (6\text{-}19a)$$

$$\beta_2 = \delta_2 + z^* = -1 + z^* \qquad (6\text{-}19b)$$

and the corresponding γ_1 an γ_2 can be obtained from expressions (6-17)

We want to compare the expressions of the means corresponding to

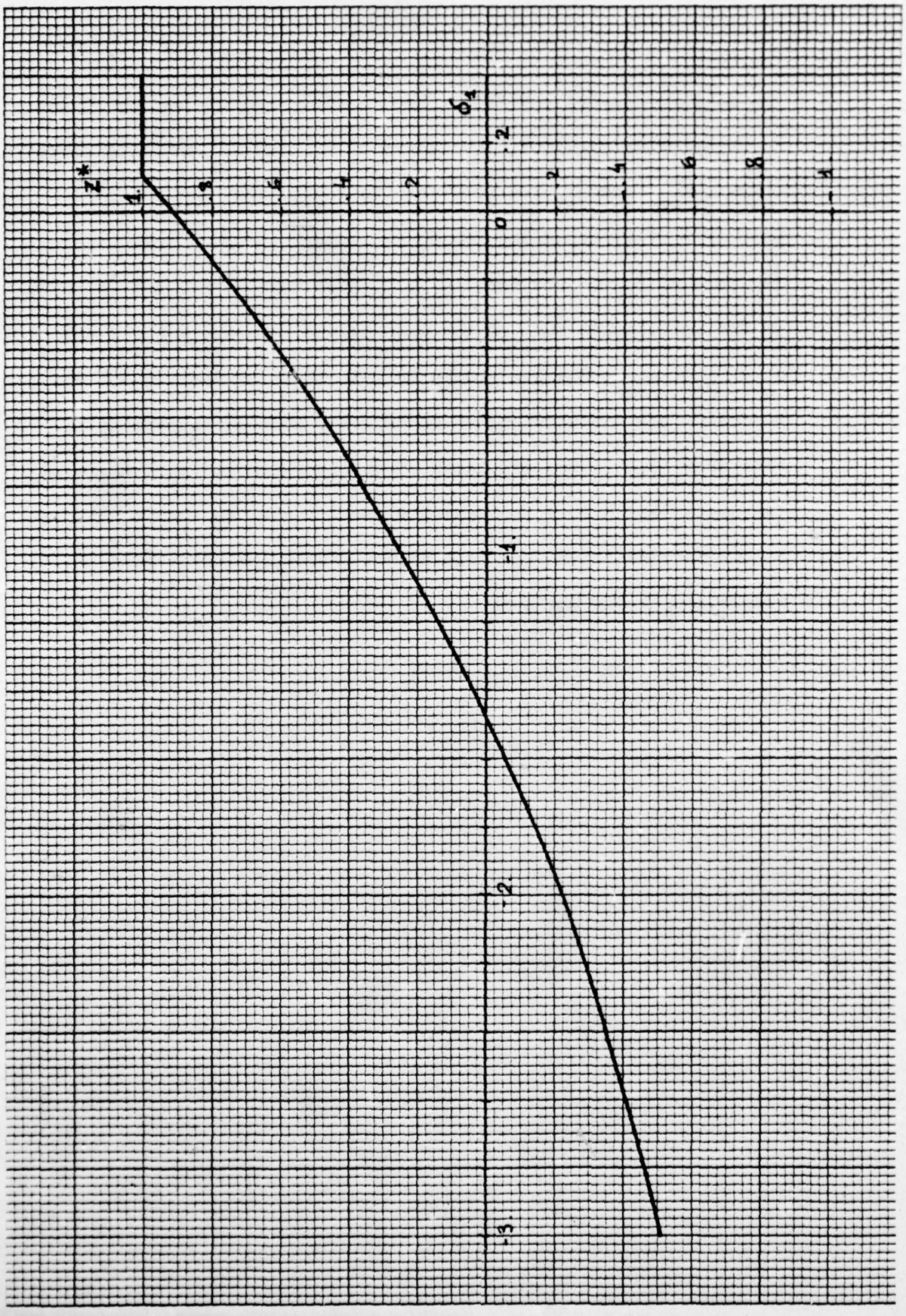


Fig. 6.3. Variation of z^* in terms of δ_1 for minimum average queue length in example 5.1.

the diffusion model and to the M/M/1 model when δ_1 varies. This is depicted in Fig. 6.4. The values of $\bar{n}_{d1}$ and $\bar{n}_{e1}$ are very close whereas $\bar{n}_{d2}$ and $\bar{n}_{e2}$ are not. This is due to the fact that the pole of $\bar{n}_{e1}$ corresponding to $\beta_1 = 0$ occurs at $\delta_1 = 1$ whereas the pole of $\bar{n}_{e2}$ occurs for $\delta_1 = 0.05$ so that $x^* = 1$ and $\beta_2 = 0$. In the same figure sum of the mean values $\bar{n}_e = \bar{n}_{31} + \bar{n}_{e2}$ and $\bar{n}_d = \bar{n}_{d1} + \bar{n}_{d2}$. Similar conclusions to those of Section 6.1 can be drawn: the diffusion model cancels the pole corresponding to zero drift due to the inclusion of an upper barrier that prevents the queue length from increasing without limit. As the drifts become more negatives the mean values corresponding to both models are nearly the same.

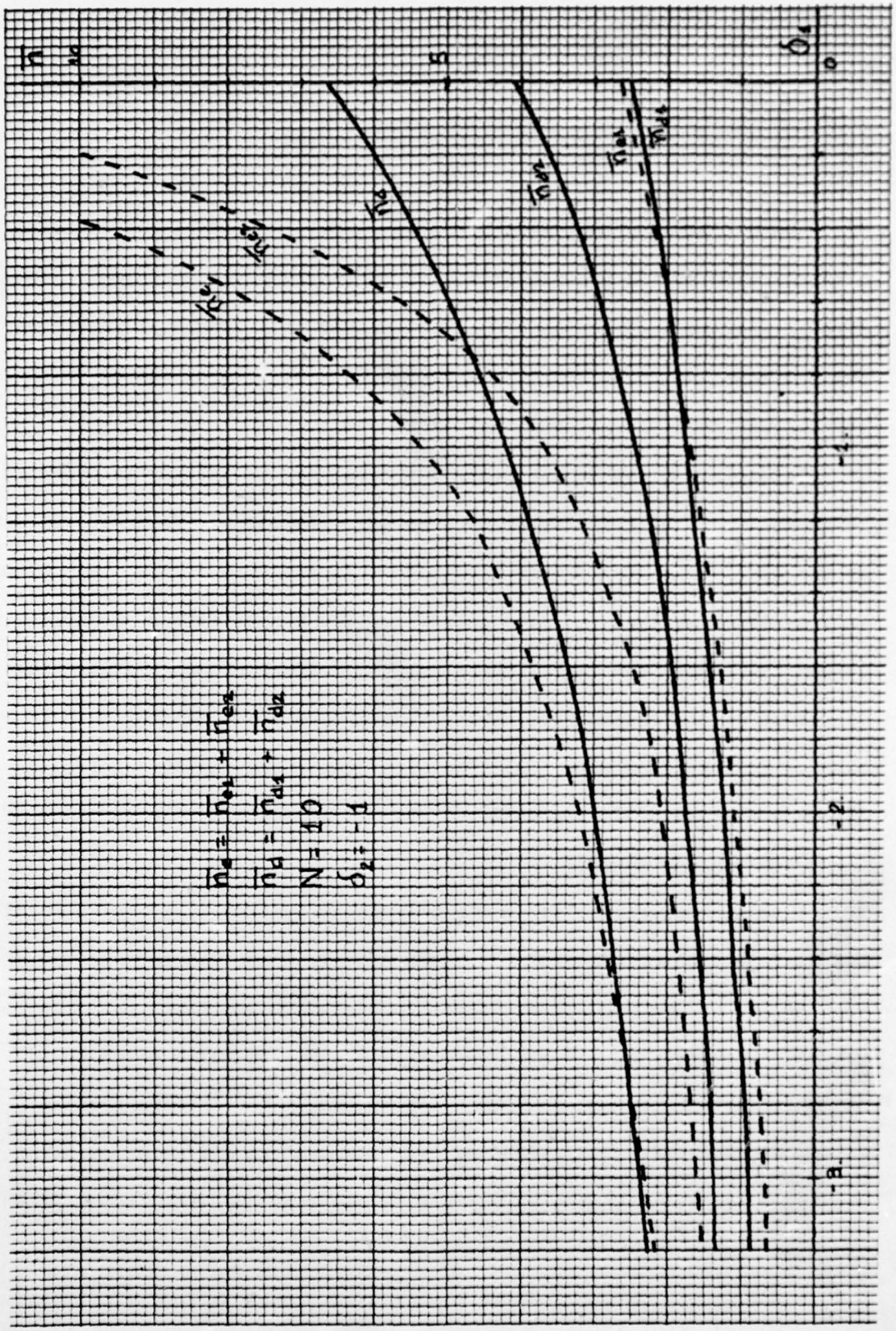


Fig. 6.4. Comparison between the diffusion model results (solid lines) and the M/M/1 model (dashed lines) for example 5.1.

7.- CONCLUSIONS AND SUGGESTIONS FOR FURTHER WORK

The diffusion approximation can be useful for modeling a computer communication system because it allows to consider the queueing processes like if they were of continuous nature rather than discrete. Therefore the increase in the number of states for complex systems can be avoided simplifying then the mathematical treatment.

In Section 2 the requirements a process has to satisfy in order to be approximated by a diffusion process was established. This was obtained as the limit of a random walk processes when the size of the steps became very small as well as the time intervals. The probabilities of jumping up and downwards were nearly the same, the difference being a quantity dependent on the square root of the time interval and tending then to zero as that becomes very small. The size step θ and the time intervals Δt were related by an expression: $\Delta t = K\theta^2$ so that the variance of the process makes sense.

This approximation led to the diffusion equation (2-5) which relates the distribution of the continuous-state continuous-time process by mean of its derivatives with respect to the state and time. Our interest was addressed to find the steady-state distribution which gives an idea of what is to be expected in the long run and after the transient period has died off. The procedure was to solve the diffusion equation when the derivative with respect to time was zero and with the appropiate boundary conditions.

The same can be applied for a multidimensional diffusion process

representing all the single diffusion processes that take place in a network of queues.

It is to be remarked that the boundaries (called "barriers" in diffusion literature) come from the physical fact that the queueing processes represent the messages waiting to be served and therefore a lower barrier must exist since the number of messages cannot be less than zero. An upper berrier was provided too for each process because finite length buffers were assumed.

In order to optimize the system operation a performance criterion had to be taken. This was to consider the sum of all the expected queue lengths whose expressions are quite straightforward once the length distribution is calculated via diffusion equation.

In Section 4 a general model for routing messages in a computer communication network was established. Then a procedure to calculate the system parameters is provided. The system parameters are all the coefficients of the diffusion equation, namely the mean per unit time of each queue and the elements of the covariance matrix per unit time.

Once the system parameters are known, the expression of the overall mean can be found and the problem consists of minimizing that expression by properly chosing the system parameters and subject to the specific constraints of the problem.

In Section 5 several examples are shown for ilustration for systems of 2, 4 and 6 queues. The last one though, is a modification of the general model that was established in Section 4. The difference is that the computer network is considered as a computer load sharing

system rather than a general routing network. Nevertheless the diffusion model can be applied to this case too by considering as generalized "channels" not only the transmission links but the computer processor themselves.

Finally, a comparison is made between the established diffusion model and the exponential M/M/1 for a single queue and for a system of two queues. The main difference between the two models arises from the fact that the M/M/1 has no limit for the queue length and therefore a pole appears for equal arrival and service rates (null drift). The more nagative the drift becomes the more both models resemble and this resemblance is more evident for larger queue capacity N.

Suggestions for further work

It was said in Section 4.5 that although the minimization problem was dependent on the elements of $\underline{\beta}$ (vector drift per unit time) and $\underline{\Lambda}$ (covariance matrix per unit time) we considered that one as fixed and carried the minimization over the elements of $\underline{\beta}$. This was because the computations involved became easier. It would be desiderable to accomplish the minimization problem by including in it the elements of $\underline{\Lambda}$. The difficulty arises from the fact that even though the constraints of the problem are linear, the function to be minimized is not linear but exponential.

It could also be interesting to find other algorithms for the minimization problem that take advantage of the system structure in order to investigate and eventually provide faster convergence.

REFERENCES

[1] D.G. Cantor & M.G. Gerla, "Optimal Routing in a Packet-switched Computer Network", IEEE Tr. on Computers, Vol. C-23, No. 10, oct. 1974.

[2] D. Cox & H. Miller, The Theory of Stochastic Processes Wiley. N. York, 1965

[3] W. Feller, An Introduction to Probability Theory and its Applications, Vol. II, Wiley, N. York 1966

[4] D.P. Gaver, "Diffusion Approximations and Models for Certain Congestion Problems", J. Appl. Prob. S, 1968.

[5] D.P. Gaver & G.S. Shedler, "Processor Utilization in Multiprograming Systems via Diffusion Approximations" IBM Research Rep. RJ-938, Nov 1971

[6] W.J. Gordon & G.F. Newell, "Closed Queueing Systems with Exponential Servers", Oper. Research 15, 2 April 1967

[7] D. Gross & C.M. Harris, Fundamentals of Queueing Theory Wiley. N. York, 1974

[8] J.R. Jackson, "Networks of Waiting Lines", Oper. Research Vol. 5, 1957

[9] J.R. Jackson, "Jobshop-Like Queueing Systems" Management Science, Vol, 10, 1963

[10] J.F.C. Kingman, "The Heavy Traffic Approximation in the Theory of Queues", Proc. of the Symposium on Congestion Theory", W.L. Smith & W.E. Wilkinson, Eds. U. of North Carolina Press, 1965.

[11] L. Kleinrock, Communication Nets, Mc. Graw Hill 1964

[12] L. Kleinrock, Queueing Systems, Vol. I: Theory, J. Wiley 1975

[13] H. Kobayashi, "Application of the Diffusion Approximation to Queueing Networks I: Equilibrium Distributions" Jour. ACM, April 1974

[14] H. Kobayashi, "Application of the Diffusion Approximation to Queueing Networks II: Nonequilibrium Distribution and Computer Modeling", Jour. ACM, July 1974

[15] J.D.C. Little, "A Proof for the Queueing Formula L = λw" Opr. Research 9, 1968

[16] G.F. Newell, "Queues with Dependent Arrival Rates" J. Appl. Probability 5, 1968

[17] E. Parzen, Stochastic Processes, Holden-Day, S. Francisco 1962

[18] M.J.D. Powell, "An Efficient Method for Finding the Minimum of a Function of Several Variables Without Calculating Derivatives", Comp. Jour. Vol. 7, 1974

[19] M. Reiser and H. Kobayashi, "Accuracy of the Diffusion A Approximation for Some Queueing Systems" IBM, J, Res. Develop. 18, 2 March 1974

[20] M. Reiser and H. Kobayashi "Queueing Networks with Multiple Closed Chains: Theory and Computational Algorithms" IBM Res. Rep. RC 4919, Jul. 1974

[21] W.D. Roome & H.C. Torng. "Modeling and Design of Computer Networks with Distributed Computer Facilities", Proceed. of the 1974 Symposium, Computer Networks: Trends and Applications, IEEE Computer Society, 1974

[22] T.L. Saaty, Elements of Queueing Theory with Applications, Mc. Graw Hill, N. York, 1961

[23] A. Segall, "New Analytical Models for Dynamic Routing in Computer Networks", Nat. Tel Conference, New Orleans 1975

[24] A Sommerfeld, Partial Differential Equations in Physics, Academic Press, N. York 1949

[25] A.L. Sweet & J.C. Hardin, "Solution for Some Diffusion Processes with two Barriers", J. Appl. Prob. 7, 1970

[26] E.F. Wunderlich, "Load Sharing in a Computer-Communication Network", M.I.T. M.S. Thesis , Sept. 1975

[27] W.I. Zangwill,"Minimizing a Function Without Calculating Derivatives" Comp. Jour. Vol 10, 1967

Distribution List

Defense Documentation Center Cameron Station Alexandria, Virginia 22314	12 copies
Assistant Chief for Technology Office of Naval Research, Code 200 Arlington, Virginia 22217	1 copy
Office of Naval Research Information Systems Program Code 437 Arlington, Virginia 22217	2 copies
Office of Naval Research Code 1021P Arlington, Virginia 22217	6 copies
Office of Naval Research Branch Office, Boston 495 Summer Street Boston, Massachusetts 02210	1 copy
Office of Naval Research Branch Office, Chicago 536 South Clark Street Chicago, Illinois 60605	1 copy
Office of Naval Research Branch Office, Pasadena 1030 East Green Street Pasadena, California 91106	1 copy
New York Area Office (ONR) 715 Broadway - 5th Floor New York, New York 10003	1 copy
Naval Research Laboratory Technical Information Division, Code 2627 Washington, D.C. 20375	6 copies

Dr. A. L. Slafkosky
Scientific Advisor
Commandant of the Marine Corps (Code RD-1)
Washington, D.C. 20380 — 1 copy

Office of Naval Research
Code 455
Arlington, Virginia 22217 — 1 copy

Office of Naval Research
Code 458
Arlington, Virginia 22217 — 1 copy

Naval Electronics Laboratory Center
Advanced Software Technology Division
Code 5200
San Diego, California 92152 — 1 copy

Mr. E. H. Gleissner
Naval Ship Research & Development Center
Computation and Mathematics Department
Bethesda, Maryland 20084 — 1 copy

Captain Grace M. Hopper
NAICOM/MIS Planning Branch (OP-916D)
Office of Chief of Naval Operations
Washington, D.C. 20350 — 1 copy

Mr. Kin B. Thompson
Technical Director
Information Systems Division (OP-91T)
Office of Chief of Naval Operations
Washington, D.C. 20350 — 1 copy

Advanced Research Projects Agency
Information Processing Techniques
1400 Wilson Boulevard
Arlington, Virginia 22209 — 1 copy