SOME CONTRIBUTIONS TO THE FIELDS OF INSENSITIVITY AND QUEUEING THEORY

by

Michael P. Rumsewicz, B.Sc.(Hons.) (Maths.Sc.)

Thesis submitted for the degree of
Doctor of Philosophy
in the Department of Applied Mathematics,
University of Adelaide

February, 1988.

CONTENTS

SUMMARY

This thesis is in two parts.

In PART ONE, insensitivity in stochastic processes is studied.

Chapter 1 contains a review of the literature and introduces the terminology and ideas involved.

In Chapter 2 we summarise the results of previous authors that are required in Chapters 3 and 4.

Insensitivity in processes with age dependent routing is studied in Chapter 3. Conditions are found for classes of age dependent processes to have the same equilibrium distribution. These conditions are related to the property of partial balance in systems which "average" the age dependent routing probabilities. The analysis is then applied to networks of queue, semi-Markov processes and interruption processes. Average residence times and systems with age dependent speeds are also examined.

In Chapter 4 we consider insensitivity in generalised semi-Markov processes which do not necessarily possess instantaneous attention. An alternative proof is given of a theorem of Taylor (1987), and results on the relationship between generalised balance and insensitivity are extended to such processes.

In PART TWO of this thesis, closed, two node, priority queueing networks are examined.

Chapter 5 contains a review of the literature on priority queueing systems.

In Chapter 6, we study closed, two node networks with a variety of priority disciplines. When both queues use pre-emptive priority, results of Morris (1981)

are extended to the case of state dependent service rates. The equilibrium distribution is also found for networks with one pre-emptive and one nonpre-emptive queue and the solution compared to an approximation used by Morris (1981). An insensitivity result is derived for pre-emptive networks with priorities reversed.

Chapter 7 contains conclusions and some ideas for further research.

SIGNED STATEMENT

This thesis contains no material which has been accepted for the award of any other degree or diploma in any University.

To the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

I consent to this thesis being made available for photocopying and loan.

Michael P. Rumsewicz

ACKNOWLEDGEMENTS

I would sincerely like to thank my supervisor, Bill Henderson, for his patience, guidance and harassment during my time as a postgraduate. Thanks also go to Peter Taylor for many long discussions and Gillian Read for being a source of laughs. Finally, I wish to thank parents for giving me this opportunity and Carolyn for her eternal support and encouragement over the last few years.

PART ONE

INSENSITIVITY IN STOCHASTIC PROCESSES

CHAPTER 1 : INTRODUCTION

Many stochastic processes have features that do not depend upon the shapes of the governing distributions. For example, the probability of emptiness in the $G/G/1$ queue depends only on the means of the arrival and service time distributions. Characteristics like this allow information on a general system to be gained from the more readily analysed purely Markov process. In this part of the thesis, we seek criteria under which the equilibrium distribution of the processes under examination are insensitive, that is, depend on the lifetime distributions only through their means.

Erlang (1917), looked at a simple model of a telephone exchange, the famous $M/G/K$ Erlang loss system, and showed that the equilibrium distribution for the number of busy servers is the same when either negative exponential or deterministic distributions with the same mean are used. Later, it was shown that the same result also holds with Erlang distributed service times, and hence it was postulated that this was true for arbitrary distributions with the same mean.

Fortet (1950) proved this postulate but did not show the uniqueness of the solution obtained. This was performed by Sevast'yanov (1957), with the assumption that the successive service times are independent, an assumption later removed by König and Matthes (1963). Takacs (1969) looked at the problem in terms of the number of busy servers a customer sees upon arrival and found similar results. The results of Takacs were extended to state dependent service and arrival rates by Brumelle (1978).

A related process, the Engset system, was shown by Cohen (1957) to be insensitive while König (1965) also showed insensitivity but allowed successive service times to be generated by a stationary point process.

The evolution of insensitivity theory gained impetus with the introduction of the Generalised Semi-Markov Process (GSMP) by Matthes (1962). A GSMP is defined on a set of states $x \in \Omega$. Associated with each state $x \in \Omega$ are active elements from

the set $S = S^* \cup S'$ where $S^* \cap S' = \phi$ and $|S^*|$ is finite. If $s \in S^*$ then the lifetime of s is generally distributed whereas if $s \in S'$ it is negative exponentially distributed. It is assumed that no two elements of S^* are activated or die simultaneously and the residual lifetimes of the remaining active elements of S^* remain unchanged upon the death of $s \in S^*$. With probability $p(x, s, x')$ the process moves from state x to state x' upon the death of element $s \in S$.

Matthes showed that a GSMP is insensitive if and only if a particular set of partial balance equations (his set of equations Z) are satisfied. That is, the equilibrium distribution depends only on the means of the lifetime distributions for elements in S^* when, in the purely Markov process, the flow into a state due to the creation of $s \in S^*$ is equal to the flow out of that state due to the death of s.

This was a major step in insensitivity theory as many physical processes can be modelled within the GSMP framework. Hence, insensitivity criteria became available for processes such as the Erlang loss system and the Engset system.

Extensions to this work, which include allowing successive lifetimes to be generated by stationary point processes, were done by König and Matthes (1963), König (1965) and König, Matthes and Nawrotzki (1967).

König and Jansen (1974) widened the scope of GSMPs by having the active lifetimes worked off at state dependent speeds, allowing processes with temporary interruptions to be modelled. In analysing such systems they assumed what they term the e_1 *property*. This property (later called instantaneous attention by Schassberger (1977)) simply means that any generally distributed lifetime must be worked on with positive speed as soon as it is created. Under this assumption, König and Jansen (1974) demonstrated that a GSMP is insensitive if and only if a certain set of partial balance equations hold for the associated purely Markov process. These partial balance equations have the same interpretation as in Matthes' (1962) work and differ only through the introduction of the speeds. Jansen, König and Nawrotzki

(1979) and Franken, Arndt, König and Schmidt (1982) extended this work to the case where lifetimes are generated by stationary point processes.

König and Jansen (1974) pointed out that processes not possessing instantaneous attention may, by the addition of suitable extra states, be converted to a process that does have this property and hence permit the utilisation of their analysis. Taylor (1987) adopted a direct approach to this problem which is discussed later in more detail.

The original proofs of the above results are quite complex and require a knowledge of stationary point processes. Schassberger (1977, 1978a, 1978b) produced simplified proofs of the results of Matthes (1962) and König and Jansen (1974) by employing the method of phases and using a weak convergence argument to show that the results hold for general distributions. Whitt (1980) gave a rigorous justification for the use of this technique to model general distributions in GSMPs.

König and Jansen (1974) and Schassberger (1978b) also showed that the supplemented equilibrium density (using residual lifetime) breaks down into a product of two terms, one for the discrete part of the state and one for the continuous part, the latter factorising into terms for each active element of S^* in the state. That is, the process behaves as if the discrete and continuous parts of the state are independent.

Henderson (1983a) introduced the property of generalised balance. A GSMP (having instantaneous attention) possesses this property if in any state the rate at which each active element dies in forward time is equal to the rate at which it dies in reverse time. It was shown by Henderson that GSMPs with speeds are insensitive if and only if they have generalised balance. Moreover, Henderson showed that this is true if and only if the equilibrium density using residual supplementary variables is the same as that using spent supplementary variables. He derived further necessary and sufficient conditions for insensitivity which relate the forward time process to the

reverse time process. These results are extended in Chapter 4 to allow for processes not having instantaneous attention.

A new structure for the examination of insensitivity was developed by Whittle (1985). Using this structure, Whittle provided a new and simpler proof of the insensitivity if and only if partial balance result. Taylor (1987) noted that it is implicitly assumed that the process under consideration possesses instantaneous attention and that active elements with generally distributed lifetimes may not die and be immediately recreated. Subsequently, Taylor (1987) reproved Whittle's results without these assumptions. Schassberger (1986) pointed out that there exists an equivalence between Whittle's structure and GSMPs. Further results on insensitivity may be found in Whittle (1986).

Henderson and Taylor (1987) considered a modified form of a GSMP with speeds by allowing generally distributed lifetimes to be terminated by a negative exponentially distributed interruption. The general lifetime then dies and the process moves to some new state dependent upon the manner in which the lifetime died, that is, naturally or due to an interruption. They derived necessary and sufficient conditions for the insensitivity of such systems which are related to the partial balance equations for an associated Markov process. The routing probabilities of this Markov process are the average routing probabilities of the interruption process. Such processes are special cases of the processes with age dependent routing to be considered in Chapter 3.

As mentioned previously, Taylor (1987) took a direct approach to the analysis of processes that do not have instantaneous attention. He found necessary and sufficient conditions for insensitivity in processes with one generally distributed lifetime and sufficient conditions for processes with many general elements. These conditions have an interesting interpretation. Whereas previous results had found a balance exists between the birth and death of generally distributed elements, Taylor showed that

the important property is a balance between the first time an element of S^* is worked on with positive speed and the death of this element. The supplemented equilibrium density using spent lifetimes is also given. In Chapter 4 we give an alternative analysis of such processes which highlights the physical, rather than mathematical, reasons behind Taylor's result.

Much work has been done, using a variety of techniques, in the area of insensitivity in queueing networks. Jackson (1957) showed that the equilibrium distribution for an open queueing network with Poisson arrivals and negative exponential service times could be broken up into a product over the nodes of the network, each term in the product corresponding to the equilibrium distribution of that node when considered in isolation with a modified arrival rate. The proof of this result was performed by showing that the proposed solution satisfied the global balance equations. A similar result for closed networks was shown by Gordon and Newell (1967).

The first of the insensitivity results in queueing networks was produced by Baskett, Chandy, Muntz and Palacios (1975). They considered open, closed and mixed networks of queues with multiple customer types arriving in Poisson streams. The nodes are permitted to be of four types, namely,

(1) a single server queue operating under a last come, first served, pre-emptive resume discipline,

(2) an infinite server queue,

(3) a processor sharing queue, in which a single server devotes an equal amount of service to each customer at the queue,

(4) a single server operating under a first come first served discipline.

For nodes of types (1), (2) and (3) the service times are allowed to be dependent on the type of customer, but must have rational Laplace transforms. For nodes of type (4), all customers must have a negative exponentially distributed service time with the same mean.

The condition that the service time distributions have rational Laplace transforms allows a Coxian description of the progress of the customer at those nodes. By describing the state of the system by the number of customers of each type at each queue, together with the possible stages of service, Baskett, *et al*, showed that the equilibrium distribution is given by a product over the nodes of the network. Using an appropriate summation, it was shown that the marginal distribution for the number of customers of each type at each queue again takes on a product form, and this distribution depends on the service times only through their means.

These results were proved by showing that the proposed product form satisfies a certain set of local balance equations which equate the rate of flow into a state by a customer entering a particular stage of service with the rate of flow out of that state due to a customer leaving that stage of service. When appropriately summed, these local balance equations give the global balance equations.

Kelly (1976) looked at networks of queues from an entirely new perspective. He firstly considered general queueing networks with negative exponential service times. By proposing an equilibrium distribution of product form and a set of reversed time transition rates, Kelly showed that the equations

$$\pi(C)q(C,D) = \pi(D)q'(D,C)$$

and

$$\sum_D q(C,D) = \sum_D q'(C,D)$$

were satisfied, where $\pi(\cdot)$ is the proposed equilibrium distribution and $q(C,D)$ (respectively $q'(C,D)$) is the forward (respectively reversed) time transition for moving from state C to state D. Satisfying these equations shows that the "guessed" equilibrium distribution and transition rates are, in fact, correct.

Kelly also introduced the concept of symmetric queues. The queue is said to be symmetric if, for every $n > 0$, the proportion of the service facility's total effort,

$\gamma(l,n)$, devoted to the customer in position l, $l = 1,\ldots,n$, when n customers are present, is equal to the probability, $\delta(l,n)$, that an arrival finding $n-1$ customers present moves into position l. Kelly showed that if, at symmetric queues, the service times come from finite mixtures of Erlang distributions, the equilibrium distribution depends only upon the mean service times, not the shape of the distribution. His conjecture that the same results hold for general service time distributions was proved by Barbour (1976). By appropriately defining the parameters of the system, the networks of Baskett, *et al*, fit into the above structure.

Chandy, Howard and Towsley (1977) showed similar results by using continuous supplementary variables for the remaining service requirement of each customer. It should be noted that the conditions under which their theorem is proved are somewhat restrictive as they require that the queues of closed networks be locally balanced when considered in isolation with Poisson arrivals. In Chapter 6, Section 4, an example is given of a closed network that appears to have such a product form solution, yet whose nodes are not necessarily locally balanced in isolation.

A queue is said to be *quasi-reversible* (Kelly (1979)) if its state $\mathbf{x}(t)$ at time t is a stationary Markov process with the property that $\mathbf{x}(t_0)$ is independent of

(1) arrival times of type c customers subsequent to time t_0,

(2) departure times of type c customers prior to time t_0.

All of the networks described above are made up solely of quasi-reversible queues. Kelly (1979) showed that such systems have a product form equilibrium distribution which, at symmetric queues, depends only on the mean of the service time distributions. He also showed that a queue is quasi-reversible if and only if the total service effort devoted to type j customers at queue i, when n customers are present, takes on the form $\phi_i(\mathbf{n}(i) - \mathbf{e}_{ij})/\phi_i(\mathbf{n}(i))$, where $\mathbf{n}(i)$ is a vector of the number of each customer type at queue i and $\mathbf{n}(i) - \mathbf{e}_{ij}$ is the state with one less type j customer present.

Chandy and Martin (1983) extended the definition of symmetric queues by allowing γ and δ to be functions of the state of the entire network; queue i is then said to be symmetric if $\delta_i(l, \mathbf{n}) = \gamma_i(l, \mathbf{n})$, where $\mathbf{n}$ is the state of the network. They showed that a more general product form of solution exists for all routing probabilities and arrival rates if and only if the queueing discipline is balanced and either

(1) the service times of all customers at a queue are negative exponentially distributed with the same mean, or

(2) the queue is symmetric.

The balanced condition implies that the total service effort of queue i devoted to type j customers takes on the form $\phi(\mathbf{n} - \mathbf{e}_{ij})/\phi(\mathbf{n})$ for all i and j where $\phi(.)$ is an arbitrary, non-negative function and $\mathbf{n} - \mathbf{e}_{ij}$ is the state with one less type j customer in queue i than in state $\mathbf{n}$ (c.f. Kelly (1979)).

Noetzel (1979) also looked at networks of queues but considered customers in arrival order, rather than inserting them in positions other than the end of the queue. It was shown that the product form holds for disciplines of the Last Batch Processor Sharing type. There does, however, exist an equivalence between the structure of Noetzel and the symmetric queues of Kelly. Other work on insensitivity and product form has been performed by Hordijk and van Dijk (1981, 1983a, 1983b).

All of the work on GSMPs described earlier assumes that the number of generally distributed elements is finite. As a result, closed networks can be modelled as GSMPs but open networks cannot. In a GSMP each element that is alive must have its own, unique label, with newly created elements randomly selecting one of the unused labels. To completely describe an open network, a countably infinite number of labels are required and hence random selection becomes impossible.

Barbour (1982) tackled this problem by showing that an open network may be realised as the limit of a suitable sequence of closed networks in such a way that the insensitivity properties of the closed network are transferred to the open network.

The comment is made, however, that it is probably easier to treat open networks as a class of problems in their own right.

In an attempt to deal with this shortcoming, Schassberger (1986) showed that it is possible to construct a GSMP with relabelling, in which, upon the death of an element with a generally distributed lifetime, all elements with the same distribution may be relabelled. This then permits the modelling of open networks.

There are a number of examples of finite queueing systems which are insensitive but may not be described as GSMPs nor fall within the networks of queues framework described above, as the governing lifetimes are not necessarily independent (in fact, they may be constrained to be identical). Such processes do not fall within the structure to be considered in this thesis but further information may be found in Jacobi (1965), Chaiken and Ignall (1972), Wolff and Wrightson (1976), Jansen (1980) and Henderson (1983b).

In Chapter 2 we present the basic notation and structures to be used in Chapters 3 and 4, together with a statement of some results of previous authors.

In Chapter 3 we shall extend the notion of GSMPs with speeds to allow routing probabilities which, upon the death of a generally distributed element, are dependent on the amount of service that has been received by that element. This then allows a more comprehensive extension of Semi-Markov Processes, in which the new state is chosen as a function of the time spent in the previous state.

The standard supplemented global balance equations are modified to include both the spent and residual lifetimes of each generally distributed element. In this way, correlations between the spent and residual equilibrium distributions of insensitive processes follow quite naturally. The theory is then applied to networks of queues, interruption processes and Semi-Markov Processes.

Finally, we look at average residence times in GSMPs and consider the extension of the above to processes with age dependent speeds.

In Chapter 4, processes not possessing instantaneous attention are examined and an alternative proof is given of a theorem of Taylor (1987). We then proceed to extend theorems of Henderson (1983a) to such processes.

CHAPTER 2 : NOTATION AND PRELIMINARY RESULTS

2.1 Introduction

In this chapter we introduce the basic notation to be used in this part of the thesis and review the main results on insensitivity in Generalised Semi-Markov Processes that will be employed. This shall be done in two parts. The first part, Section 2.2, considers continuous time stochastic processes with one generally distributed lifetime. This will enable us to easily introduce the concept of age dependent routing in Chapter 3 and will provide the structure for the processes to be examined in Chapter 4. The second part, Section 2.3, extends the class of processes to include those with many generally distributed lifetimes.

2.2 Insensitivity in processes with one general lifetime

Consider an irreducible, stationary process P on a set Ω, and let A be a finite subset of Ω with $\overline{A}$ its complement. Incorporated in each state $x \in \Omega$ are active elements from the set $S = S' \cup \{t\}, t \notin S'$, which decay at rate $c(s,x)$, $s \in S$. If $s \in S'$, the lifetime of s has a negative exponential distribution with mean λ_s^{-1}. If $x \in A$ then it also has associated with it the element t, whose lifetime is generally distributed with differentiable distribution function $G(.)$, mean μ^{-1}, density function $g(.)$ and hazard function $h(.)$. When an active element $s \in S$ dies, the process moves to state $x' \in \Omega$ with probability $p(x,s,x')$. For notational ease, partition A into disjoint subsets A_1 (states in A where the generalised lifetime is being worked off with positive speed, that is, $c(t,x) > 0$) and A_0 (states in A where the generalised lifetime is not being worked off, that is, $c(t,x) = 0$).

Definition 2.1 : The system P is insensitive if the equilibrium distribution for time spent in each state is dependent on the general distribution only through its mean.

Let π_x denote the probability of being in state x when $G(y) = 1 - \exp(-\mu y)$, that is, when P is a purely Markov process.

The global balance equations for this purely Markov process are

$$\pi_x \sum_{s \in S' \cap x} \lambda_s c(s,x) = \sum_{x' \in \overline{A}} \sum_{s \in S' \cap x'} \pi_{x'} \lambda_s c(s,x') p(x',s,x) + \sum_{x' \in A_1} \pi_{x'} \mu c(t,x') p(x',t,x) \qquad \forall x \in \overline{A}, \tag{2.1}$$

$$\pi_x \left[\mu c(t,x) + \sum_{s \in S' \cap x} \lambda_s c(s,x) \right] = \sum_{x' \in \Omega} \sum_{s \in S' \cap x'} \pi_{x'} \lambda_s c(s,x') p(x',s,x) + \sum_{x' \in A_1} \pi_{x'} \mu c(t,x') p(x',t,x), \qquad \forall x \in A_1 \tag{2.2}$$

and

$$\pi_x \sum_{s \in S' \cap x} \lambda_s c(s,x) = \sum_{x' \in \Omega} \sum_{s \in S' \cap x'} \pi_{x'} \lambda_s c(s,x') p(x',s,x) + \sum_{x' \in A_1} \pi_{x'} \mu c(t,x') p(x',t,x), \quad \forall x \in A_0. \tag{2.3}$$

Definition 2.2 : The process P has the property of instantaneous attention if t is worked off with positive speed as soon as it is created, that is $p(x,s,x') = 0$ for $x \in \overline{A}, x' \in A_0, s \in S'$ and $p(x,t,x') = 0$ for $x \in A_1, x' \in A_0$.

Definition 2.3 : P is said to be partially balanced with respect to t if, for each $x \in \Omega$, the flux of creating t in state x is equal to the flux at which t dies in state x.

Appropriate partial balance equations may be written as

$$\sum_{x' \in A_1} \pi_{x'} \mu c(t,x') p(x',t,x) + \sum_{x' \in \overline{A}} \sum_{s \in S' \cap x'} \pi_{x'} \lambda_s c(s,x') p(x',s,x) = 0 \quad \forall x \in A_0 \tag{2.4}$$

and

$$\pi_x \mu c(t,x) = \sum_{x' \in A_1} \pi_{x'} \mu c(t,x') p(x',t,x) + \sum_{x' \in \overline{A}} \sum_{s \in S' \cap x'} \pi_{x'} \lambda_s c(s,x') p(x',s,x) \qquad \forall x \in A_1. \tag{2.5}$$

The terms in equation (2.4) are obviously non-negative and hence may only hold when $p(x,t,x') = 0$ for each $x \in A_1, x' \in A_0$ and $p(x,s,x') = 0$ for each $x \in \overline{A}, x' \in A_0$ and $s \in S' \cap x$.

That is, the process P can only be partially balanced if it possesses instantaneous attention. Most authors have assumed this property in their work. In order to keep the process complexity to a minimum, we shall do likewise in Chapter 3. Taylor (1987), however, examines systems not necessarily possessing instantaneous attention and finds necessary and sufficient conditions for the insensitivity of such processes. This will be discussed in more detail in Chapter 4.

Note also that if equation (2.5) holds, then

$$(2.6) \qquad \pi_x \sum_{s \in S' \cap x} \lambda_s c(s,x) = \sum_{x' \in A} \sum_{s \in S' \cap x'} \pi_{x'} \lambda_s c(s,x') p(x',s,x), \qquad \forall x \in A_1.$$

This is shown by subtracting equation (2.5) from equation (2.2).

König and Jansen (1974) proved the following theorem.

Theorem 2.1

Let P be a process possessing instantaneous attention. Then P is insensitive if and only if the partial balance relations (2.5) and (2.6) hold.

Proof : See König and Jansen (1974). ■

2.3 Processes with many generally distributed lifetimes

In this section we consider processes with many generally distributed lifetimes.

The following notation is essentially that of Henderson (1983a). Let P be an irreducible, stationary stochastic process on a set of states $x \in \Omega$. Let S be a set of elements such that $S = S' \cup S^*$, with S' and S^* disjoint sets and $|S^*|$ finite. Incorporated in each state $x \in \Omega$ are active elements from the set S which decay at rate $c(s,x), s \in S$. If $s \in S'$, the lifetime of s has a negative exponential distribution

with mean λ_s^{-1}. If $s \in S^*$, the lifetime of s has a differentiable distribution function $G_s(.)$ with mean μ_s^{-1}, density function $g_s(.)$ and hazard function $h_s(.)$. It will be assumed that when the process moves from x to x' upon the death of s that no two active elements from S^* are activated or die simultaneously and the residual lifetimes of the remaining elements from $x \cap S^*$ remain unchanged. When an active element $s \in S$ dies, the process moves to state $x' \in \Omega$ with probability $p(x, s, x')$. Define

$$\Gamma_s(x) = \{x' | x' \cap S^* = x \cap S^* - \{s\}\},$$

$$\Lambda_{ss'}(x) = \{x' | x' \cap S^* - \{s'\} = x \cap S^* - \{s\}\}, s \neq s',$$

$$U_s(x) = \{x' | x' \cap S^* = (x \cap S^*) \cup \{s\}\},$$

$$\theta(x) = \{x' | x' \cap S^* = x \cap S^*\}.$$

These have the following interpretations.

$\Gamma_s(x)$ is the set of states that have the same active general lifetimes as x except that s has been removed.

$\Lambda_{ss'}(x)$ is the set of states that have the same active general lifetimes as x except that the s lifetime has been replaced by the s' lifetime.

$U_s(x)$ is the set of states with the s lifetime added to those of x.

$\theta(x)$ is the set of states with the same active general lifetimes as x.

Definition 2.4 : The system P is insensitive if the equilibrium distribution for time spent in each state is dependent on arbitrary $G_s(.)$ only through μ_s for every $s \in S^*$.

Definition 2.5 : The system P has the property of instantaneous attention if, for each $s \in S^*$, s is worked off with positive speed as soon as it is created.

For the case when S^* contains more than one element, we shall make the assumption that the processes to be examined possess instantaneous attention.

Let π_x denote the probability of being in state x when $G_s(y) = 1 - \exp(-\mu_s y)$ for all $s \in S^*$, that is, when P is a purely Markov process.

The global balance equations for this purely Markov process are

$$
\begin{aligned}
&\pi_x \left[\sum_{s \in S^* \cap x} \mu_s c(s,x) + \sum_{s \in S' \cap x} \lambda_s c(s,x) \right] \\
&= \sum_{x' \in \theta(x)} \sum_{s \in S^* \cap x} \pi_{x'} \mu_s c(s,x') p(x',s,x) \\
&\quad + \sum_{s \in S^* \cap x} \sum_{s' \in S^* - x} \sum_{x' \in \Lambda_{ss'}(x)} \pi_{x'} \mu_{s'} c(s',x') p(x',s',x) \\
&\quad + \sum_{s \in S^* \cap x} \sum_{x' \in \Gamma_s(x)} \sum_{s' \in S' \cap x'} \pi_{x'} \lambda_{s'} c(s',x') p(x',s',x) \\
&\quad + \sum_{x' \in \theta(x)} \pi_{x'} \sum_{s \in S' \cap x'} \lambda_s c(s,x') p(x',s,x) \\
&\quad + \sum_{s \in S^* - x} \sum_{x' \in U_s(x)} \pi_{x'} \mu_s c(s,x') p(x',s,x).
\end{aligned} \tag{2.7}
$$

As in Section 2.2, we say that P is partially balanced if the flux of creating $s \in S^*$ in state x is equal to the flux at which s dies in state x, for all $x \in \Omega$.

For each $s \in S^* \cap x$, appropriate partial balance equations may be written as

$$
\begin{aligned}
\pi_x \mu_s c(s,x) &= \sum_{x' \in \theta(x)} \pi_{x'} \mu_s c(s,x') p(x',s,x) \\
&\quad + \sum_{x' \in \Gamma_s(x)} \sum_{s' \in S' \cap x'} \pi_{x'} \lambda_{s'} c(s',x') p(x',s',x) \\
&\quad + \sum_{s' \in S^* - x} \sum_{x' \in \Lambda_{ss'}(x)} \pi_{x'} \mu_{s'} c(s',x') p(x',s',x)
\end{aligned} \tag{2.8}
$$

and

$$
\begin{aligned}
\pi_x \sum_{s \in S' \cap x} \lambda_s c(s,x) &= \sum_{x' \in \theta(x)} \pi_{x'} \sum_{s \in S' \cap x'} \lambda_s c(s,x') p(x',s,x) \\
&\quad + \sum_{s \in S^* - x} \sum_{x' \in U_s(x)} \pi_{x'} \mu_s c(s,x') p(x',s,x).
\end{aligned} \tag{2.9}
$$

Let $\mathcal{P}_s$ be the matrix of routing probabilities $p(x,s,x')$ for each $s \in S^*$.

Theorem 2.2 (König and Jansen (1974))

The process P is insensitive if and only if equations (2.8) and (2.9) hold.

Proof : See König and Jansen (1974). ■

A number of results exist which relate the equilibrium distribution using spent lifetime to that using residual lifetime. Let $\pi_x^S(\mathbf{y})$ (respectively $\pi_x^R(\mathbf{y})$) denote the equilibrium density for P being in state $x \in \Omega$ with vector of spent (respectively residual) lifetimes $\mathbf{y}$, for $s \in S^*$.

Theorem 2.3

The following statements are equivalent.

(1) P is insensitive.

(2) $\pi_x^S(\mathbf{y}) = \pi_x^R(\mathbf{y})$ for all $x \in \Omega$.

(3) $\pi_x^S(\mathbf{y}) = \pi_x \prod_{s \in S^* \cap x} \mu_s(1 - G_s(y_s))$, where $\{\pi_x, x \in \Omega\}$ is the equilibrium distribution of the purely Markov process.

Proof : See Henderson (1983a). ■

CHAPTER 3 : PROCESSES WITH AGE DEPENDENT ROUTING

3.1 Introduction

Many systems have the characteristic that they are insensitive, that is, the equilibrium distribution is dependent upon a set of governing distributions only through their means. As discussed in Chapter 1, numerous authors have found necessary and sufficient conditions for insensitivity of Generalised Semi-Markov Processes (GSMPs) and networks of queues. Most of those authors have shown that GSMPs with instantaneous attention are insensitive if and only if they exhibit the property of partial balance, as defined in Chapter 2. Henderson (1983a) found alternative necessary and sufficient conditions for insensitivity by examining the processes in reverse time.

All of these schemes involved the assumption that the probabilities $p(x, s, x')$ for moving from a state x to some other state x' through the death of lifetime s are independent of the length of the lifetimes.

Henderson and Taylor (1987) and Taylor (1987) made the first forays into relaxing this assumption by studying a class of *interruption* processes in which generally distributed lifetimes may be interrupted by events which occur in Poisson streams. The routing probabilities are dependent on the nature of the death (that is, "natural" or interrupted) and hence are dependent on the age of the lifetime.

In this chapter their investigation is extended to the consideration of processes where the routing probabilities, upon the death of a generally distributed lifetime, are functions of the age of the lifetime. It is shown that the equilibrium distribution in insensitive processes without age dependent routing is the same as that of some age dependent processes. As a result, the theory of insensitivity is not only generalised, but wide classes of processes are found for which the supplemented equilibrium distribution has a simple and elegant form.

To illustrate the ideas introduced in this chapter, we deal firstly with processes having only one generally distributed lifetime, and then move to processes with many generally distributed elements.

In Section 3.2, necessary and sufficient conditions are derived for classes of processes with one generally distributed lifetime and age dependent routing to have the same equilibrium distribution. In Section 3.3 the analysis is extended to processes with many generally distributed lifetimes while networks of queues, with examples, are studied in Section 3.4. Applications of the theory to interruption processes and Markov renewal processes are examined in Sections 3.5 and 3.6, and Section 3.7 provides some insight as to why the techniques used in this chapter actually work. In Section 3.8 a theorem of Barbour and Schassberger (1981) is extended to include age dependent processes while Section 3.9 considers the extension to age dependent speeds.

3.2 Age dependent routing in processes with one general lifetime

Let P be the process with one generally distributed lifetime defined in Section 2.2 and assume that it has instantaneous attention.

In the description of P it is assumed that all of the routing probabilities are constant. We shall now relax this assumption and allow some of these to be functions of y. In particular, let $p(x,t,x',y)$ be the routing probability of moving from state x to x' through the death of t at age y, and define $\mathcal{P}(y)$ to be the matrix of these routing probabilities for all $y > 0$. We will call the process with these age dependent probabilities $\overline{P}$.

Let y be a supplementary variable for the spent service time of the generally distributed lifetime. We denote the probability density that the process $\overline{P}$ is in states x and (x,y) by $\overline{\pi}_x$ and $\overline{\pi}_x(y)$ respectively.

The global supplemented balance equations of the process $\overline{P}$ are

(3.1)
$$\overline{\pi}_x \sum_{s\in S'\cap x} \lambda_s c(s,x) = \sum_{x'\in\overline{A}} \sum_{s\in S'\cap x'} \overline{\pi}_{x'} \lambda_s c(s,x') p(x',s,x) + \sum_{x'\in A_1} \int_0^\infty \overline{\pi}_{x'}(y) h(y) c(t,x') p(x',t,x,y)\,dy \qquad \forall x \in \overline{A},$$

$$\overline{\pi}_x(y)\sum_{s\in S'\cap x}\lambda_s c(s,x)+\left[h(y)\overline{\pi}_x(y)+\frac{d}{dy}\overline{\pi}_x(y)\right]c(t,x)$$
$$=\sum_{x'\in A}\sum_{s\in S'\cap x'}\overline{\pi}_{x'}(y)\lambda_s c(s,x')p(x',s,x) \qquad \forall x\in A_1, \tag{3.2}$$

$$\overline{\pi}_x(y)\sum_{s\in S'\cap x}\lambda_s c(s,x)=\sum_{x'\in A}\sum_{s\in S'\cap x'}\overline{\pi}_{x'}(y)\lambda_s c(s,x')p(x',s,x) \qquad \forall x\in A_0 \tag{3.3}$$

and

$$\overline{\pi}_x(0)c(t,x)=\sum_{x'\in\overline{A}}\sum_{s\in S'\cap x}\overline{\pi}_{x'}\lambda_s c(s,x')p(x',s,x)$$
$$+\sum_{x'\in A_1}\int_0^\infty\overline{\pi}_{x'}(y)h(y)c(t,x')p(x',t,x,y)dy \qquad \forall x\in A_1. \tag{3.4}$$

For a given $\overline{P}$ process choose $\mathcal{P}=\int_0^\infty\mathcal{P}(y)dG(y)$ and hence construct an associated P process with routing matrix $\mathcal{P}$. Label this process the $Q(\overline{P})$ process, which for the sake of simplicity will be called the Q process. Note that the irreducibility of Q follows from $\overline{P}$ being irreducible.

It may well appear that the equilibrium distribution of Q and $\overline{P}$ are identical as one is simply the "average" of the other. If true, this would be of great benefit as the Q process is obviously easier to analyse than $\overline{P}$. However, the equilibrium distribution of age dependent systems is not necessarily the same as that of the associated average process and hence it is interesting to find conditions under which both distributions are the same. A simple example is given to illustrate that differences occur.

Example 3.1

Consider the system illustrated in Figure 3.1. Let $\Omega=\{1,2,3,4\}$, $A=\{1,2\}$ and $\overline{A}=\{3,4\}$. Suppose that the length of time spent in A is negative exponentially distributed, mean 0.25. If the element t dies in state 1 the process moves to state 3. On the other hand, if t dies in state 2, after spending a period of time y in A, the process moves to state 3 with probability $\exp(-\gamma y)$, $(\gamma>0)$, and state 4 with

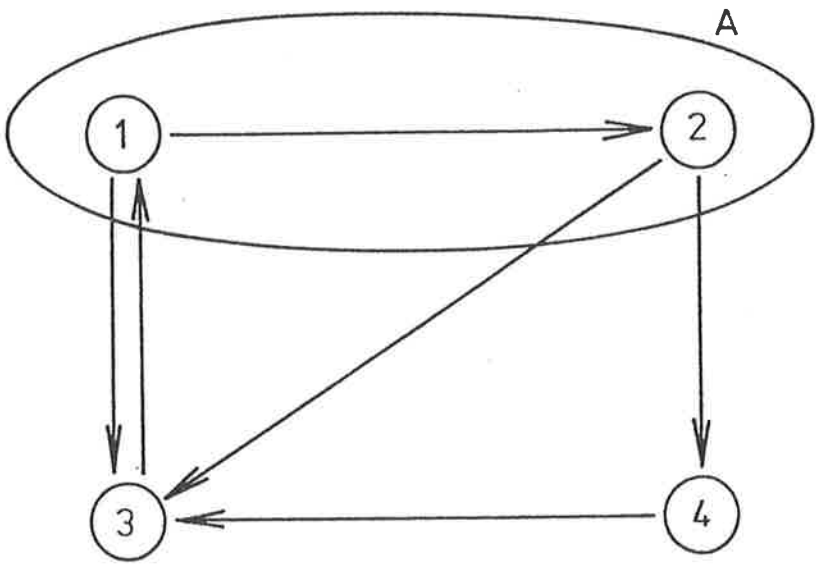


Figure 3.1. State transition diagram for process described in Example 3.1.

probability $1-\exp(-\gamma y)$. The process spends a negative exponentially distributed period of time, with unit mean, in state 3 (respectively 4) and then moves to state 1 (respectively 3). While in state 1 an internal transition may occur (at unit rate) and the process moves to state 2.

The balance equations for this process are

$$\frac{d}{dy}\pi_1(y) + 5\pi_1(y) = 0,$$

$$\frac{d}{dy}\pi_2(y) + 4\pi_2(y) = \pi_1(y),$$

$$\pi_1(0) = \pi_3$$

and

$$\pi_4 = \int_0^\infty 4\pi_2(y)(1 - \exp(-\gamma y))dy.$$

These equations have solution

$$\pi_1(y) = B_1 \exp(-5y) \quad \Rightarrow \quad \pi_1 = \frac{B_1}{5},$$

$$\pi_2(y) = B_1(\exp(-4y) - \exp(-5y)) \quad \Rightarrow \quad \pi_2 = \frac{B_1}{20},$$

$$\pi_3 = B_1,$$

and

$$\pi_4 = 4B_1\left[\frac{1}{20} - \frac{1}{4+\gamma} + \frac{1}{5+\gamma}\right],$$

where B_1 is a normalising constant.

The routing probabilities for the associated Q process are

$$p(2,3) = \int_0^\infty 4\exp(-\gamma y)\exp(-4y)dy = \frac{4}{4+\gamma}$$

and

$$p(2,4) = 1 - p(2,3) = \frac{\gamma}{4+\gamma}$$

The balance equations of Q are

$$\frac{d}{dy}\pi_1(y) + 5\pi_1(y) = 0,$$

$$\frac{d}{dy}\pi_2(y) + 4\pi_2(y) = \pi_1(y),$$

$$\pi_1(0) = \pi_3$$

and

$$\pi_4 = \int_0^\infty 4\pi_2(y)dy\frac{\gamma}{4+\gamma}.$$

The solution to this system of equations is

$$\pi_1(y) = B_2\exp(-5y) \quad \Rightarrow \quad \pi_1 = \frac{B_2}{5},$$

$$\pi_2(y) = B_2(\exp(-4y) - \exp(-5y)) \quad \Rightarrow \quad \pi_2 = \frac{B_2}{20},$$

$$\pi_3 = B_2,$$

and

$$\pi_4 = B_2\frac{\gamma}{5(4+\gamma)},$$

where B_2 is a normalising constant.

The two distributions are identical if $B_1 = B_2$. Elementary algebraic manipulation shows that this only occurs when $\gamma = 0$ or $\gamma = \infty$, both cases corresponding to fixed routing probabilities. Hence the distribution of the age dependent process and the associated Q process are not generally the same. □

Φ-insensitivity and the equilibrium distribution

Assume henceforth that π_x refers to the equilibrium distribution of the Q process whenever this distribution exists.

Let Φ be any set of pairs $(\mathcal{P}(.), G(.))$ with the properties

(1) $\int_0^\infty \mathcal{P}(y)dG(y) = \mathcal{P}$ for fixed $\mathcal{P}$.

(2) The mean of $G(.)$ is μ^{-1} for fixed μ^{-1}.

(3) $(\mathcal{P}, G(.)) \in \Phi$ for all distribution functions $G(.)$ satisfying (2).

Definition 3.1 : The process $\overline{P}$ is Φ-insensitive if for all $(\mathcal{P}(.), G(.)) \in \Phi$, the equilibrium distribution is invariant.

Theorem 3.1

The following statements are equivalent

(a) $\overline{P}$ is Φ-insensitive.

(b) The associated Q process is insensitive.

(c) The equilibrium distribution of $\overline{P}$ is

$$\overline{\pi}_x = \pi_x, \quad \forall x \in \overline{A}, \tag{3.5}$$

$$\overline{\pi}_x(y) = \pi_x \mu(1 - G(y)), \quad \forall x \in A. \tag{3.6}$$

Proof : We show (a)⇒(b)⇒(c)⇒(a).

(a)⇒(b) : If $\overline{P}$ is Φ-insensitive the particular choice $\mathcal{P}(y) = \mathcal{P}$ for all $y > 0$ produces a $\overline{P}$ process identical to its associated Q process. With this choice the pair $(\mathcal{P}, G(.))$ is in Φ for any $G(.)$ with mean μ^{-1} and, since $\overline{P}$ is Φ-insensitive, the equilibrium distribution is the same for each $G(.)$. Hence Q is insensitive.

(b)⇒(c) : If Q is insensitive then, by Theorem 2.1, equations (2.5) and (2.6) are satisfied. Substituting from equations (3.5) and (3.6) into equation (3.1) gives, for

$x \in \overline{A}$,

$$
\begin{aligned}
\pi_x \sum_{s \in S' \cap x} \lambda_s c(s,x) &= \sum_{x' \in \overline{A}} \sum_{s \in S' \cap x'} \pi_{x'} \lambda_s c(s,x') p(x', s, x) \\
&\quad + \sum_{x' \in A_1} \int_0^\infty \pi_{x'} \mu (1 - G(y)) h(y) c(t, x') p(x', t, x, y) dy \\
&= \sum_{x' \in \overline{A}} \sum_{s \in S' \cap x'} \pi_{x'} \lambda_s c(s,x') p(x', s, x) \\
&\quad + \sum_{x' \in A_1} \pi_{x'} \mu \int_0^\infty c(t, x') p(x', t, x, y) dG(y) \\
&= \sum_{x' \in \overline{A}} \sum_{s \in S' \cap x'} \pi_{x'} \lambda_s c(s,x') p(x', s, x) \\
&\quad + \sum_{x' \in A_1} \pi_{x'} \mu c(t, x') p(x', t, x).
\end{aligned} \tag{3.7}
$$

This is equation (2.1).

Substituting equation (3.6) into equation (3.2) gives, for $x \in A_1$,

$$
\begin{aligned}
\pi_x \mu (1 - G(y)) \sum_{s \in S' \cap x} \lambda_s c(s,x) &+ \left[h(y) \pi_x \mu (1 - G(y)) - \pi_x \mu g(y) \right] c(t,x) \\
&= \sum_{x' \in A} \sum_{s \in S' \cap x'} \pi_{x'} \mu (1 - G(y)) \lambda_s c(s,x') p(x', s, x).
\end{aligned} \tag{3.8}
$$

Using $h(y)(1-G(y)) = g(y)$ and dividing both sides by $\mu(1-G(y))$ gives equation (2.6).

Substituting into equation (3.3) and dividing both sides by $\mu(1 - G(y))$ gives (2.3) for $x \in A_0$.

Finally, substituting into equation (3.4) gives

$$
\begin{aligned}
\pi_x \mu c(t,x) &= \sum_{x' \in \overline{A}} \sum_{s \in S' \cap x} \pi_{x'} \lambda_s c(s,x') p(x', s, x) \\
&\quad + \sum_{x' \in A_1} \pi_{x'} \mu \int_0^\infty (1 - G(y)) h(y) c(t, x') p(x', t, x, y) dy \\
&= \sum_{x' \in \overline{A}} \sum_{s \in S' \cap x} \pi_{x'} \lambda_s c(s,x') p(x', s, x) \\
&\quad + \sum_{x' \in A_1} \pi_{x'} \mu c(t, x') p(x', t, x) \qquad \forall x \in A_1,
\end{aligned} \tag{3.9}
$$

which is equation (2.5). Hence equations (3.5) and (3.6) give the equilibrium distribution of $\overline{P}$.

(c)⇒(a) : $\overline{P}$ has equilibrium density given by equations (3.5) and (3.6) with $\{\pi_x, x \in \Omega\}$ the equilibrium distribution of the associated Q process. The $\{\pi_x, x \in \Omega\}$ depends upon $G(.)$ only through its mean and the routing probabilities of Q, that is, through μ and $\int_0^\infty \mathcal{P}(y)dG(y)$. Keeping this value constant keeps $\{\pi_x, x \in \Omega\}$ constant and restricts consideration of the general distributions to the set Φ. Hence $\overline{P}$ is Φ-insensitive. ■

Example 3.2

Consider a process with one generally distributed lifetime and age dependent routing probabilities taking the form

$$p(x,s,x',y) = \begin{cases} p_1(x,s,x') & \text{for } 0 < y \leq t_1 \\ p_2(x,s,x') & \text{for } t_1 < y \leq t_2 \\ \vdots & \\ p_n(x,s,x') & \text{for } t_{n-1} < y \leq t_n \\ p_{n+1}(x,s,x') & \text{for } y > t_n. \end{cases}$$

The average routing probabilities are

$$p(x,s,x') = G_s(t_1)p_1(x,s,x') + (G_s(t_2) - G_s(t_1))p_2(x,s,x') + \ldots + (1 - G_s(t_n))p_{n+1}(x,s,x').$$

For fixed $p_i(x,s,x'), i = 1,\ldots,n+1$, $p(x,s,x')$ can be kept at a constant value by selecting $G_s(.)$ from the set

$$\{H(.)|H(t_1) = a_1, H(t_2) = a_2, \ldots, H(t_n) = a_n, \text{ and } \int_0^\infty (1 - H(y))dy = \mu^{-1}\}$$

where a_i, for $i = 1,\ldots,n$, and μ are fixed values. This process then has an equilibrium distribution in the form of equations (3.5) and (3.6) for such distribution functions if equations (2.5) and (2.6) hold for the associated Q process. □

3.3 Age dependent processes with many generally distributed lifetimes

In the previous section we illustrated the concept of Φ-insensitivity in processes with only one generally distributed lifetime. We now extend that analysis to processes with many generally distributed lifetimes and introduce some new results concerning the equilibrium densities of insensitive systems.

Let P be the process described in Section 2.3.

We now create the process $\overline{P}$ by modifying P so as to allow age dependent routing probabilities. For each $s \in S^*$, let $p(x, s, x', y_s)$ be the probability of moving from state x to x' through the death of s when its age was y_s and denote the probability density that $\overline{P}$ is in state x with vector of spent lifetimes $\mathbf{y}$ and residual lifetimes $\mathbf{z}$, for elements in $x \cap S^*$ by $\overline{\pi}_x(\mathbf{y}, \mathbf{z})$. Using the notation introduced in Section 2.3, the supplemented global balance equations of $\overline{P}$ may be written as : For each $s \in x \cap S^*$,

(3.10)

$$\begin{aligned}
&\overline{\pi}_x(\mathbf{y},0_s,\mathbf{z},z_s)c(s,x) \\
&\quad = \sum_{x' \in \Gamma_s(x)} \overline{\pi}_{x'}(\mathbf{y},\mathbf{z}) \sum_{s' \in S' \cap x'} \lambda_{s'} c(s',x') p(x',s',x) g_s(z_s) \\
&\qquad + \sum_{s' \in S^* - x} \sum_{x' \in \Lambda_{ss'}(x)} \int_0^\infty \overline{\pi}_{x'}(\mathbf{y}, y_{s'}, \mathbf{z}, 0_{s'}) c(s',x') p(x',s',x,y_{s'}) dy_{s'} g_s(z_s) \\
&\qquad + \sum_{x' \in \theta(x)} \int_0^\infty \overline{\pi}_{x'}(\mathbf{y}, y_s, \mathbf{z}, 0_s) c(s,x') p(x',s',x,y_s) dy_s g_s(z_s)
\end{aligned}$$

and

(3.11)

$$\begin{aligned}
&\sum_{s \in S^* \cap x} c(s,x) \left(\frac{\partial}{\partial y_s} - \frac{\partial}{\partial z_s} \right) \overline{\pi}_x(\mathbf{y},\mathbf{z}) + \overline{\pi}_x(\mathbf{y},\mathbf{z}) \sum_{s \in S' \cap x} \lambda_s c(s,x) \\
&\quad = \sum_{s \in S^* - x} \sum_{x' \in U_s(x)} \int_0^\infty \overline{\pi}_{x'}(\mathbf{y}, y_s, \mathbf{z}, 0_s) c(s,x') p(x',s,x,y_s) dy_s \\
&\qquad + \sum_{x' \in \theta(x)} \overline{\pi}_{x'}(\mathbf{y},\mathbf{z}) \sum_{s \in S' \cap x'} \lambda_s c(s,x') p(x',s,x),
\end{aligned}$$

where $(\mathbf{y}, 0_s, \mathbf{z}, z_s)$ is a vector of lifetimes with the additional lifetime $s \in S^*$ having just been created and given residual lifetime z_s and $(\mathbf{y}, y_s, \mathbf{z}, 0_s)$ is a vector of lifetimes

with the additional lifetime $s \in S^*$ about to die, having received an amount of service y_s.

Define $\mathcal{P}_s(y)$ to be the matrix of age dependent routing probabilities $p(x, s, x', y)$ for all $y > 0$ and $s \in S^*$.

As in Section 3.2, for a given $\overline{P}$ process choose $\mathcal{P}_s = \int_0^\infty \mathcal{P}_s(y) dG_s(y)$ and hence construct an associated P process with routing matrices $\mathcal{P}_s$. Label this P process the $Q(\overline{P})$ process, which, again for the sake of simplicity, will be called the Q process.

Assume henceforth that π_x refers to the equilibrium distribution of the Q process whenever this distribution exists.

Define, for each $s \in S^*$, Φ_s to be any set of pairs $(\mathcal{P}_s(.), G_s(.))$ with the properties

(1) $\int_0^\infty \mathcal{P}_s(y) dG_s(y) = \mathcal{P}_s$ for fixed $\mathcal{P}_s$.

(2) The mean of $G_s(.)$ is μ_s^{-1} for fixed μ_s^{-1}.

(3) $(\mathcal{P}_s, G_s(.)) \in \Phi_s$ for all distribution functions $G_s(.)$ satisfying (2).

Definition 3.2 : The process $\overline{P}$ is Φ-insensitive if for all $(\mathcal{P}_s(.), G_s(.)) \in \Phi_s$, $s \in S^*$, the equilibrium distribution is invariant.

Theorem 3.2

The following statements are equivalent

(a) $\overline{P}$ is Φ-insensitive.

(b) The associated Q process is insensitive.

(c) The equilibrium density of $\overline{P}$ is

$$\overline{\pi}_x(\mathbf{y}, \mathbf{z}) = \pi_x \prod_{s \in S^* \cap x} \mu_s g_s(y_s + z_s). \tag{3.12}$$

Proof : We show (a)⇒(b)⇒(c)⇒(a).

(a)⇒(b) : If $\overline{P}$ is Φ-insensitive the particular choice $\mathcal{P}_s(y) = \mathcal{P}_s$ for all $y > 0$ and $s \in S^*$ produces a $\overline{P}$ process identical to its associated Q process. With this choice the pair $(\mathcal{P}_s, G_s(.))$ is in Φ_s for any $G_s(.)$ with mean μ_s^{-1}. Hence Q is insensitive.

(b)⇒(c) : If Q is insensitive then Theorem 2.2 implies that equations (2.8) and (2.9) are satisfied. Substituting equation (3.12) into equation (3.10) we get

$$
\begin{aligned}
&\pi_x c(s,x)\mu_s g_s(z_s) \prod_{s'\in S^*\cap x-\{s\}} \mu_{s'} g_{s'}(y_{s'}+z_{s'}) \\
&= \sum_{x'\in\Gamma_s(x)} \pi_{x'} \prod_{r\in S^*\cap x'} \mu_r g_r(y_r+z_r) \sum_{s'\in S'\cap x'} \lambda_{s'} c(s',x') p(x',s',x) g_s(z_s) \\
&+ \sum_{s'\in S^*-x} \sum_{x'\in\Lambda_{ss'}(x)} \pi_{x'} \prod_{r\in S^*\cap x'-\{s'\}} \mu_r g_r(y_r+z_r) \\
&\qquad \cdot \int_0^\infty \mu_{s'} g_{s'}(y_{s'}) c(s',x') p(x',s',x,y_{s'}) dy_{s'}\, g_s(z_s) \\
&+ \sum_{x'\in\theta(x)} \pi_{x'} \prod_{r\in S^*\cap x'-\{s\}} \mu_r g_r(y_r+z_r) \\
&\qquad \cdot \int_0^\infty \mu_s g_s(y_s) c(s,x') p(x',s,x,y_s) dy_s\, g_s(z_s).
\end{aligned} \tag{3.13}
$$

Cancelling $g_s(z_s)\prod_{s'\in S^*\cap x-\{s\}} \mu_{s'} g_{s'}(y_{s'}+z_{s'})$ from both sides and recalling that $\int_0^\infty p(x',s,x,y_s)dG_s(y_s) = p(x',s,x)$ gives equation (2.8), which holds as Q is insensitive.

Substituting into equation (3.11) gives

$$
\begin{aligned}
&\pi_x \prod_{r\in S^*\cap x} \mu_r g_r(y_r+z_r) \sum_{s\in S'\cap x} \lambda_s c(s,x) \\
&= \sum_{s\in S^*-x} \sum_{x'\in U_s(x)} \pi_{x'} \prod_{s'\in S^*\cap x'-\{s\}} \mu_{s'} g_{s'}(y_{s'}+z_{s'}) \int_0^\infty \mu_s g_s(y_s) c(s,x') p(x',s,x,y_s) dy_s \\
&\quad + \sum_{x'\in\theta(x)} \pi_{x'} \prod_{s'\in S^*\cap x} \mu_{s'} g_{s'}(y_{s'}+z_{s'}) \sum_{s\in S'\cap x'} \lambda_s c(s,x') p(x',s,x).
\end{aligned} \tag{3.14}
$$

Cancelling $\prod_{s'\in S^*\cap x} \mu_{s'} g_{s'}(y_{s'}+z_{s'})$ from both sides gives equation (2.9), which also holds due to the insensitivity of Q. Hence equation (3.12) is the equilibrium density of $\overline{P}$.

(c)⇒(a) : $\overline{P}$ has equilibrium density given by equation (3.12) with $\{\pi_x, x\in\Omega\}$ the equilibrium distribution of the associated Q process. The $\{\pi_x, x\in\Omega\}$ depends upon $\{G_s(.), s\in S^*\}$ only through the means of these distributions and the routing

probabilities of Q, that is, through μ_s and $\int_0^\infty \mathcal{P}_s(y)dG_s(y)$ for all $s \in S^*$. Keeping these values constant keeps $\{\pi_x, x \in \Omega\}$ constant and restricts consideration of the general distributions to the sets Φ_s for all $s \in S^*$. Hence $\overline{P}$ is Φ-insensitive. ■

Discussion of Theorem 3.2

It is important to realise that Φ-insensitivity requires the standard insensitivity of König and Jansen (1974) to be a special case. That is, when the routing probabilities are independent of the age of the dying lifetime, the distributions $G_s(.)$, for each $s \in S^*$, may be varied arbitrarily (provided the mean is fixed) without affecting the equilibrium distribution.

In general, however, Φ-insensitivity is not of this form because, for a given age dependent routing matrix $\mathcal{P}_s(.)$, $G_s(.)$ may *not* be chosen arbitrarily. Consider a particular process $\overline{P}$ with routing probabilities $\mathcal{P}_s(.)$ (not constant) and lifetime distributions $G_s(.)$ specified for each $s \in S^*$. An associated Q process may be constructed for this system in the previously described fashion. Suppose now that new lifetime distributions are chosen (with the same means as before). The routing probabilities of the Q process associated with this system are not necessarily the same as those of the original Q process and hence the equilibrium distributions may differ. So for Φ-insensitivity we are looking at a restricted set of distribution functions associated with particular age dependent routing probabilities; we are *not* looking at all distribution functions with the same mean.

A consequence of Theorem 3.2 is that given an insensitive process, P, any age dependent process which has P as its associated Q process will have the same supplemented equilibrium distribution. We can therefore find many age dependent processes whose physical behaviour are quite different yet retain the same equilibrium distribution because they have the same mean routing probabilities. On the other hand, a system with age dependent features could be found to have a solution of the form (3.12) by checking the associated Q process for insensitivity. This procedure

is used in Section 3.4 to find the equilibrium distribution for classes of networks of queues with age dependent routing.

Note also that a further consequence of Theorem 3.2 is that $\overline{P}$ is Φ-insensitive if and only if the partial balance equations (2.8) and (2.9) of the associated Q process are satisfied.

Corollary 3.3

Let $\hat{\pi}_x(\mathbf{y},\mathbf{z})$ be the probability that $\overline{P}$ is in state $x \in \Omega$ and that, for $s \in S^* \cap x$, the spent (respectively residual) lifetime of s is greater than y_s (respectively z_s).

If $\overline{P}$ is Φ-insensitive then

$$\hat{\pi}(\mathbf{y},\mathbf{z}) = \pi_x \prod_{s \in S^* \cap x} \mu_s \int_{y_s+z_s}^{\infty} (1 - G_s(u_s))du_s. \tag{3.15}$$

Proof : By definition,

$$\hat{\pi}_x(\mathbf{y},\mathbf{z}) = \int_{\mathbf{y}}^{\infty} \int_{\mathbf{z}}^{\infty} \overline{\pi}_x(\mathbf{u},\mathbf{v})d\mathbf{v}d\mathbf{u}. \tag{3.16}$$

Substituting for $\overline{\pi}_x(\mathbf{u},\mathbf{v})$ gives

$$\begin{aligned} \hat{\pi}_x(\mathbf{y},\mathbf{z}) &= \int_{\mathbf{y}}^{\infty} \int_{\mathbf{z}}^{\infty} \left(\pi_x \prod_{s \in S^* \cap x} \mu_s g_s(u_s + v_s) \right) d\mathbf{v}d\mathbf{u} \\ &= \pi_x \prod_{s \in S^* \cap x} \mu_s \int_{y_s}^{\infty} (1 - G_s(u_s + z_s))du_s \\ &= \pi_x \prod_{s \in S^* \cap x} \mu_s \int_{y_s+z_s}^{\infty} (1 - G_s(u_s))du_s, \end{aligned} \tag{3.17}$$

as required. ■

Remember that insensitivity in the usual sense is a special case of Φ-insensitivity and therefore the above result also holds for all insensitive processes. By putting $\mathbf{z}$ equal to the zero-vector, the integrated version of the product form given in Theorem 2.3 is derived.

Corollary 3.4

Let $\overline{\pi}_x^S(\mathbf{y})$ (respectively $\overline{\pi}_x^R(\mathbf{y})$) denote the equilibrium density for $\overline{P}$ being in state $x \in \Omega$ with vector of spent (respectively residual) lifetimes $\mathbf{y}$, for $s \in S^*$. If $\overline{P}$ is Φ-insensitive then

$$\overline{\pi}_x^R(\mathbf{y}) = \overline{\pi}_x^S(\mathbf{y}) = \pi_x \prod_{s \in S^* \cap x} \mu_s(1 - G_s(y_s)). \tag{3.18}$$

Proof : By definition,

$$\overline{\pi}_x^S(\mathbf{y}) = \int_0^\infty \overline{\pi}_x(\mathbf{y}, \mathbf{v}) d\mathbf{v}. \tag{3.19}$$

Substituting for $\pi_x(\mathbf{y}, \mathbf{v})$ and integrating gives the desired form of $\overline{\pi}_x(\mathbf{y})$. Using the symmetry of $\overline{\pi}_x(\mathbf{y}, \mathbf{z})$ completes the proof. ■

Note that this is a generalised form of the result given in Theorem 2.3. We are now saying that a product form equilibrium density exists not only for the process with fixed routing probabilities, but also for the age dependent process.

Theorem 3.5

If a process $\overline{P}$ has the property that

$$\overline{\pi}_x^R(\mathbf{y}) = \overline{\pi}_x^S(\mathbf{y}), \tag{3.20}$$

then $\overline{P}$ is Φ-insensitive.

Proof : It suffices to consider the case where there is only one element, t say, in S^*. Without loss of generality let $\overline{\pi}_x^R(y) = \overline{\pi}_x^S(y) = \overline{\pi}_x(y)$. Using the notation of Section 3.2, the supplemented global balance equations employing both spent and residual lifetimes may be written as

$$\begin{aligned} \overline{\pi}_x \sum_{s \in S' \cap x} \lambda_s c(s, x) = & \sum_{x' \in \overline{A}} \sum_{s \in S' \cap x'} \overline{\pi}_{x'} \lambda_s c(s, x') p(x', s, x) \\ & + \sum_{x' \in A_1} \int_0^\infty \overline{\pi}_{x'}(y, 0) c(t, x') p(x', t, x, y) dy, \qquad \forall x \in \overline{A}, \end{aligned} \tag{3.21}$$

(3.22)

$$\overline{\pi}_x(y,z)\sum_{s\in S'\cap x}\lambda_s c(s,x)+\left(\frac{\partial}{\partial y}-\frac{\partial}{\partial z}\right)\overline{\pi}_x(y,z)c(t,x)$$
$$=\sum_{x'\in A}\sum_{s\in S'\cap x'}\overline{\pi}_{x'}(y,z)\lambda_s c(s,x')p(x',s,x),\qquad \forall x\in A_1,$$

(3.23)

$$\overline{\pi}_x(y,z)\sum_{s\in S'\cap x}\lambda_s c(s,x)=\sum_{x'\in A}\sum_{s\in S'\cap x'}\overline{\pi}_{x'}(y,z)\lambda_s c(s,x')p(x',s,x),\qquad \forall x\in A_0$$

and

(3.24)

$$\overline{\pi}_x(0,z)c(t,x)=\sum_{x'\in\overline{A}}\sum_{s\in S'\cap x}\overline{\pi}_{x'}\lambda_s c(s,x')p(x',s,x)g(z)$$
$$+\sum_{x'\in A_1}\int_0^\infty\overline{\pi}_{x'}(y,0)c(t,x')p(x',t,x,y)dyg(z),\qquad \forall x\in A_1.$$

Integrating equation (3.22) with respect to y (with an appropriate change of notation) gives

(3.25)

$$\overline{\pi}_x^R(y)\sum_{s\in S'\cap x}\lambda_s c(s,x)-\overline{\pi}_x(0,y)c(t,x)-\frac{d}{dy}\overline{\pi}_x^R(y)c(t,x)$$
$$=\sum_{x'\in A}\sum_{s\in S'\cap x'}\overline{\pi}_x^R x'(y)\lambda_s c(s,x')p(x',s,x),\qquad \forall x\in A_1.$$

Similarly, integrating with respect to z gives

(3.26)

$$\overline{\pi}_x^S(y)\sum_{s\in S'\cap x}\lambda_s c(s,x)+\overline{\pi}_x(y,0)c(t,x)+\frac{d}{dy}\overline{\pi}_x^S(y)c(t,x)$$
$$=\sum_{x'\in A}\sum_{s\in S'\cap x'}\overline{\pi}_{x'}^S(y)\lambda_s c(s,x')p(x',s,x),\qquad \forall x\in A_1.$$

The term $\overline{\pi}_x(y,0)$ is the density that the lifetime has been alive for time y and is about to die. However, this is equal to $\overline{\pi}_x^S(y)h(y)$. Conversely, $\overline{\pi}_x(0,y)$ is the equilibrium density that the lifetime has just been created and given residual lifetime y. This may alternatively be expressed as $\pi_x^R(0)g(y)$.

Using these observations, together with relationship (3.20) and subtracting equation (3.25) from equation (3.26) yields

(3.27)

$$2\frac{d}{dy}\overline{\pi}_x(y)=-\overline{\pi}_x(0)g(y)-\overline{\pi}_x(y)h(y).$$

Equation (3.27) has solution given by $\overline{\pi}_x(y) = \overline{\pi}_x(0)(1 - G(y))$, and hence $\overline{\pi}_x(y) = \pi_x\mu(1 - G(y))$. This implies that $\overline{P}$ is Φ-insensitive. Note also that equations (3.23) and (3.24) are also satisfied by this solution. ■

Together, Corollary 3.4 and Theorem 3.5 provide an extension of Theorem 2.3 to processes with age dependent routing. The ideas developed in this section now give an approach for the analysis of systems with age dependent routing, namely by first considering the insensitivity of the associated Q process. This approach will be exploited in Sections 3.4, 3.5 and 3.6.

3.4 Networks of queues with age dependent routing

Consider a network comprising of a set of labelled queues $\mathcal{N} = \{1, 2, \ldots, N\}$, with customers labelled by type from the set $\mathcal{T} = \{1, 2, \ldots, T\}$.

The state of the network can be represented by a vector $\mathbf{c}$ which gives the type of customer in each position of each queue. Let $n(i,t)$ be the number of type t customers in queue i. Define $\mathbf{n} = (n(i), i = 1, \ldots, N)$, where $n(i) = \sum_{t=1}^{T} n(i,t)$, to be the "macrostate" giving the total number of customers at each node, irrespective of type and $\mathbf{e}_i$ to be an N-vector with a one in the i^{th} position and zeroes elsewhere.

Assume that service facility i works at a rate $\mu_i(\mathbf{n}) = \phi(\mathbf{n}-\mathbf{e}_i)/\phi(\mathbf{n})$ for arbitrary $\phi(.)$ and a proportion $\gamma_i(l, \mathbf{n})$ of this effort is dedicated to the customer in position l when the macrostate is $\mathbf{n}$.

Customers of type t arrive at queue i from outside the network in a Poisson stream of rate $\lambda(i,t)$. If, after the arrival of a customer to queue i the macrostate is $\mathbf{n}$, the customer moves into position l with probability $\delta_i(l, \mathbf{n})$. Note that

$$\sum_{l=1}^{n(i)} \gamma_i(l, \mathbf{n}) = \sum_{l=1}^{n(i)} \delta_i(l, \mathbf{n}) = 1. \tag{3.28}$$

When a customer leaves position l in queue i the customers in positions $l + 1, \ldots, n(i)$ move to positions $l, \ldots, n(i) - 1$ respectively. Similarly, when a customer

moves into position l the customers in positions $l, \ldots, n(i)-1$ move into positions $l+1, \ldots, n(i)$.

Definition 3.3 : Queue i is symmetric if and only if $\gamma_i(l, \mathbf{n}) = \delta_i(l, \mathbf{n})$ for all $l, \mathbf{n}$.

Let the set of symmetric queues be $\mathcal{I} \subseteq \mathcal{N}$.

If $i \in \mathcal{I}$ then type t customers at queue i have service time drawn from the general distribution $G_{it}(.)$ with mean $M(i,t)$. If $i \in \mathcal{N} - \mathcal{I}$, the service time is negative exponentially distributed with mean $M(i)$.

A type t customer, on leaving queue i, for $i \in \mathcal{N} - \mathcal{I}$, changes to a type s customer and moves to queue j with probability $p(i,t;j,s)$ or leaves the network with probability $r(i,t)$. When $i \in \mathcal{I}$ and the customer has received an amount of service y, the corresponding routing probabilities are given by $p(i,t,y;j,s)$ and $r(i,t,y)$ respectively.

Theorem 3.6

A network of the type described above has equilibrium distribution

$$\pi(\mathbf{c}) = C\phi(\mathbf{n}) \prod_{t=1}^{T} \left[\prod_{i \in \mathcal{N}-\mathcal{I}} [y(i,t)M(i)]^{n(i,t)} \right] \left[\prod_{k \in \mathcal{I}} [y(k,t)M(k,t)]^{n(k,t)} \right], \tag{3.29}$$

where $y(i,t)$ satisfies

$$y(i,t) = \lambda(i,t) + \sum_{j=1}^{N} \sum_{s=1}^{T} y(j,s) q(j,s;i,t), \quad i \in \mathcal{N}, t \in \mathcal{T}, \tag{3.30}$$

and, for $i \in \mathcal{N}$ and $s, t \in \mathcal{T}$,

$$q(j,s;i,t) = \begin{cases} \int_0^\infty p(j,s,y;i,t) dG_{js}(y) & \text{for } j \in \mathcal{I} \\ p(j,s;i,t) & \text{for } j \in \mathcal{N} - \mathcal{I}, \end{cases} \tag{3.31}$$

with C a normalising constant.

Proof : The queueing network described above is an age dependent process as defined in Section 3.3. The associated Q process is also a queueing network but with

routing probabilities given by equation (3.31). Chandy and Martin (1983) showed that these types of queueing networks have equilibrium distribution given by equation (3.29) and are therefore insensitive. By Theorem 3.2 the associated Q process and the $\overline{P}$ process therefore have the same equilibrium distribution. ■

Remarks

(1) Using Theorem 3.2 the supplemented equilibrium distribution can be derived from equation (3.29) by including an additional product of terms of the form $(1 - G_{it}(y))/M(i,t)$ for each customer at a symmetric queue.

(2) The routing probabilities $p(i,t,y;j,s)$ are functions of the service requirements and not the time spent being served. Although the service requirement and time spent being served appear to be identical, the former is the amount of service requested which, unlike the latter, does not take into account the work rate of the service facility.

For example, consider a single server queue employing a Last Come First Served discipline and suppose that an arriving customer demands y units of service. This is worked off at unit rate until either the service is completed or another customer arrives at the queue. If the latter occurs, the new customer moves directly into service and hence the period of time spent at the queue by the first customer is greater than their service requirement. The routing probabilities for this customer are functions of y, not of the time spent at the queue.

(3) Theorem 3.6 may also be proved by setting up the supplemented balance equations for the queueing network described and substituting the supplemented product form described in (1), although this is obviously much more laborious. The problem of applying results based upon a GSMP framework to open networks has been considered by Barbour (1982) and Schassberger (1986), as discussed in greater detail in Chapter 1. The underlying structure of the process described in Section 3.3 is the same as that used in Schassberger's paper (only

the routing probabilities take on a different nature) and thus his analysis also applies here. There is no difficulty dealing with finite systems and hence the analysis may be applied without modification for closed networks.

Example 3.3

Let us consider any network of the kind considered by Chandy and Martin (1983), that is, a collection of quasi-reversible nodes and a constant routing matrix.

In many real situations travel time between nodes plays an important role in the behaviour of the queueing network. A simple way to model this feature is for each customer to visit an infinite server queue after a service completion. It is reasonable to assume that the period spent travelling will influence the customer's future route through the network. For example, customers who encounter unforeseen delays in travelling between venues may alter their planned schedule. Fortunately, an infinite server queue is symmetric and therefore the routing probabilities from these queues may be made functions of the travel time without losing the simplicity of a product form solution. □

Example 3.4

In some situations, if too much time is spent in service at one node of the network, the customer may feel that they do not have enough time to complete a series of tasks. As a result, the customer may shorten their route by abandoning some of their planned future destinations.

Consider a service network consisting of two infinite server queues. Customers from outside the network arrive at queue i in Poisson streams with parameter λ_i, $(i = 1, 2)$, and their service time is distributed $G_i(.)$ with mean μ_i^{-1}. Those completing service at queue 1 move to queue 2 if their service time was less than or equal to W, otherwise they leave the network. Customers completing service at queue 2 immediately depart the network.

Let n_i denote the number of customers in queue i, y the time taken to complete service at queue 1 and $\pi(n_1, n_2)$ the probability of being in state (n_1, n_2). The age dependent routing probability $p_{12}(y)$ of going from queue 1 to 2 is

$$p_{12}(y) = \begin{cases} 1 & \text{if } y \leq W \\ 0 & \text{otherwise.} \end{cases}$$

Let $p_{12} = \int_0^\infty p_{12}(y) dG_1(y) = G_1(W)$. By Theorem 3.6, the equilibrium distribution is given by

$$\pi(n_1, n_2) = D \prod_{i=1}^{n_1} \frac{\lambda_1}{i\mu_1} \prod_{j=1}^{n_2} \frac{p_{12}\lambda_1 + \lambda_2}{j\mu_2},$$

for all

$$G_1(.) \in \{G(.) | G(W) = p_{12} \text{ and } G(.) \text{ has mean } \mu_1^{-1} \},$$

where p_{12} is now considered fixed and D is a normalising constant. □

Example 3.5

Bunday and Scraton (1980) found the supplemented equilibrium distribution for a machine interference problem with R repairers. Their model can be interpreted as a closed queueing network consisting of M machines cycling between a $\cdot/G/\infty$ queue and a $\cdot/M/R$ queue. Bunday and Scraton's model may be generalised not only by incorporating different types of machines which are repaired at different depots but also by allowing the nature of the servicing to be dependent upon the length of time the machine has been operational. For example, whereas routine breakdowns are attended to by the regular repairers, a machine which has operated for a long time without malfunction is sent elsewhere for a general overhaul. While a network comprising more than the two queues stipulated above is required for both generalisations the product form equilibrium distribution is still retained.

Suppose there are $n(t)$ machines of type t, $(t = 1, \ldots, T)$ and T repair and T overhaul centres (labelled $1, \ldots, T$). If a type t machine works for time W_t it is immediately sent to overhaul centre t. On the other hand, if the machine breaks

down after working for time $y < W_t$ it is sent to repair centre t as a type 1 repair with probability $p_t(y)$ and a type 2 repair with probability $1 - p_t(y)$. Upon being repaired, the type 1 repairs are sent for overhaul while type 2 repairs are sent back to work. Typically, we may imagine that $p_t(y)$ is an increasing function in y, as the longer the machine works the more likely it is that it will need to be overhauled.

Machines of type t work for a generally distributed period of time $G_t(.)$ with mean μ_t^{-1}, while repair (respectively overhaul) times are negative exponentially distributed with mean λ_t^{-1} (respectively γ_t^{-1}). There are J_t (respectively K_t) repair (respectively overhaul) people at repair (respectively overhaul) centre t where a first come first served discipline operates.

Denote by node 0 the $\cdot/G/\infty$ queue which models the machines that are currently working and by i_t the number of machines of type t at that node. Let j_{kt} be the number of type k ($k = 1, 2$) repairs at service centre t and m_t the number of machines at overhaul centre t. The state of the system may then be described by the vector

$$\mathbf{n} = \{i_t, j_{1t}, j_{2t}, m_t | t = 1, \ldots, T\},$$

that is, $\mathbf{n}$ gives the number of machines in each part of the network. The equilibrium distribution is given by

$$\pi(\mathbf{n}) = D \prod_{t=1}^{T} \left(\frac{y_{0t}}{\mu_t}\right)^{i_t} \left(\frac{y_{1t}}{\lambda_t}\right)^{j_{1t}} \left(\frac{y_{2t}}{\lambda_t}\right)^{j_{2t}} \frac{1}{i_t! F(j_{1t} + j_{2t}, J_t)} \left(\frac{y_{3t}}{\gamma_t}\right)^{m_t} \frac{1}{F(m_t, K_t)},$$

for feasible states $\mathbf{n}$, where, for $t = 1, \ldots, T$,

$$y_{1t} = p_t y_{0t},$$

$$y_{2t} = (G_t(W_t) - p_t) y_{0t},$$

$$y_{3t} = y_{1t} + (1 - G_t(W_t)) y_{0t},$$

with $y_{0t} > 0$ arbitrary, $p_t = \int_0^{W_t} p_t(y) dG_t(y)$, D a normalising constant and

$$F(j,k) = \begin{cases} j! & j = 0, \ldots, k \\ k! k^{j-k} & j > k. \end{cases}$$

□

3.5 Interruption processes

Henderson and Taylor (1987) examined systems which they term interruption processes. In these processes the generally distributed lifetimes may either die naturally (with hazard function $h_{s1}(y)$ and corresponding distribution function $G_{s1}(y)$) or be interrupted by events which occur in Poisson streams (with rate α_s). The state changes from x to x' with probability $q(x,s,x')$ if the lifetime s dies naturally and with probability $r(x,s,x')$ if it dies as the result of an interruption. Hence the routing probabilities may be considered functions of the age of the lifetime and are given by

$$p(x,s,x',y) = \frac{h_{s1}(y)}{h_{s1}(y)+\alpha_s} q(x,s,x') + \frac{\alpha_s}{h_{s1}(y)+\alpha_s} r(x,s,x'), \quad \forall s \in S^*. \tag{3.32}$$

In the notation used earlier,

$$h_s(y) = h_{s1}(y) + \alpha_s, \tag{3.33}$$

$$G_s(y) = 1 - \exp[-\int_0^y (h_{s1}(u)+\alpha_s)du] \tag{3.34}$$

and

$$\begin{aligned} p(x,s,x') &= \int_0^\infty p(x,s,x',y) dG_s(y) \\ &= q(x,s,x') \int_0^\infty h_{s1}(y) \exp[-\int_0^y (h_{s1}(u)+\alpha_s)du]dy \\ &\quad + r(x,s,x') \int_0^\infty \alpha_s \exp[-\int_0^y (h_{s1}(u)+\alpha_s)du]dy \\ &= q(x,s,x')\hat{G}_{s1}(\alpha_s) + r(x,s,x')(1-\hat{G}_{s1}(\alpha_s)), \end{aligned} \tag{3.35}$$

where $\hat{G}_{s1}(\alpha_s)$ is the Laplace-Stieltjes transform of $G_{s1}(.)$ evaluated at α_s, that is,

$$\hat{G}_s(\alpha_s) = \int_0^\infty \exp(-\alpha_s y) dG_s(y). \tag{3.36}$$

Using Theorem 3.2 we require that the mean of $G_s(.)$ be fixed, μ_s^{-1} say, and that $p(x,s,x')$ is fixed for each $s \in S^*$. The mean is given by

$$\mu_s^{-1} = \int_0^\infty (1-G_s(y))dy = \frac{1}{\alpha_s}(1-\hat{G}_{s1}(\alpha_s)). \tag{3.37}$$

To fix both $p(x,s,x')$ and μ_s^{-1} we need $\hat{G}_{s1}(\alpha_s)$ constant, that is, only one parameter of $G_{s1}(.)$ needs to be specified to satisfy both conditions.

Then using Theorems 3.2 and 3.5 the equilibrium distribution of the interruption process is

$$\overline{\pi}_x(\mathrm{y}) = \pi_x \prod_{s \in S^* \cap x} \mu_s(1 - G_{s1}(y)) \exp(-\alpha_s y_s), \tag{3.38}$$

if and only if (2.8) and (2.9) hold for the associated Q process. This is one of the results of Henderson and Taylor (1987).

Henderson and Taylor (1987) show that if the process $\overline{P}$ is *interruption matched* and $\hat{G}_s(\alpha_s)$ is held constant, then if, for all $G_{s1}(.)$ with $\hat{G}_{s1}(\alpha_s)$ fixed, the equilibrium distribution is unchanged, then the partial balance equations of the associated Q process hold. The interruption matched assumption is not a serious liability. It simply means that if there is a set of states such that, when a lifetime $s \in S^*$ dies naturally in one of these states it is immediately reborn in the set, then the same must hold true if that lifetime dies due to an interruption in that set. In most physical processes the former does not occur and hence no difficulty normally arises.

Our results in this direction are slightly weaker, as we also assume that the age dependent process has the same equilibrium distribution as the associated Q process, and hence that the associated Q process is insensitive.

In the other direction, however, we have a stronger result, as we show that if Q is partially balanced then all age dependent processes (not only interruption processes with routing probabilities $r(x,s,x')$ and $q(x,s,x')$ for $s \in S^*$) having Q as their "average" process will have a product form equilibrium density. In addition, their equilibrium distribution will be the same as that of Q.

Taylor (1987) also looked at generally distributed interruptions. Using the framework of this section such a generalisation occurs naturally.

Generally distributed interruptions

Suppose now that interruptions occur according to a general distribution, $G_{s2}(.)$ with hazard function $h_{s2}(.)$. The age dependent routing probabilities are given by

(3.39)
$$p(x,s,x',y) = \frac{h_{s1}(y)}{h_{s1}(y)+h_{s2}(y)}q(x,s,x') + \frac{h_{s2}(y)}{h_{s1}(y)+h_{s2}(y)}r(x,s,x') \quad \forall s \in S^*.$$

We hence obtain

(3.40)
$$h_s(y) = h_{s1}(y) + h_{s2}(y),$$

(3.41)
$$G_s(y) = 1 - \exp[-\int_0^y (h_{s1}(u) + h_{s2}(u))du],$$

(3.42)
$$\begin{aligned} p(x,s,x') &= \int_0^\infty p(x,s,x',y)dG_s(y) \\ &= q(x,s,x')\int_0^\infty h_{s1}(y)\exp[-\int_0^y (h_{s1}(u)+h_{s2}(u))du]dy \\ &\quad + r(x,s,x')\int_0^\infty h_{s2}(y)\exp[-\int_0^y (h_{s1}(u)+h_{s2}(u))du]dy \\ &= q(x,s,x')Pr[\text{lifetime dies naturally}] + r(x,s,x')Pr[\text{lifetime interrupted}]. \end{aligned}$$

So application of Theorem 3.2 requires that

(3.43)
$$\int_0^\infty \exp[-\int_0^y (h_{s1}(u)+h_{s2}(u))du]dy = \mu_s^{-1}$$

and that Pr[lifetime dies naturally] is fixed. Note that the means of $G_{s1}(.)$ and $G_{s2}(.)$ do not explicitly appear but that the mean of the minimum of $G_{s1}(.)$ and $G_{s2}(.)$ does. The supplemented equilibrium distribution of the process is then given by

$$\overline{\pi}_x(\mathbf{y}) = \pi_x \prod_{s \in S^* \cap x} \mu_s(1 - G_{s1}(y_s))(1 - G_{s2}(y_s))$$

if and only if equations (2.8) and (2.9) hold for the associated Q process.

This result can, of course, be extended to the case of having many generally distributed interruptions.

Henderson and Taylor (1987) also introduced the concept of *n-parameter insensitivity*. That is, the equilibrium distribution may depend on n parameters of the governing lifetime distributions. Interruption processes and insensitivity with respect to the mean (as in Matthes (1962)) both illustrate the concept of single parameter insensitivity, the former showing that the specified parameter need not necessarily be the mean of the distribution. The more general structure introduced here makes it much simpler to give examples of this phenomenon, as illustrated by Example 3.2 which is $n+1$-parameter insensitive.

3.6 Markov renewal processes

Below we give a brief description of a Markov renewal process. For more detail, consult Çinlar (1975), Chapter 10.

Let (X,T) be a Markov renewal process with state space Ω and semi-Markov kernel K, that is, K is the family of probabilities

$$K = \{K(i,j,t) | i,j \in \Omega, t \in R^+\}, \tag{3.44}$$

where

$$K(i,j,t) = Pr[X_{n+1} = j, T_{n+1} - T_n \leq t | X_n = i]. \tag{3.45}$$

Define

$$K(i,j) = \lim_{t \to \infty} K(i,j,t), \tag{3.46}$$

and $\overline{K}$ the matrix whose $(i,j)^{th}$ element is $K(i,j)$. $K(i,j)$ is the probability that the process will move to state j when it leaves state i if no information concerning the period spent in i is known. As noted by Çinlar, $\overline{K}$ is the transition matrix for some Markov chain with state space Ω.

The minimal semi-Markov process $Z = \{Z_t | t \geq 0\}$ associated with (X,T) is

$$Z_t = X_n \quad \text{if } T_n \leq t \leq T_{n+1}. \tag{3.47}$$

This semi-Markov process is an example of the age dependent routing processes described in Section 3.3.

Associated with each state $i \in \Omega$ is a lifetime, s_i say, with distribution function $G_i(y) = \sum_{j\in\Omega} K(i,j,y)$ with mean μ_i^{-1}. Exactly one lifetime is active in any state.

We now need to calculate the age dependent routing probabilities $p(i,j,y)$. The hazard function for $G_i(.)$ is given by

$$h_i(y) = \frac{\frac{d}{dy}G_i(y)}{1 - G_i(y)}. \tag{3.48}$$

The hazard function associated with moving from state i to state j when the lifetime has received an amount of service, y, is

$$h_{ij}(y) = \frac{\frac{d}{dy}K(i,j,y)}{1 - G_i(y)}. \tag{3.49}$$

Hence the age dependent routing probabilities are given by

$$p(i,j,y) = \frac{h_{ij}(y)}{h_i(y)} = \frac{\frac{d}{dy}K(i,j,y)}{\frac{d}{dy}G_i(y)}. \tag{3.50}$$

We have now described the semi-Markov process in terms of a GSMP with age dependent routing.

We now construct the associated Q process. The Q process has state space Ω and routing probabilities

$$\begin{aligned} p(i,j) &= \int_0^\infty p(i,j,y)dG_i(y) \\ &= \int_0^\infty \frac{\frac{d}{dy}K(i,j,y)}{\frac{d}{dy}G_i(y)}dG_i(y) \\ &= \int_0^\infty dK(i,j,y) \\ &= \lim_{y\to\infty} K(i,j,y) \\ &= K(i,j). \end{aligned} \tag{3.51}$$

This is expected since $p(i,j)$ is the average probability of moving from i to j.

As only one lifetime is active in any state, the partial balance equations of the Q process are precisely the global balance equations and hence Q is insensitive. If ν is a solution to $\nu = \nu\overline{K}$ then $\{\pi_i = \nu_i/\mu_i, i \in \Omega\}$ (suitably normalised) is the equilibrium distribution of Q and (by application of Theorem 3.2) of the semi-Markov process Z. This is Theorem 5.22 of Çinlar, (1975). Using Theorem 3.2 the supplemented steady state distribution of Z is given by

$$\overline{\pi}_i(y,z) = \pi_i \mu_i g_i(y+z), \quad i \in \Omega. \tag{3.52}$$

3.7 Why does the Q process work?

Immediately prior to Example 3.1 we mentioned that the equilibrium distribution for age dependent processes is not necessarily the same as that of the associated Q process and used this as the motivation behind our analysis. It is interesting to see why insensitive Q processes lead to age dependent systems with the same equilibrium distribution.

Consider the process $\overline{P}$ defined in Section 3.2. The supplemented balance equations (3.1) to (3.4) may be rewritten as

$$\overline{\pi}_x \sum_{s \in S' \cap x} \lambda_s c(s,x) = \sum_{x' \in \overline{A}} \sum_{s \in S' \cap x'} \overline{\pi}_{x'} \lambda_s c(s,x') p(x',s,x) + \sum_{x' \in A_1} \int_0^\infty \overline{\pi}_{x'}(y) h(y) dy c(t,x') \hat{p}(x',t,x), \qquad \forall x \in \overline{A}, \tag{3.53}$$

$$\overline{\pi}_x(y) \sum_{s \in S' \cap x} \lambda_s c(s,x) + \left[h(y)\overline{\pi}_x(y) + \frac{d}{dy}\overline{\pi}_x(y) \right] c(t,x) = \sum_{x' \in A} \sum_{s \in S' \cap x'} \overline{\pi}_{x'}(y) \lambda_s c(s,x') p(x',s,x), \qquad \forall x \in A_1, \tag{3.54}$$

$$\overline{\pi}_x(y) \sum_{s \in S' \cap x} \lambda_s c(s,x) = \sum_{x' \in A} \sum_{s \in S' \cap x'} \overline{\pi}_{x'}(y) \lambda_s c(s,x') p(x',s,x), \qquad \forall x \in A_0 \tag{3.55}$$

and

$$\overline{\pi}_x(0)c(t,x) = \sum_{x' \in \overline{A}} \sum_{s \in S' \cap x} \overline{\pi}_{x'} \lambda_s c(s,x') p(x',s,x)$$

$$+ \sum_{x' \in A_1} \int_0^\infty \overline{\pi}_{x'}(y) h(y) dy c(t,x') \hat{p}(x',t,x), \qquad \forall x \in A_1, \tag{3.56}$$

where

$$\hat{p}(x,t,x') = \frac{\int_0^\infty \overline{\pi}_x(y) h(y) p(x,t,x',y) dy}{\int_0^\infty \overline{\pi}_x(y) h(y) dy}. \tag{3.57}$$

These equations obviously have the same solution as equations (3.1) to (3.4), and are, in fact, the supplemented balance equations for some process with constant routing probabilities.

Thus, there exists some process with constant routing probabilities which has the same supplemented equilibrium distribution as $\overline{P}$. The problem lies in determining the exact nature of $\hat{p}(x,t,x')$. Obviously, if $\hat{p}(x,t,x')$ is dependent upon $\pi_x(.)$ there is nothing to be gained by writing the balance equations of $\overline{P}$ in the above form, as finding $\hat{p}(x,t,x')$ would then be just as difficult as solving the supplemented balance equations.

The integrand of the numerator in expression (3.57) may be interpreted as the density that the process moves from state x, after t has been alive for time y, into state x'. The numerator may thus be interpreted as the total flow from state x to state x', while the numerator is the total flow out of state x. We may thus interpret $\hat{p}(x,t,x')$ as the "true" average routing probability from state x to state x'.

Note that in general it is not true that $\hat{p}(x,t,x')$ is given by $\int_0^\infty p(x,t,x',y) dG(y)$.

Example 3.1 revisited

For this example,

$$\begin{aligned}
\hat{p}(2,3) &= \frac{\int_0^\infty B_1(\exp(-4y) - \exp(-5y)) 4 \exp(-\gamma y) dy}{\int_0^\infty B_1(\exp(-4y) - \exp(-5(y)) 4 dy} \\
&= 20 \left(\frac{1}{4+\gamma} - \frac{1}{5+\gamma} \right) \\
&\neq p(2,3).
\end{aligned}$$

As noted earlier, using $p(2,3)$ and $p(2,4)$ as the average routing probabilities does not give the correct equilibrium distribution. Using $\hat{p}(2,3)$ and $\hat{p}(2,4)$ does, however, give the right answer. □

So the question arises :

Under what conditons is $\hat{p}(x,t,x')$ easily found?

Well, if the equilibrium distribution of $\overline{P}$ takes on the form of equations (3.5) and (3.6), then substituting into equation (3.57) gives

$$\begin{aligned}\hat{p}(x,t,x') &= \frac{\int_0^\infty \pi_x \mu(1-G(y))h(y)p(x,t,x',y)dy}{\int_0^\infty \pi_x \mu(1-G(y))h(y)dy} \\ &= \frac{\int_0^\infty p(x,t,x',y)dG(y)}{\int_0^\infty dG(y)} \\ &= \int_0^\infty p(x,t,x',y)dG(y),\end{aligned} \tag{3.58}$$

which is the routing probability used in the construction of the Q process. So the "averaging" procedure employed actually gives the correct form of $\hat{p}(x,t,x')$ whenever the $\overline{P}$ process has a product form solution.

3.8 Insensitive average residence times in Φ-insensitive processes

Consider the processes $\overline{P}$ and Q as described in Section 3.3 (at this stage saying nothing about the insensitivity of Q) and assume also that they are ergodic. Let $s \in S$ and denote by ν_s the average length of time between two successive births of s and by γ_s the average length of time between the birth of s and its death. Also define $\overline{\pi}_x$ to be the probability that $\overline{P}$ is in state x.

Lemma 3.7 (Barbour and Schassberger (1981))

ν_s and γ_s are given by

$$\nu_s = \begin{cases} \lambda_s^{-1}\left(\sum_{x\in\Omega_s}\overline{\pi}_x c(s,x)\right)^{-1} & \text{for } s \in S' \\ \mu_s^{-1}\left(\sum_{x\in\Omega_s}\overline{\pi}_x c(s,x)\right)^{-1} & \text{for } s \in S^* \end{cases} \tag{3.59}$$

and

$$\gamma_s = \nu_s \sum_{x \in \Omega_s} \overline{\pi}_x, \tag{3.60}$$

where $\Omega_s = \{x | s \in x\}$.

Proof : See Barbour and Schassberger (1981). ■

Remark : The use of age dependent routing has no effect in the proof of Lemma 3.7 as the only requirements are that $\overline{P}$ be stationary and ergodic. As pointed out by Barbour and Schassberger, this lemma says that the average requested lifetime of s is the product of the average actual lifetime and the average rate at which s is worked off.

We may now extend Theorem 2 of Barbour and Schassberger to processes with age dependent routing probabilities. Let $\gamma_s(y)$ be the expected lifetime of element s when it requests y units of service, and $\nu_s(y)$ the expected time between the birth of s when it requests y units of service and it's next birth.

Theorem 3.8

Let $\overline{P}$ be Φ-insensitive. Then, for each $s \in S^*$,

$$\gamma_s(y) = \mu_s \gamma_s y \tag{3.61}$$

and

$$\nu_s(y) = \mu_s \gamma_s y + \nu_s - \gamma_s. \tag{3.62}$$

Proof : The proof basically follows that of Barbour and Schassberger and as such will only be outlined here, with special note being given to the differences.

We create a new process $\overline{P}^*$ from $\overline{P}$ by replacing each state $x \in \Omega$ by two states $(x, 1)$ and $(x, 2)$. These two states correspond, respectively, to lifetime s requesting less than or greater than y units of service when it was last created. We shall denote

by s_1 and s_2, respectively, the lifetimes corresponding to s in $\overline{P}$ requesting less than or greater than y units of service. The lifetime distributions governing these are thus given by (for $z \geq 0$)

$$G_{s_1}(z) = \begin{cases} G_s(z)/G_s(y) & \text{for } z \leq y \\ 1 & \text{for } z > y \end{cases} \tag{3.63}$$

and

$$G_{s_2}(z) = \begin{cases} 0 & \text{for } z \leq y \\ (G_s(z) - G_s(y))/(1 - G_s(y)) & \text{for } z > y. \end{cases} \tag{3.64}$$

When element s_i, $(i = 1, 2)$, is created, the age dependent routing probabilities of $\overline{P}^*$ are

$$p((x,j),t,(x',i),z) = p(x,t,x',z)\beta_i \quad \text{for} \quad x' \in \Lambda_{ts}(x), \tag{3.65}$$

$$p((x,j),s_j,(x',i),z) = p(x,s,x',z)\beta_i \quad \text{for} \quad x' \in \theta(x) \tag{3.66}$$

and

$$p((x,j),s,(x',i),z) = p(x,s,x',z)\beta_i \quad \text{for} \quad x' \in \Gamma_s(x), s \in x \cap S' \tag{3.67}$$

where $\beta_1 = G_s(y)$ and $\beta_2 = 1 - G_s(y)$, with all other routing probabilities and speeds defined in the natural way.

Let Q^* be the Q process associated with $\overline{P}^*$. The routing probabilities for the Q^* process are

$$p((x,1),s_1,(x',i)) = \begin{cases} \frac{\int_0^y p(x,s,x',z)dG_s(z)}{G_s(y)}\beta_i & \text{for } x, x' \in \Omega_s \\ \frac{\int_0^y p(x,s,x',z)dG_s(z)}{G_s(y)} & \text{for } x \in \Omega_s, x' \notin \Omega_s, i = 1 \end{cases} \tag{3.68}$$

and

$$p((x,2),s_2,(x',i)) = \begin{cases} \frac{\int_y^\infty p(x,s,x',z)dG_s(z)}{1-G_s(y)}\beta_i & \text{for } x, x' \in \Omega_s \\ \frac{\int_y^\infty p(x,s,x',z)dG_s(z)}{1-G_s(y)} & \text{for } x \in \Omega_s, x' \notin \Omega_s, i = 2, \end{cases} \tag{3.69}$$

again, with all other average routing probabilities defined in the obvious way (c.f. Section 3.3).

Barbour and Schassberger verified that if Q is insensitive then Q^* is also insensitive, with equilibrium distribution given by

$$\pi_{(x,i)} = \begin{cases} \beta_i \overline{\pi}_x \mu_s / \mu_{s_i} & x \in \Omega_s, i = 1,2 \\ \beta_i \overline{\pi}_x & x \notin \Omega_s, i = 1,2, \end{cases} \tag{3.70}$$

where $\mu_{s_i}^{-1}$ is the mean of $G_{s_i}(.)$.

As in Barbour and Schassberger (1981),

$$\begin{aligned} \gamma_{s_i} &= \mu_s \gamma_s / \mu_{s_i} \\ &= \frac{\int_0^y \mu_s \gamma_s z dG_s(z)}{G_s(y)}, \end{aligned} \tag{3.71}$$

from (3.63). As y is arbitrary, $\mu_s \gamma_s y$ can be taken as $\gamma_s(y)$, thus giving equation (3.61).

Equation (3.62) is derived using precisely the same arguments as Barbour and Schassberger (1981). ■

Having derived the quantity $\gamma_s(y)$, it is then easy to apply the theory to particular examples, the most obvious of which are the queueing networks of Section 3.4.

By using a GSMP description of a network, $\gamma_s(y)$ would then be interpreted as the expected time a customer spends at the queue given that they request y units of service. More importantly, the actual time spent at the queue is directly proportional to the service requirement.

Barbour and Schassberger (1981) give queueing examples to illustrate the use of the theory. Further work and examples may be found in Jansen (1984).

3.9 A note on the use of age dependent speeds

In this chapter we have only considered the possibility of allowing the routing probabilities to be functions of the age of the lifetime.

We now modify the process $\overline{P}$ described in Section 3.3 by also allowing the speed at which a generally distributed lifetime is worked off to be a function of the age of the lifetime. This is useful in modelling processes where servers may call upon extra facilities if a customer is taking too long to be served.

We alter $\overline{P}$ by replacing $c(s,x)$ by $c(s,x,y)$ for each $s \in S^*$ where y is the amount of service that s has already received.

The supplemented global balance equations (using spent lifetime) of $\overline{P}$ then become

(3.72)

$$\begin{aligned}\overline{\pi}_x(\mathbf{y},0_s)c(s,x,0) = &\sum_{x'\in\Gamma_s(x)} \overline{\pi}_{x'}(\mathbf{y}) \sum_{s'\in S'\cap x'} \lambda_{s'}c(s',x')p(x',s',x) \\ &+ \sum_{s'\in S^*-x} \sum_{x'\in\Lambda_{ss'}(x)} \int_0^\infty \overline{\pi}_{x'}(\mathbf{y},y_{s'})c(s',x',y_{s'})h_{s'}(y_{s'})p(x',s',x,y_{s'})dy_{s'} \\ &+ \sum_{x'\in\theta(x)} \int_0^\infty \overline{\pi}_{x'}(\mathbf{y},y_s)h_s(y_s)c(s,x',y_s)p(x',s,x,y_s)dy_s\end{aligned}$$

and

(3.73)

$$\begin{aligned}&\sum_{s\in S^*\cap x} c(s,x,y_s)\frac{\partial}{\partial y_s}\overline{\pi}_x(\mathbf{y}) + \overline{\pi}_x(\mathbf{y})\left[\sum_{s\in S^*\cap x} h_s(y_s)c(s,x,y_s) + \sum_{s\in S'\cap x} \lambda_s c(s,x)\right] \\ &\quad = \sum_{s\in S^*-x} \sum_{x'\in U_s(x)} \int_0^\infty \overline{\pi}_{x'}(\mathbf{y},y_s)h_s(y_s)c(s,x',y_s)p(x',s,x,y_s)dy_s \\ &\quad + \sum_{x'\in\theta(x)} \overline{\pi}_{x'}(\mathbf{y}) \sum_{s\in S'\cap x'} \lambda_s c(s,x')p(x',s,x).\end{aligned}$$

Define $C_s(y)$ to be the diagonal matrix of age dependent speeds $c(s,x,y)$ for all $y \geq 0$ and $s \in S^*$.

Define, for each $s \in S^*$, Θ_s to be any set of triples $(\mathcal{P}_s(.), C_s(.), G_s(.))$ with the properties

(1) $\int_0^\infty C_s(y)dG_s(y) = C_s(0) = C_s$ for fixed C_s.

(2) $\int_0^\infty C_s(y)\mathcal{P}_s(y)dG_s(y) = C_s\mathcal{P}_s$ for fixed $\mathcal{P}_s$.

(3) The mean of $G_s(.)$ is μ_s^{-1} for fixed μ_s^{-1}.

(4) $(\mathcal{P}_s, C_s, G_s(.)) \in \Theta_s$ for all distribution functions $G_s(.)$ satisfying (3).

The associated Q process is formed in the same way as in Section 3.3, but now with speeds $c(s,x) = \int_0^\infty c(s,x,y)dG_s(y)$ for each $s \in S^*$ and $x \in \Omega$.

Definition 3.4 : The process $\overline{P}$ is Θ-insensitive if for all $(\mathcal{P}_s(.), C_s(.), G_s(.)) \in \Theta_s$, $s \in S^*$, the equilibrium distribution is invariant.

We may now give a generalised form of Theorem 3.2 which allows the use of age dependent speeds.

Theorem 3.9

The following statements are equivalent

(a) $\overline{P}$ is Θ-insensitive.

(b) The associated Q process is insensitive.

(c) The equilibrium distribution of $\overline{P}$ is

$$\overline{\pi}_x(\mathbf{y}) = \pi_x \prod_{s \in S^* \cap x} \mu_s(1 - G_s(y_s)). \tag{3.74}$$

Proof : The proof is similar to that used in Theorem 3.2 and will therefore not be given. ■

Obviously, Theorem 3.2 may be derived from Theorem 3.9 simply by choosing processes with constant speeds. We have chosen to give our results in this fashion as it removes unnecessary distraction from main ideas behind the result. Also, in practice it is very rare, when truly age dependent speeds are used, that the average speed shall be equal to the initial speed and hence these results will not prove as useful as those presented earlier.

CHAPTER 4 : INSENSITIVITY IN SYSTEMS WITH ZERO SPEEDS

4.1 Introduction

In Chapter 3 we dealt with systems possessing instantaneous attention and age dependent routing. In practice, however, many systems allow customers to wait before being allowed into service. For this reason we now turn our attention to such processes and, for ease of explanation, restrict the analysis to the case of constant routing probabilities. It should be noted that the results of Chapter 3 may be extended to the processes to be considered here.

In the literature it has been shown that if partial balance (as defined in Chapter 2) holds then the process under examination is insensitive and has instantaneous attention. However, it is well known that the latter is not a necessary condition for insensitivity.

König and Jansen (1974) called this the e_1 property (the term instantaneous attention being coined by Schassberger (1977)) and mention that processes not possessing this may, by the addition of suitable extra states, be converted to processes that do. Schassberger (1978b) and Whittle (1985) also assume instantaneous attention in their works on insensitivity.

Taylor (1987), on the other hand, adopted a direct approach to this problem and derived necessary and sufficient conditions for such systems to be insensitive. Previous results for systems possessing instantaneous attention arise as special cases. Taylor attacked the problem by setting up supplemented global balance equations using spent lifetimes and finding the equilibrium distribution. In Section 4.2 we shall instead consider the residual lifetime supplemented global balance equations and, by incorporating the ideas of both Taylor and König and Jansen, derive the equilibrium distribution with residual supplementary variables and hence provide an alternative proof of Taylor's main result on processes with zero speeds.

As discussed in Chapter 1, and reexamined in Section 2.3, Henderson (1983a) found that for systems possessing instantaneous attention a necessary and sufficient condition for the insensitivity of the process is that the residual and spent supplementary variable equilibrium distribution be the same. For systems without instantaneous attention this does not hold but a similar relationship is found.

4.2 Systems without instantaneous attention

We shall employ the process P described in Section 2.2. In the following, state changes which involve the death of the generally distributed lifetime, t, shall be referred to as *external* transitions, while all other state changes will be called *internal* transitions.

It shall prove convenient to adopt the matrix notation of Taylor (1987) for the analysis of this system.

We define the following matrices.

$$[Q_{\overline{AA}}]_{xx'} = \begin{cases} -\sum_{\substack{f\in\Omega \\ f\neq x}} \sum_{s\in S'\cap x} \lambda_s c(s,x)p(x,s,f) & \text{if } x = x' \in \overline{A} \\ \sum_{s\in S'\cap x} \lambda_s c(s,x)p(x,s,x') & \text{if } x \neq x' \text{ and } x,x' \in \overline{A}, \end{cases}$$

$$[Q_{\overline{A}i}]_{xx'} = \sum_{s\in S'\cap x} \lambda_s c(s,x)p(x,s,x'), \quad x \in \overline{A} \text{ and } x' \in A_i, (i=0,1),$$

$$[Q_{1\overline{A}}]_{xx'} = c(t,x)p(x,t,x'), \quad x \in A_1 \text{ and } x' \in \overline{A},$$

$$[Q^E_{1i}]_{xx'} = c(t,x)p(x,t,x'), \quad x \in A_1 \text{ and } x' \in A_i, (i=0,1),$$

$$[Q^I_{11}]_{xx'} = \begin{cases} -\sum_{\substack{f\in A \\ f\neq x}} \sum_{s\in S'\cap x} \lambda_s c(s,x)p(x,s,f) & \text{if } x = x' \in A_1 \\ \sum_{s\in S'\cap x} \lambda_s c(s,x)p(x,s,x') & \text{if } x \neq x' \text{ and } x,x' \in A_1, \end{cases}$$

$$[C]_{xx'} = \begin{cases} c(t,x) & \text{for } x = x' \in A_1 \\ 0 & \text{for } x \neq x' \text{ and } x,x' \in A_1, \end{cases}$$

$$[Q_{01}]_{xx'} = \sum_{s \in S' \cap x} \lambda_s c(s,x) p(x,s,x'), \quad x \in A_0 \text{ and } x' \in A_1,$$

$$[Q_{00}]_{xx'} = \begin{cases} -\sum_{\substack{f \in A \\ f \neq x}} \sum_{s \in S' \cap x} \lambda_s c(s,x) p(x,s,f) & \text{if } x = x' \in A_0 \\ \sum_{s \in S' \cap x} \lambda_s c(s,x) p(x,s,x') & \text{if } x \neq x' \text{ and } x, x' \in A_0. \end{cases}$$

We partition the vector Π, the equilibrium distribution for the system with all lifetimes negative exponentially distributed, into $(\Pi_{\overline{A}}, \Pi_1, \Pi_0)$ according to the sets $\overline{A}$, A_1 and A_0.

The global balance equations (2.1), (2.2) and (2.3) for the purely Markov process may be written as

$$\Pi_{\overline{A}} Q_{\overline{A}\overline{A}} + \Pi_1 \mu Q_{1\overline{A}} = 0, \tag{4.1}$$

$$\Pi_1 [Q_{11}^I + \mu Q_{11}^E - \mu C] + \Pi_0 Q_{01} + \Pi_{\overline{A}} Q_{\overline{A}1} = 0 \tag{4.2}$$

and

$$\Pi_0 Q_{00} + \Pi_1 [Q_{10}^I + \mu Q_{10}^E] + \Pi_{\overline{A}} Q_{\overline{A}0} = 0. \tag{4.3}$$

The partial balance equations (2.4) to (2.6) may be written as

$$\Pi_0 Q_{00} + \Pi_1 Q_{10}^I = 0, \tag{4.4}$$

$$\Pi_1 Q_{11}^I + \Pi_0 Q_{01} = 0, \tag{4.5}$$

$$\Pi_1 \mu Q_{10}^E + \Pi_{\overline{A}} Q_{\overline{A}0} = 0 \tag{4.6}$$

and

$$\Pi_1 \mu [Q_{11}^E - C] + \Pi_{\overline{A}} Q_{\overline{A}1} = 0. \tag{4.7}$$

As noted in Section 2.2, the partial balance equation (4.6) can only be satisfied for processes possessing instantaneous attention. Theorem 2.1 may now be rewritten

as: If P possesses instantaneous attention, then it is insensitive if and only if equations (4.4), (4.5) and (4.7) hold.

For $x \in A$, denote by $\overline{\pi}_x$ and $\overline{\pi}^R_x(y)$ the probability of being in states x and (x,y) respectively, where y is the residual lifetime of t. The supplemented global balance equations, using residual lifetime, are

$$\begin{aligned}\overline{\pi}_x \sum_{s\in S'\cap x} &= \lambda_s c(s,x) \sum_{x'\in\overline{A}} \sum_{s\in S'\cap x} \overline{\pi}_{x'}\lambda_s c(s,x')p(x',s,x) \\ &+ \sum_{x'\in A_1} \overline{\pi}^R_{x'}(0)c(t,x')p(x',t,x), \qquad \forall x \in \overline{A},\end{aligned} \tag{4.8}$$

$$\begin{aligned}-\frac{d}{dy}\overline{\pi}^R_x(y)c(t,x) + \overline{\pi}^R_x(y) &\sum_{s\in S'\cap x} \lambda_s c(s,x) \\ &= \sum_{x'\in\overline{A}} \sum_{s\in S'\cap x'} \overline{\pi}_{x'}\lambda_s c(s,x')p(x',s,x)g(y) \\ &+ \sum_{x'\in A} \sum_{s\in S'\cap x'} \overline{\pi}^R_{x'}(y)\lambda_s c(s,x')p(x',s,x) \\ &+ \sum_{x'\in A_1} \overline{\pi}^R_{x'}(0)c(t,x')p(x',t,x) \qquad \forall x \in A_1\end{aligned} \tag{4.9}$$

and, for $x \in A_0$,

$$\begin{aligned}\overline{\pi}^R_x(y) \sum_{s\in S'\cap x} \lambda_s c(s,x) = &\sum_{x'\in\overline{A}} \overline{\pi}_{x'} \sum_{s\in S'\cap x'} \lambda_s c(s,x)p(x',s,x)g(y) \\ &+ \sum_{x'\in A_1} \overline{\pi}^R_{x'}(0)c(t,x)p(x',t,x)g(y) \\ &+ \sum_{x'\in A} \sum_{s\in S'\cap x'} \overline{\pi}^R_{x'}(y)c(s,x')p(x',s,x).\end{aligned} \tag{4.10}$$

By defining $\overline{\Pi}^R_i(y), (i=0,1)$ as the vector of probability densities for being in states of $A_i, (i=0,1)$ with residual sojourn time y, equations (4.8) to (4.10) may be written in matrix form as

$$\overline{\Pi}_{\overline{A}} Q_{\overline{A}A} + \overline{\Pi}^R_1(0)Q_{1\overline{A}} = 0, \tag{4.11}$$

$$-\frac{d}{dy}\overline{\Pi}^R_1(y)C = \overline{\Pi}^R_0(y)Q_{01} + \overline{\Pi}^R_1(y)Q^I_{11} + \overline{\Pi}^R_1(0)Q^E_{11}g(y) + \overline{\Pi}_{\overline{A}}Q_{\overline{A}1}g(y) \tag{4.12}$$

and

$$\overline{\Pi}_0^R(y)Q_{00} + \overline{\Pi}_1^R(y)Q_{10}^I + \overline{\Pi}_{\overline{A}}Q_{\overline{A}0}g(y) + \overline{\Pi}_1^R(0)Q_{10}^E g(y) = 0. \tag{4.13}$$

Theorem 4.1 (Taylor (1987))

A necessary and sufficient condition for the process P to be insensitive is that

$$\Pi_1 Q_1 = 0, \tag{4.14}$$

where $Q_1 = Q_{11}^I - Q_{10}^I(Q_{00})^{-1}Q_{01}$.

Remarks

(1) Note that in the case of instantaneous attention, equation (4.14) may be derived by Gauss-Jordan reduction on equations (4.4) and (4.5).

(2) Taylor proved Theorem 4.1 directly via the use of the supplemented global balance equations based upon spent lifetime. Here we provide an alternative proof using an extended state space (as proposed by König and Jansen (1974)) and residual lifetime as the supplementary variable.

(3) As a consequence of equation (4.14) it is also true that

$$-\Pi_1 Q_2 = 0, \tag{4.15}$$

where

$$\begin{aligned} Q_2 =& \mu C - \mu Q_{11}^E + \mu Q_{1\overline{A}}Q_{\overline{A}\overline{A}}^{-1}Q_{\overline{A}1} + \mu Q_{10}^E Q_{00}^{-1}Q_{01} \\ & - \mu Q_{1\overline{A}}Q_{\overline{A}\overline{A}}^{-1}Q_{\overline{A}0}Q_{00}^{-1}Q_{01}. \end{aligned} \tag{4.16}$$

This may be shown as follows : Noting that $Q_{\overline{A}\overline{A}}$ and Q_{00} are non-conservative q-matrices (Taylor (1987)), it then follows that they are invertible (by application of the Perron-Fröbenius Theorem). Hence equations (4.1) and (4.3) may be rearranged to read

$$\Pi_{\overline{A}} = -\Pi_1 \mu Q_{1\overline{A}}Q_{\overline{A}\overline{A}}^{-1} \tag{4.17}$$

and

$$\Pi_0 = -\left[\Pi_1[Q_{10}^I + \mu Q_{10}^E] + \Pi_{\overline{A}} Q_{\overline{A}0}\right] Q_{00}^{-1}. \tag{4.18}$$

Substituting these expressions into equation (4.2), and using equation (4.14), gives equation (4.15).

Both Q_1 and $-Q_2$ are conservative q-matrices (Taylor (1987)). Q_1 can be interpreted as being the q-matrix of the process P restricted to states in A_1 where only internal transitions are allowed. When an internal transition moves the process to a state in A_0, time is "suspended" until the process moves back into A_1. Similarly, $-Q_2$ is the q-matrix of the process P restricted to states in A_1 but now only allowing external transitions (that is, transitions which involve the death and subsequent rebirth of the generally distributed lifetime). As a consequence, Π_1 is an invariant measure for both of these subprocesses of P. This shall be discussed in more detail later in this Chapter.

Proof : Try a solution to the residual supplemented global balance equations of the form

$$\overline{\Pi}_{\overline{A}} = \Pi_{\overline{A}}, \tag{4.19}$$

$$\overline{\Pi}_1^R(y) = \Pi_1 \mu (1 - G(y)) \tag{4.20}$$

and

$$\overline{\Pi}_0^R(y) = -\Pi_1 Q_{10}^I Q_{00}^{-1} \mu (1 - G(y)) + [\Pi_0 + \Pi_1 Q_{10}^I Q_{00}^{-1}] g(y), \tag{4.21}$$

where $\Pi = (\Pi_{\overline{A}}, \Pi_1, \Pi_0)$ is the equilibrium distribution of the process P.

Substituting, from relations (4.19) to (4.21) into equation (4.11) gives equation (4.1).

Substituting into equation (4.12) gives

$$\begin{aligned}\mathbf{\Pi}_1\mu g(y)C = &-\mathbf{\Pi}_1 Q_{10}^I Q_{00}^{-1} Q_{01}\mu(1-G(y)) + [\mathbf{\Pi}_0 + \mathbf{\Pi}_1 Q_{10}^I Q_{00}^{-1}]g(y)Q_{01} \\ &+ \mathbf{\Pi}_1\mu(1-G(y))Q_{11}^I + \mathbf{\Pi}_1\mu Q_{11}^E g(y) + \mathbf{\Pi}_{\overline{A}} Q_{\overline{A}1} g(y).\end{aligned} \tag{4.22}$$

Using equation (4.14), we have

$$\mathbf{\Pi}_1[-Q_{10}^I Q_{00}^{-1} Q_{01} + Q_{11}^I]\mu(1-G(y)) = 0 \tag{4.23}$$

and, by also utilising equation (4.2),

$$[\mathbf{\Pi}_1[Q_{11}^I + \mu Q_{11}^E - \mu C] + \mathbf{\Pi}_0 Q_{01} + \mathbf{\Pi}_{\overline{A}} Q_{\overline{A}1}]g(y) = 0. \tag{4.24}$$

Hence equation (4.12) is satisfied by the proposed solution.

Substituting into the left hand side of equation (4.13) gives

$$\begin{aligned}&-\mathbf{\Pi}_1 Q_{10}^I Q_{00}^{-1} Q_{00}\mu(1-G(y)) + [\mathbf{\Pi}_0 + \mathbf{\Pi}_1 Q_{10}^I Q_{00}^{-1}]Q_{00}g(y) \\ &+ \mathbf{\Pi}_1 Q_{10}^I \mu(1-G(y)) + \mathbf{\Pi}_{\overline{A}} Q_{\overline{A}0} g(y) + \mathbf{\Pi}_1\mu Q_{10}^E g(y).\end{aligned} \tag{4.25}$$

Simplifying, we obtain

$$[\mathbf{\Pi}_0 Q_{00} + \mathbf{\Pi}_1[Q_{10}^I + \mu Q_{10}^E] + \mathbf{\Pi}_{\overline{A}} Q_{\overline{A}0}]g(y) \tag{4.26}$$

which is equal to zero by equation (4.3).

Hence equations (4.19), (4.20) and (4.21) give the equilibrium distribution of P. Integrating gives the insensitivity of the process, thus showing the sufficiency of equation (4.14).

For the necessity of equation (4.14) we utilise the results of König and Jansen (1974) by converting the process P to a system V which exhibits instantaneous attention. This idea has been put forward by König and Jansen but not explicitly used.

The process P may be represented diagrammatically as in Figure 4.1. To create V we add extra states to those of P and appropriately define the parameters of the

new process (V is represented diagrammatically as in Figure 4.2). This is done in the following fashion.

States of V :

(1) For each state in $\overline{A}$ (respectively A_1) of P there exists a one to one correspondence with states in $\overline{A}_V$ (respectively A_{1V}) of V.

(2) For states in A_0 of P there is a one to two correspondence with states in $A_{ZV} \cup A_{0V}$ in V, distinguished by whether or not t has been worked off with positive speed. That is, if V is in state $x \in A_{ZV}$ (corresponding to state $w \in A_0$) then this means that P is in state w, but the generally distributed lifetime has not yet been worked on with positive speed. On the other hand, if V is in state $x \in A_{0V}$ (corresponding to state $w \in A_0$) then this means that P is in state w and the generally distributed lifetime has been worked on with positive speed.

We denote by Ω_V the states of V and note that it is possible that some of these states are transient.

Lifetimes of V :

The set S of lifetimes of V is the same as the set of lifetimes of P. If $x \in \Omega_V - A_{ZV}$ corresponds to state $w \in \Omega$ then x has precisely the same lifetimes associated with it as the state w has in the P process. For $x \in A_{ZV}$, corresponding to w in A_0, the same lifetimes are active except for the element t, which is not considered to have been created.

Speeds and routing probabilities of V :

Denote by the subscript V the speeds and routing probabilities of the V process. If $x, x' \in \Omega_V$ correspond to $w, w' \in \Omega$ respectively, we define

$$c_V(s,x) = c(s,w) \quad \forall x \in \Omega_V, s \in S \cap x,$$

$$p_V(x,s,x') = p(w,s,w') \quad \forall x \in \overline{A}_V, x' \in \overline{A}_V \cup A_{1V} \cup A_{ZV}, s \in S' \cap x,$$

$$p_V(x,s,x') = p(w,s,w') \quad \forall x, x' \in A_{1V} \cup A_{0V}, s \in S' \cap x,$$

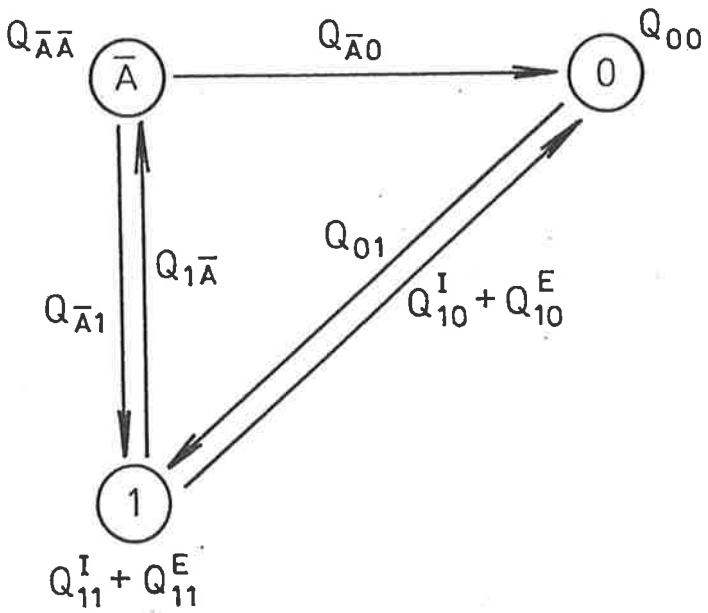


Figure 4.1. Schematic representation of the process P with the appropriate transition matrices marked.

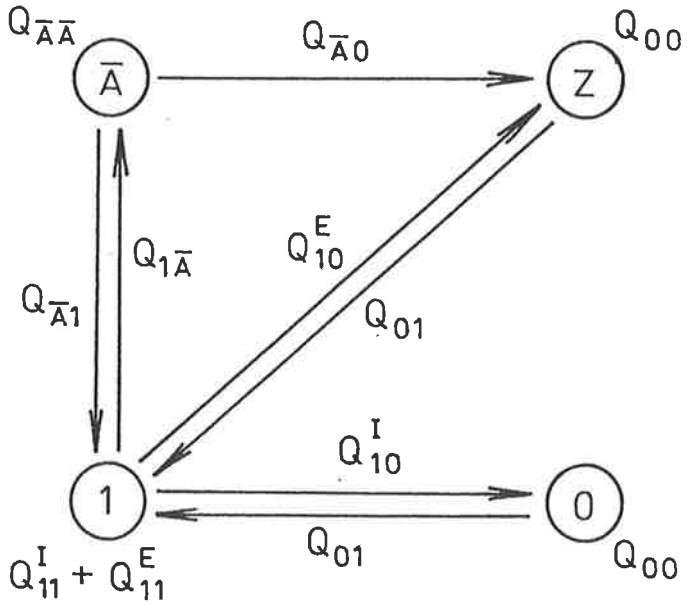


Figure 4.2. Schematic representation of the process V with the appropriate transition matrices marked.

$$p_V(x,s,x') = p(w,s,w') \quad \forall x \in A_{ZV}, x' \in A_{1V} \cup A_{ZV}, s \in S' \cap x,$$

$$p_V(x,t,x') = p(w,t,w') \quad \forall x \in A_{1V}, x' \in \overline{A}_V \cup A_{1V} \cup A_{ZV}.$$

All other routing probabilities are zero.

So for systems with instantaneous attention $p_V(x,s,x') = 0$ for $x \in \overline{A}_V \cup A_{1V}, x' \in A_{ZV}$. Thus the states of A_{ZV} are transient and play no part in the equilibrium analysis.

By defining $\underline{\Pi}_i^R(y), (i = 0,1)$ as the vector of probability densities for being in states of $A_{iV}, (i = 0,1)$ with residual sojourn time y, the residual supplemented balance equations of V are

$$\underline{\Pi}_{\overline{A}} Q_{\overline{A}A} + \underline{\Pi}_1^R(0) Q_{1\overline{A}} = 0 \tag{4.27}$$

$$\underline{\Pi}_Z Q_{00} + \underline{\Pi}_1^R(0) Q_{10}^E + \underline{\Pi}_{\overline{A}} Q_{\overline{A}0} = 0 \tag{4.28}$$

$$-\frac{d}{dy}\underline{\Pi}_1^R(y)C = \underline{\Pi}_0^R(y)Q_{01} + \underline{\Pi}_1^R(y)Q_{11}^I + \underline{\Pi}_1^R(0)Q_{11}^E g(y) + \underline{\Pi}_Z Q_{01} g(y) + \underline{\Pi}_{\overline{A}} Q_{\overline{A}1} g(y) \tag{4.29}$$

$$\underline{\Pi}_0^R(y)Q_{00} + \underline{\Pi}_1^R(y)Q_{10}^I = 0 \tag{4.30}$$

Lemma 4.2

P is insensitive if and only if V is insensitive.

Proof : Remembering that there exists a one to one correspondence between the states of $\overline{A} \cup A_1$ and the states of $\overline{A}_V \cup A_{1V}$, and that the periods spent in these states is identical in the two processes, we must then have $\underline{\Pi}_1 = \overline{\Pi}_1$ and $\underline{\Pi}_{\overline{A}} = \overline{\Pi}_{\overline{A}}$. Similarly, the period spent in states of A_0 is made up of the periods spent in the corresponding states of $A_{0V} \cup A_{ZV}$ and hence $\overline{\Pi}_0 = \underline{\Pi}_0 + \underline{\Pi}_Z$.

Therefore, if V is insensitive the equilibrium distribution of V depends on $G(.)$ only through its mean and hence P is insensitive.

On the other hand, if P is insensitive then $(\overline{\Pi}_1, \overline{\Pi}_0, \overline{\Pi}_{\overline{A}})$ depends upon $G(.)$ only through its mean and hence $\underline{\Pi}_1$ and $\underline{\Pi}_{\overline{A}}$ are invariant. Integrating equation (4.30) for y going from 0 to infinity gives that $\underline{\Pi}_0$ is also invariant. As $\underline{\Pi}_Z = \overline{\Pi}_0 - \underline{\Pi}_0$, the insensitivity of V follows. ■

Let $\hat{\Pi} = (\hat{\Pi}_{\overline{A}}, \hat{\Pi}_1, \hat{\Pi}_0, \hat{\Pi}_Z)$ be the equilibrium distribution of V when all lifetimes are negative exponentially distributed.

The global balance equations of V with purely negative exponential lifetimes may now be written as

$$\hat{\Pi}_{\overline{A}} Q_{\overline{A}\overline{A}} + \hat{\Pi}_1 \mu Q_{1\overline{A}} = 0, \tag{4.31}$$

$$\hat{\Pi}_Z Q_{00} + \hat{\Pi}_1 \mu Q^E_{10} + \hat{\Pi}_{\overline{A}} Q_{\overline{A}0} = 0, \tag{4.32}$$

$$\hat{\Pi}_1 [Q^I_{11} + \mu Q^E_{11} - \mu C] + \hat{\Pi}_Z Q_{01} + \hat{\Pi}_0 Q_{01} + \hat{\Pi}_{\overline{A}} Q_{\overline{A}1} = 0, \tag{4.33}$$

$$\hat{\Pi}_0 Q_{00} + \hat{\Pi}_1 Q^I_{10} = 0. \tag{4.34}$$

The complementary process V now has instantaneous attention, hence using Theorem 2.1, V is insensitive only if $\hat{\Pi}$ satisfies the partial balance equations. That is,

$$\hat{\Pi}_0 Q_{00} + \hat{\Pi}_1 Q^I_{10} = 0, \tag{4.35}$$

$$\hat{\Pi}_1 Q^I_{11} + \hat{\Pi}_0 Q_{01} = 0 \tag{4.36}$$

and

$$\hat{\Pi}_1 \mu [Q^E_{11} - C] + \hat{\Pi}_{\overline{A}} Q_{\overline{A}1} + \hat{\Pi}_Z Q_{01} = 0. \tag{4.37}$$

As noted in Section 2.2, equation (4.37) holds immediately if equations (4.35) and (4.36) are satisfied.

Gauss-Jordan reduction on equations (4.35) and (4.35) gives (4.14), completing the proof of the theorem. ■

4.3 Extending generalised balance to systems with zero speeds

In this section we explore the relationship between the residual and spent lifetime probability densities and the insensitivity of the processes under consideration. Part of this relationship stems from the property of generalised balance introduced by Henderson (1983a). We shall extend a theorem of Henderson on insensitivity in processes with instantaneous attention to systems without this feature.

We again deal with processes with one generally distributed lifetime and employ the same processes P and V as in Section 4.2.

The following definitions are from Henderson (1983a). Let

1. $Z_R(u) = (x(u), y_R(u))$ where $x(u)$ is the state of the process at time u and $y_R(u)$ is the residual sojourn time in A.

2. $Z_S(u) = (x(u), y_S(u))$ where $x(u)$ is the state of the process at time u and $y_S(u)$ is the spent sojourn time in A.

3. The process P has the property of *generalised balance* if and only if

$$\begin{aligned} &\lim_{\Delta\to 0^-} \frac{1}{\Delta}[1 - Pr[Z_R(u+\Delta) = (x, y - c(t,x)\Delta) | Z_R(u) = (x,y)]] \\ &\quad = \lim_{\Delta\to 0^-} \frac{1}{\Delta}[1 - Pr[Z_S(u-\Delta) = (x, y - c(t,x)\Delta) | Z_S(u) = (x,y)]] \end{aligned} \tag{4.38}$$

and

$$\begin{aligned} &\lim_{\Delta\to 0^+} \frac{1}{\Delta}[1 - Pr[Z_R(u+\Delta) = (x, y - c(t,x)\Delta) | Z_R(u) = (x,y)]] \\ &\quad = \lim_{\Delta\to 0^+} \frac{1}{\Delta}[1 - Pr[Z_S(u-\Delta) = (x, y - c(t,x)\Delta) | Z_S(u) = (x,y)]] \end{aligned} \tag{4.39}$$

This property equates the rates at which lifetimes are dying in forward and reversed time. As noted by Henderson, processes consisting only of negative exponentially distributed elements will automatically possess generalised balance as the

above equations then reduce to the global balance equations of the process. Note also that no change has been made to the definition of generalised balance even though the assumption of instantaneous attention has been dropped.

Henderson defined the above quantities for processes with many generally distributed lifetimes, but we shall only consider the case with one generally distributed element.

We employ the natural notation that superscript R (respectively S) refers to the state description using residual (respectively spent) sojourn times.

Theorem 4.3 (Henderson (1983a))

The following conditions are equivalent in a GSMP:

(a) The process is insensitive with respect to the generally distributed lifetime t.

(b) $\overline{\Pi}_1^R(y) = \overline{\Pi}_1^S(y)$ and $\overline{\Pi}_0^R(y) = \overline{\Pi}_0^S(y)$.

(c) Generalised balance is satisfied.

Proof : See Henderson (1983a). ■

This theorem states that the equilibrium density using residual lifetime is exactly the same as the density when spent lifetime is employed if the process being considered is insensitive and has instantaneous attention.

This relationship does not hold, however, when the process does not possess instantaneous attention, but we can prove a similar theorem.

Theorem 4.4

The following conditions are equivalent.

(a) The process is insensitive with respect to the distribution of nominal sojourn time in A.

(b) $\overline{\Pi}_1^R(y) = \overline{\Pi}_1^S(y)$.

(c) Generalised balance is satisfied.

Proof : As in Henderson (1983a) we show (a)⇔(b) and (a)⇔(c).

The supplementary global balance equations (with spent lifetime) are

$$\overline{\Pi}_{\overline{A}} Q_{\overline{A}A} + \int_0^\infty \overline{\Pi}_1^S(y) h(y) dy Q_{1\overline{A}} = 0, \tag{4.40}$$

$$\overline{\Pi}_1^S(y) Q_{11}^I + \overline{\Pi}_0^S(y) Q_{01} = [h(y)\overline{\Pi}_1^S(y) + \frac{d}{dy}\overline{\Pi}_1^S(y)]C, \tag{4.41}$$

$$\overline{\Pi}_0^S(y) Q_{00} + \overline{\Pi}_1^S(y) Q_{10}^I = 0, \tag{4.42}$$

$$\overline{\Pi}_{\overline{A}} Q_{\overline{A}1} + \int_0^\infty \overline{\Pi}_1^S(y) h(y) dy Q_{11}^E + \overline{\Pi}_0^S(0) Q_{01} = \overline{\Pi}_1^S(0) C \tag{4.43}$$

and

$$\overline{\Pi}_0^S(0) Q_{00} + \overline{\Pi}_{\overline{A}} Q_{\overline{A}0} + \int_0^\infty \overline{\Pi}_1^S(y) h(y) dy Q_{10}^E = 0. \tag{4.44}$$

Note that

$$\overline{\Pi}_1^R(0) = \int_0^\infty \overline{\Pi}_1^S(y) h(y) dy. \tag{4.45}$$

Equation (4.45) arises by looking at the probability density, using either the residual or spent supplementary variable, of the lifetime t being about to die.

(b)⇒(a). If (b) holds, then

$$\overline{\Pi}_0^R(y) = \overline{\Pi}_0^S(y) + \overline{\Pi}_0^S(0) g(y). \tag{4.46}$$

Equation (4.46) holds because the probability density that the process is in some state x in A_0 with residual service time y is made up of two components, the first being that the process has moved into state x and the generally distributed lifetime has been worked on while the second is the case where it has been created but not yet worked upon. The density of the former is given by $\overline{\Pi}_0^S(y)$ (by application of

Theorem 2.3 to the V process) while the latter is given by $\hat{\Pi}_Z g(y)$, as $g(y)$ is the density that the lifetime lasts for time y. This quantity, however, is $\overline{\Pi}_0^S(0)g(y)$.

Using equation (4.45), and without loss of generality denoting $\overline{\Pi}_1^R(y)$ and $\overline{\Pi}_1^S(y)$ by $\overline{\Pi}_1(y)$, the spent supplemented global balance equations, (4.40) to (4.44) may be rewritten as

$$\overline{\Pi}_{\overline{A}} Q_{\overline{A}\overline{A}} + \int_0^\infty \overline{\Pi}_1(y)h(y)dy Q_{1\overline{A}} = 0, \tag{4.47}$$

$$\overline{\Pi}_1(y)Q_{11}^I + \overline{\Pi}_0^S(y)Q_{01} = [h(y)\overline{\Pi}_1(y) + \frac{d}{dy}\overline{\Pi}_1(y)]C, \tag{4.48}$$

$$\overline{\Pi}_0^S(y)Q_{00} + \overline{\Pi}_1(y)Q_{10}^I = 0, \tag{4.49}$$

$$\overline{\Pi}_{\overline{A}} Q_{\overline{A}1} + \overline{\Pi}_1(0)Q_{11}^E + \overline{\Pi}_0^S(0)Q_{01} = \overline{\Pi}_1(0)C \tag{4.50}$$

and

$$\overline{\Pi}_0^S(0)Q_{00} + \overline{\Pi}_{\overline{A}} Q_{\overline{A}0} + \overline{\Pi}_1(0)Q_{10}^E = 0. \tag{4.51}$$

The residual supplemented balance equations, (4.11) to (4.13), may be written in the form

$$\overline{\Pi}_{\overline{A}} Q_{\overline{A}\overline{A}} + \overline{\Pi}_1(0)Q_{1\overline{A}} = 0, \tag{4.52}$$

$$-\frac{d}{dy}\overline{\Pi}_1(y)C = \overline{\Pi}_0^R(y)Q_{01} + \overline{\Pi}_1(y)Q_{11}^I + \overline{\Pi}_1(0)Q_{11}^E g(y) + \overline{\Pi}_{\overline{A}} Q_{\overline{A}1} g(y) \tag{4.53}$$

and

$$\overline{\Pi}_0^R(y)Q_{00} + \overline{\Pi}_1(y)Q_{10}^I + \overline{\Pi}_{\overline{A}} Q_{\overline{A}0} g(y) + \overline{\Pi}_1(0)Q_{10}^E g(y) = 0. \tag{4.54}$$

Rearranging equation (4.49) and substituting into (4.48) gives

$$\overline{\Pi}_1(y)Q_1 = [h(y)\overline{\Pi}_1(y) + \frac{d}{dy}\overline{\Pi}_1(y)]C. \tag{4.55}$$

Adding $g(y)\times$(4.50) to equation (4.53) and rearranging gives

$$\begin{aligned}\frac{d}{dy}\overline{\Pi}_1(y)C + \overline{\Pi}_1(0)Cg(y) &= [\overline{\Pi}_0^S(0)g(y) - \overline{\Pi}_0^R(y)]Q_{01} - \overline{\Pi}_1(y)Q_{11}^I \\ &= -\overline{\Pi}_0^S(y)Q_{01} - \overline{\Pi}_1(y)Q_{11}^I \\ &= -\overline{\Pi}_1(y)Q_1.\end{aligned} \tag{4.56}$$

Comparing equations (4.55) and (4.56) we obtain

$$-\frac{d}{dy}\overline{\Pi}_1(y) - \overline{\Pi}_1(0)g(y) = \frac{d}{dy}\overline{\Pi}_1(y) + \overline{\Pi}_1(y)h(y), \tag{4.57}$$

which has solution

$$\overline{\Pi}_1(y) = \Pi_1(0)(1 - G(y)). \tag{4.58}$$

Let

$$\overline{\Pi}_1 = \int_0^\infty \overline{\Pi}_1(y)dy. \tag{4.59}$$

Then integrating equation (4.58) for values of y between 0 and infinity gives that $\overline{\Pi}_1(0) = \overline{\Pi}_1\mu$. Substituting for $\overline{\Pi}_1(y)$ in equations (4.48) and (4.49) then gives

$$\overline{\Pi}_1\mu(1 - G(y))Q_{11}^I + \overline{\Pi}_0^S(y)Q_{01} = 0 \tag{4.60}$$

and

$$\overline{\Pi}_1\mu(1 - G(y))Q_{10}^I + \overline{\Pi}_0^S(y)Q_{00} = 0. \tag{4.61}$$

respectively. Let

$$\overline{\Pi}_0^{S+} = \lim_{\tau\to 0}\int_\tau^\infty \overline{\Pi}_0^S(y)dy. \tag{4.62}$$

Integrating equations (4.60) and (4.61) then produces

$$\overline{\Pi}_1 Q_{11}^I + \overline{\Pi}_0^{S+}Q_{01} = 0 \tag{4.63}$$

and

$$\overline{\Pi}_1 Q_{10}^I + \overline{\Pi}_0^{S+} Q_{00} = 0 \tag{4.64}$$

which may be rearranged to give (4.14) and hence gives the insensitivity of P.

(a)⇒(b). From Theorem 4.1 we have insensitivity if and only if $\Pi_1 Q_1 = 0$. Taylor has shown that the equilibrium distribution using spent sojourn time as the supplementary variable is given by

$$\overline{\Pi}_{\overline{A}}^S = \Pi_{\overline{A}}, \tag{4.65}$$

$$\overline{\Pi}_1^S(y) = \Pi_1 \mu(1 - G(y)), \tag{4.66}$$

$$\overline{\Pi}_0^S(y) = -\Pi_1 Q_{10}^I (Q_{00})^{-1} \mu(1 - G(y)) \tag{4.67}$$

and

$$\overline{\Pi}_0^S(0) = \Pi_0 + \Pi_1 Q_{10}^I (Q_{00})^{-1}. \tag{4.68}$$

Comparing equation (4.66) with equation (4.20) gives (b).

(a)⇒(c). The insensitivity of the process gives that equations (4.19) to (4.21) and (4.65) to (4.68) are the residual and spent supplementary variable distributions respectively. By considering only the states where the generally distributed lifetime is being worked off with positive speed the proof is now the same as in Henderson (1983a).

(c)⇒(a). For states in A, the instantaneous attention assumption has no bearing in the derivation of equation (3.8) of Henderson (1983a), namely

$$\frac{d}{dy}\overline{\Pi}_1^R(y) + \overline{\Pi}_1^R(y)h(y) = 0. \tag{4.69}$$

This has solution

$$\overline{\Pi}_1^R(y) = \overline{\Pi}_1^R(0)(1 - G(y)). \tag{4.70}$$

By defining $\overline{\Pi}_1 = \int_0^\infty \overline{\Pi}_1^R(y)dy$, we have that $\overline{\Pi}_1^R(y) = \overline{\Pi}_1\mu(1 - G(y))$. Substituting into equation (4.11) gives

$$\overline{\Pi}_{\overline{A}} = -\overline{\Pi}_1\mu Q_{1\overline{A}}Q_{\overline{A}\overline{A}}^{-1}. \tag{4.71}$$

Substituting into equation (4.13) and rearranging gives

$$\overline{\Pi}_0^R(y) = -\overline{\Pi}_1\mu\left[Q_{10}^I(1 - G(y)) - Q_{1\overline{A}}Q_{\overline{A}\overline{A}}^{-1}Q_{\overline{A}0}g(y) + Q_{10}^E g(y)\right]Q_{00}^{-1}. \tag{4.72}$$

Substituting from expression (4.72) into equation (4.12) and rearranging gives

$$-\overline{\Pi}_1 Q_2 g(y) = \overline{\Pi}_1 Q_1(1 - G(y)), \tag{4.73}$$

which only holds for all $y > 0$ if $\overline{\Pi}_1 Q_1 = -\overline{\Pi}_1 Q_2 = 0$. By Theorem 4.1 we then have that P is insensitive, thus completing the proof. ■

The work on insensitivity with non-instantaneous attention to this point emphasizes that a form of partial balance is required *not* when the generally distributed lifetime is created, but when it is first worked on with positive speed. That is, the flow into a state of A_1 where t is being worked on *for the first time* must be balanced by the flow out of that state due to the death of the lifetime. This is illustrated in the physical interpretation of Theorem 4.1 and in Theorem 4.4 where we only require that the spent and residual equilibrium densities be matched on states of A_1.

This is most simply thought of as the superimposing of one process upon another, where the set of states in common have the same equilibrium measure in both processes. Physically, the process appears to be unable to distinguish whether it has moved into these states due to the death of the generally distributed lifetime or a negative exponentially distributed lifetime.

We are now in a position to extend the interpretation of Theorem 2 of Henderson (1983a) (with notation appropriately changed for the current context) to processes not having instantaneous attention. This is done by restating this theorem in terms of the process V.

We first define the following.

Let $p'(n,s,m)$ be the transition probabilities for going from state n to m due to the death of element s in the reverse time GSMS.

Let $q'(n,m)$ be the reverse time transition rate from n to m due the death of an active element in $n \cap S'$.

Theorem 4.5 (Henderson, Theorem 2, 1983a)

Suppose that the speed, $c(s,x)$, that lifetime $s \in S$ is worked on in state $x \in \Omega_V$ is the same in both forward and reverse time.

The GSMP V is insensitive with respect to the distribution of t if and only if the set of balance equations (4.74) to (4.76) are valid for all n,m,i,j. If $n \in \overline{A}_V \cup A_{ZV}$ then

$$(4.74) \qquad \sum_{s \in n \cap S'} \lambda_s = \sum_{m \in A_{1V}} \frac{p(m,t,n)}{p'(m,t,n)} \sum_{s \in n \cap S'} \lambda_s p(n,s,m) + \sum_{m \in \overline{A}_V \cup A_{ZV}} q'(n,m).$$

If $m \in \overline{A}_V \cup A_{ZV}, n \in A_{1V}$ then

$$(4.75) \qquad \pi_n c(t,n) \mu p'(n,t,m) = \pi_m \sum_{s \in m \cap S'} \lambda_s p(m,s,n).$$

If $m,n \in \overline{A}_V \cup A_{ZV}$ or $m,n \in A_{1V} \cup A_{0V}$ then

$$(4.76) \qquad \pi_n q'(n,m) = \pi_m \sum_{s \in m \cap S'} \lambda_s p(m,s,n).$$

■

For all intents and purposes, the set of states A_{ZV} may be considered to be states where t is not alive, since it is yet to have been worked on and $\overline{A}_V$ may not be entered from this set. Thus, for the process P, we then have that equation (4.75)

equates the flow out of state n into state m due to the death of t in reverse time with the forward time flow of moving out of state m into state n and working on t with positive speed for the first time.

Equation (4.76) then refers to state changes that do not involve the death of t or its being worked on for the first time with positive speed. These equations will automatically hold as they are just the statement of the relationship between the forward and reverse time process (see Theorem 1.12, Kelly (1979)).

The idea of creating an expanded state space, however, becomes impractical when more generally distributed lifetimes are included. Very quickly the state space becomes extremely large as we must keep track of all the combinations of lifetimes which have or have not been worked on with positive speed (for lifetimes not receiving instantaneous attention). Taylor (1987) has derived sufficient conditions for the insensitivity of such processes. These are equivalent to the creation of a process with instantaneous attention, (that is, a process with the expanded state space referred to above), even though they were derived directly from the original process.

PART TWO

CLOSED TWO NODE PRIORITY QUEUEING NETWORKS

CHAPTER 5 : INTRODUCTION

It is common in queueing systems to have a variety of customer types competing for the available resources of a service facility. This competition may be resolved in many ways, for example, a First In First Out (FIFO) or Last In First Out (LIFO) service discipline. In other systems, particular types may be given preference over others, that is, are given priority over other types of customers. In such systems, if a customer is in service at the time of arrival of a higher priority customer, the latter may move directly into service (known as *pre-emptive priority*) or wait until the current job is completed before moving in (*nonpre-emptive priority*).

The analytic results available on equilibrium distributions for queues with priorities is reasonably limited, with computationally useful results restricted in the main to systems with Poisson arrivals and negative exponential service times. A larger number of results are available concerning waiting time distributions, expected queue lengths and expected waiting times.

White and Christie (1958) were the first to find the equilibrium distribution for a single server pre-emptive priority queue with two types of customer arriving in Poisson streams and having negative exponentially distributed service times. The form of solution is so complex as to not be readily usable for practical purposes. For the same system, Stephan (1958) derived more readily computable formulae for the joint queue length distribution as well as finding moments of the waiting time distribution for low priority customers. Miller (1959) found the queue length distribution and expected waiting times for systems of the type studied by White and Christie except that a nonpre-emptive priority queueing discipline was utilised. The complex nature of the results are an indication of the difficulty in finding any sort of exact solution for systems featuring nonpre-emptive priorities (in fact, systems with priorities in general).

In the book "Priority Queues", Jaiswal (1968) gave an in-depth analysis of many priority queueing disciplines based upon the completion time approach. This approach

(introduced by Gaver (1962) and Keilson (1962)) considers the distribution of time taken for a low priority customer to complete service once they have started. This distribution (the completion time distribution) is then used as the service time distribution of the low priority customer in a process where they are the only customer type and the analysis is then similar to that of the ordinary $M/G/1$ queue.

Jaiswal considered the system studied by White and Christie (1958) and Stephan (1958) but allowed the service times of customers to be generally distributed. The results include joint queue length distributions during the busy period (in generating function form) and the busy period distribution. Variations of this system were also analysed, in particular, systems where there are bounds on the numbers of each type of customer present. Analyses of pre-emptive repeat-different and repeat-identical disciplines as well as nonpre-emptive schemes were also given.

Utilising supplementary variables for the time until service completion, Henderson (1969) developed a relatively simple technique for finding the generating function (in terms of Laplace Transforms) of the joint queue length in many standard priority queueing models with general service time distributions. Henderson also examined priority systems of the $G/M/1$ type, that is, where the interarrival times are generally distributed. Hokstad (1978) employed Henderson's technique in the analysis of queues where the priorities were determined by the length of the service time.

The use of generating functions and Laplace Transforms makes life difficult from a computational viewpoint. As a result, much recent work has focused on techniques (both exact and approximate) which make it easier to find performance measures such as queue length distributions, utilisations, throughputs, etc.

Brandwajn (1982) used a finite difference approach to find the equilibrium joint queue length distribution for a two level priority queue, high priority customers having pre-emptive priority, with service and interarrival times negative exponentially distributed with state dependent rate for low priority customers. As pointed out by

Brandwajn, it may be preferable to employ generating functions to find queue characteristics such as mean queue length, but the finite difference approach is probably more amenable to finding the joint queue length distribution.

It should be noted that all of the work cited above deals with single server queues, that is, queues in which the total service effort is fixed.

Much of the work on networks of priority queues involves the use of approximation techniques. These are, in the main, based on the idea of using *virtual servers*, as first introduced by Sevcik (1977) (who used the term "reduced occupancy approximation").

Sevcik looked at N node networks, with one queue a priority scheduling, single server queue with pre-emptive resume discipline. Service times are negative exponentially distributed with mean μ_i, (ν_i), $i = 1, \ldots, N$ for high (low) priority customers. A new network is created by replacing the priority queue by two servers, one dedicated to the high- and the other to the low-priority customers. These servers have respective mean service rates μ_1 (since high priority customers receive no interference) and $\nu_1(1 - \hat{\rho}_1)$, where $\hat{\rho}_1$ is an approximation for the occupancy of the CPU by high priority customers. The term $1 - \hat{\rho}_1$ is therefore an estimate of the proportion of time the priority queue is free to serve low priority jobs and hence the new service rate is, at least intuitively, a reasonable estimate of the average service rate seen by low priority customers. The equilibrium distribution was then approximated by the corresponding product form for networks with two types of customer (see Baskett, *et al*, (1975), Chandy, Howard and Towsley (1977), Chandy and Martin (1983), Kelly (1981)).

The main problem with the reduced occupancy approximation is that the mean service rate of the low priority priority server is incorrect (shown by Kaufman (1982)). Kaufman found the correct mean, thus introducing a modified reduced occupancy approximation, and showed that for various test networks this procedure provides better approximations for performance measures such as mean response time.

Schmitt (1983) extended this even further by creating a state dependent reduced occupancy approximation. Schmitt's technique involved noticing that the balance equations for the marginal distribution of low priority customers are just the balance equations for an $M/M/1$ queueing system with state dependent service rates. The problem is then to find how the service rates are to be modified for each state. If these new parameters may be found (or at least estimated) then the exact distribution for low priority customers can be obtained. Note, however, that the behaviour of customers given by such models is an approximation. For example, the departure process for low priority customers is not a Poisson stream as would be thought using this approximation. A variation of this technique can also be used to examine systems with a nonpre-emptive discipline. Kuehn and Schmitt (1985) used this method as the basis for the approximation of transit delay distributions in priority queueing networks with arbitrary state independent routing.

Ikehara and Miyazaki (1985), however, did not use any state dependent rates in their approximation. Instead, an asymptotic approximation analysis was used to determine appropriate service parameters for the system. From these, an iterative technique was developed to find utilisation factors for the two virtual servers.

As mentioned earlier, analytic results for equilibrium distributions of priority queueing networks are limited and when it comes to distributions for priority networks, exact results are restricted to only one paper. Morris (1981) considered the closed two node network with two types of customer, a single server at each queue and pre-emptive priority disciplines. Morris found the equilibrium joint queue length distribution for this system as well as the high and low priority throughputs for each customer type. An approximation was also given for the system with nonpre-emptive priority discipline at one node.

Morris also considered a variation of the closed two node queue with priorities in which the priorities are reversed at the two nodes. That is, the high priority customer

at one queue becomes the low priority customer at the other queue. The analysis of this system is very similar to that of an $M/M/1/K$ queue with state dependent arrival and service rates.

In Chapter 6 we shall examine two systems.

The first system to be considered is shown in Figure 6.1, and is analysed in Section 6.1. The process consists of a two node network with single server, pre-emptive priority queues at each node. Unlike Morris' work, we shall not assume that the service rates at each queue are constant. This system may be used to model a computer system consisting of a CPU and an input/output (I/O) device, which processes interactive (high priority) jobs as well as batch (low priority) jobs. In practice, a computer system is required to service jobs coming from people sitting at terminals as well as those which have been placed on batch queues (usually jobs that take a large amount of CPU time, that is, have a long mean service time). Since it is more important to service the requirements of those that are present (to minimise their time spent on a particular chore), their jobs are given higher priority than the batch jobs and hence the motivation for the system to be analysed.

In Section 6.2 the model is altered by employing a nonpre-emptive service discipline at the left node. The equilibrium distribution is then found. A comparison between an approximate and the exact solution for this problem is given in Section 6.3.

The second system is shown in Figure 6.2, and is analysed in Section 6.4. Here, the priorities are reversed, that is, the high priority customers at the left node become low priority customers at the right and vice versa. Again, we shall allow the service rates to be state dependent at each queue.

Section 6.5 then looks at two node closed networks with pre-emptive priority at the left node and allows all disciplines but batch servicing at the right.

For the cases where negative exponential service times are considered, the standard approach is to set up the global balance equations for the network and attempt to solve these. However, for the systems described in Sections 6.1, 6.2 and 6.3, their nature is exploited to give first order nonhomogeneous ordinary difference equations which may be solved via standard techniques, as opposed to the second order ordinary difference equations which result from calling directly on the global balance equations. It is this property which allows the extension of the results of Morris (1981) to systems with state dependent service rates.

CHAPTER 6 : CLOSED TWO NODE PRIORITY QUEUEING NETWORKS

6.1 Two nodes with pre-emptive priorities at each node

6.1.1 Network description

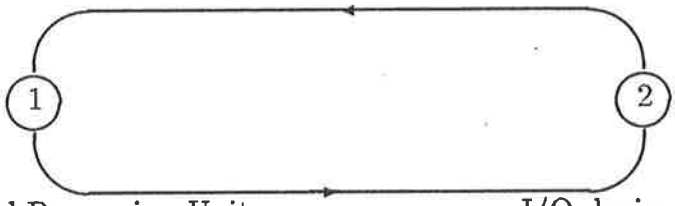
Consider a two node closed network such as that shown in Figure 6.1. There are N high and M low priority customers. We describe the state of the system by (n, m) when there are n high and m low priority customers at queue 1. Note that this is a complete description of the state of the system as customers may neither enter nor leave the network. At node i, $(i = 1, 2)$ let the service time distributions be negative exponential with unit mean, while the service rates are $\mu_i(n, m)$, ($\nu_1(0, m)$ and $\nu_2(N, m)$) for high priority customers (respectively low priority customers) at nodes 1 and 2. Since we are dealing with only single server queues, the low priority customers may only enter service when no high priority are present. Thus, the low priority service rates at each queue are, in effect, only dependent on m. High priority customers take pre-emptive priority over low priority customers at each node. The system is cyclic, that is, upon completion of service, customers move immediately to the other queue.

6.1.2 State independent rates

Morris (1981) considered the case $\mu_i(n, m) = \mu_i$, $\nu_i(n, m) = \nu_i$, $\forall n, m$. The solution is found by setting up the global balance equations for the network and deriving second order homogeneous recurrence relations. These are then solved to find the equilibrium distribution.

In this section an alternative approach using the Markov process result (see Kelly (1979), page 8) that the flux out of any set $\mathcal{A}$ is balanced by the flux into $\mathcal{A}$, that is,

$$\sum_{j \in \mathcal{A}} \sum_{k \in \mathcal{A}^c} p_j q(j, k) = \sum_{k \in \mathcal{A}^c} \sum_{j \in \mathcal{A}} p_k q(k, j), \tag{6.1}$$

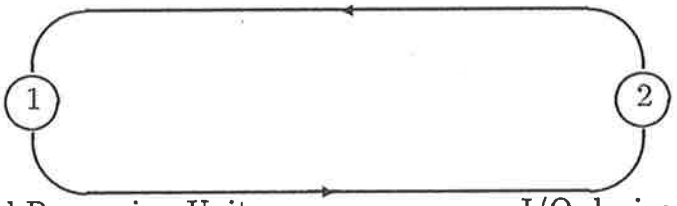


Central Processing Unit
Neg. exp. service time
Pre-emptive priority

I/O device
Neg. exp. service time
Pre-emptive priority

N high priority jobs
M low priority jobs

Figure 6.1. Two node system with pre-emptive priorities at both nodes.



Central Processing Unit
Neg. exp. service time
Nonpre-emptive priority

I/O device
Neg. exp. service time
Pre-emptive priority

N high priority jobs
M low priority jobs

Figure 6.2. Two node system with nonpre-emptive priority at node 1 and pre-emptive priority at node 2.

where p_j is the probability of being in state j, $q(j,k)$ is the rate of moving from state j to state k and $\overline{\mathcal{A}}$ is the complement of $\mathcal{A}$. This has the advantage that it allows the system with state dependent service rates to be solved (Section 6.1.3).

First consider $\mathcal{A} = \{(i,m), 0 \leq i \leq n < N\}$. This corresponds to cut **1** in Figure 6.3. The balance equation generated by a cut of this type is

$$p_{nm}\mu_2 + p_{0m}\nu_1(1-\delta_{m0}) = p_{n+1m}\mu_1 + p_{0m+1}\nu_1(1-\delta_{mM}), \tag{6.2}$$

where δ_{ij} is the Kronecker delta. Rearranging gives

$$p_{n+1m}\mu_1 - p_{nm}\mu_2 = \nu_1\{p_{0m}(1-\delta_{m0}) - p_{0m+1}(1-\delta_{mM})\}. \tag{6.3}$$

This is simply a first order nonhomogeneous recurrence relation in n for p_{nm}, the solution of which is

$$p_{nm} = A_m(\mu_2/\mu_1)^n + \nu_1\{p_{0m}(1-\delta_{m0}) - p_{0m+1}(1-\delta_{mM})\}/(\mu_1-\mu_2), \tag{6.4}$$

where

$$A_m = p_{0m} - \nu_1\{p_{0m}(1-\delta_{m0}) - p_{0m+1}(1-\delta_{mM})\}/(\mu_1-\mu_2). \tag{6.5}$$

Hence,

$$\begin{aligned} p_{nm} =& p_{0m}\{(\mu_2/\mu_1)^n + \nu_1(1-\delta_{m0})(1-(\mu_2/\mu_1)^n)/(\mu_1-\mu_2)\} \\ & - p_{0m+1}\nu_1(1-\delta_{mM})(1-(\mu_2/\mu_1)^n)/(\mu_1-\mu_2). \end{aligned} \tag{6.6}$$

A relationship between p_{0m} and p_{0m+1}, $m = 0,\ldots,M-1$, is required. This is obtained using cut **2**, i.e. $\mathcal{A} = \{(i,j), i = 0,\ldots,N, j = 0,\ldots,m\}$ which gives

$$p_{0m+1}\nu_1 = p_{Nm}\nu_2. \tag{6.7}$$

Substituting for p_{Nm} from (6.6) gives

$$\begin{aligned} p_{0m+1} &= \frac{p_{0m}\{(\mu_2/\mu_1)^N + \nu_1(1-\delta_{m0})(1-(\mu_2/\mu_1)^N)/(\mu_1-\mu_2)\}}{\nu_1/\nu_2 + \nu_1(1-(\mu_2/\mu_1)^N)/(\mu_1-\mu_2)} \\ &= p_{00}H_0H^m, \end{aligned} \tag{6.8}$$

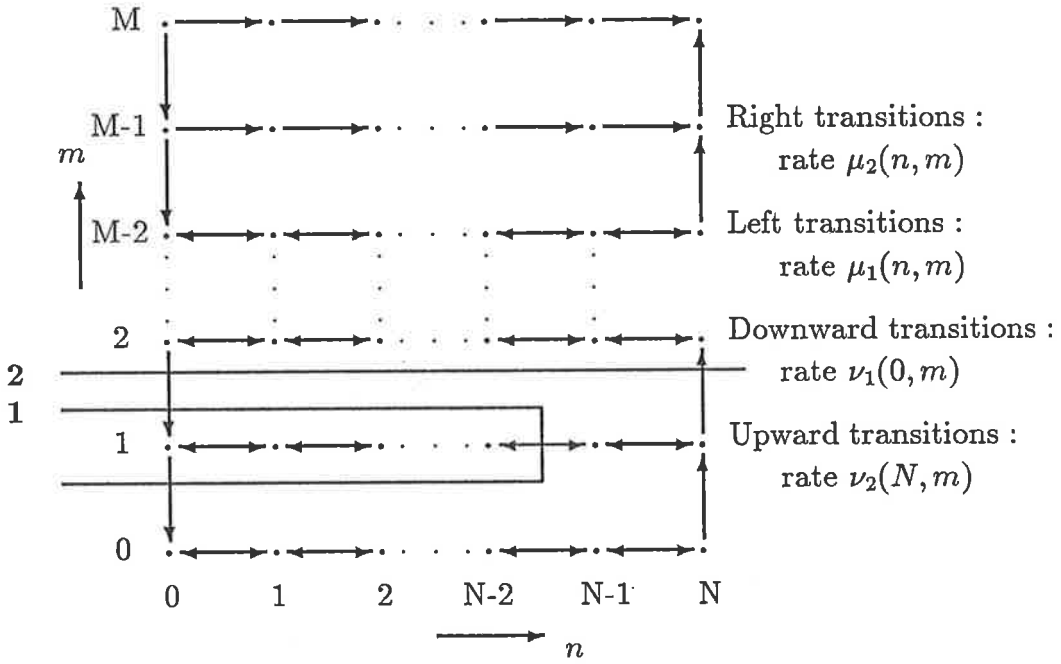


Figure 6.3. State transition diagram for two node pre-emptive priority system.

where

$$H = \frac{(\mu_2/\mu_1)^N + \nu_1(1-\delta_{m0})(1-(\mu_2/\mu_1)^N)/(\mu_1-\mu_2)}{\nu_1/\nu_2 + \nu_1(1-(\mu_2/\mu_1)^N)/(\mu_1-\mu_2)} \tag{6.9}$$

and

$$H_0 = \frac{(\mu_2/\mu_1)^N}{\nu_1/\nu_2 + \nu_1(1-(\mu_2/\mu_1)^N)/(\mu_1-\mu_2)}. \tag{6.10}$$

Substituting into (6.6) gives (for $m = 1, \ldots, M$)

$$p_{nm} = p_{00} H_0 H^{m-1} \left(\left(\frac{\mu_2}{\mu_1} \right)^n + \frac{\nu_1}{\mu_1} \left(1 - H(1-\delta_{mM}) \left(\frac{1-(\frac{\mu_2}{\mu_1})^n}{1-\frac{\mu_2}{\mu_1}} \right) \right) \right) \tag{6.11}$$

and (for $m = 0$)

$$p_{n0} = p_{00} \left(\left(\frac{\mu_2}{\mu_1} \right)^n - \frac{\nu_1 H_0}{\mu_1} \left(\frac{1-(\frac{\mu_2}{\mu_1})^n}{1-\frac{\mu_2}{\mu_1}} \right) \right). \tag{6.12}$$

The solution is completed by finding p_{00} from the normalisation equation

$$\sum_{n=0}^{N} \sum_{m=0}^{M} p_{nm} = 1. \tag{6.13}$$

This gives

$$p_{00} = \left[\left(H_0 \left(\frac{1-H^M}{1-H} \right) + 1 \right) \left(\frac{1-(\frac{\mu_2}{\mu_1})^{N+1}}{1-\frac{\mu_2}{\mu_1}} \right) \right]^{-1}. \tag{6.14}$$

Note that the above equations are discontinuous for $\mu_1 = \mu_2$ or $H = 1$. In these cases we can use the same argument to show that, for appropriately defined x, we replace $(1-x^k)/(1-x)$ by k.

Note that (6.3) is equivalent to equation (6) of Morris (1981) (after substitution of the equation for p_{1m}), but the approach is simpler. The second cut gives the relation between the p_{0m} probabilities immediately, as opposed to solving a second order recurrence relation.

Unlike the form for p_{00} given by Morris (1981), (6.14) is computationally efficient. Also, its form suggests a different method of determining p_{00}. Define $\sum_{m=0}^{M} p_{nm} = p_n$.

to be the marginal distribution for the number of high priority customers at node 1. Making use of the fact that low priority customers do not interfere with those of high priority, $p_{0\cdot}$ is given by the standard solution for a closed two node network with one customer type, i.e.

$$p_{0\cdot} = \left(\frac{1 - (\frac{\mu_2}{\mu_1})^{N+1}}{1 - \frac{\mu_2}{\mu_1}} \right)^{-1}. \tag{6.15}$$

Using $\sum_{m=0}^{M} p_{0m} = p_{0\cdot}$ and substituting for p_{0m} from (6.8) we have

$$\begin{aligned} p_{0\cdot} =& p_{00} + \sum_{m=0}^{N-1} p_{00} H_0 H^m \\ =& p_{00} \left(1 + H_0 \sum_{m=0}^{N-1} H^m \right) \\ =& p_{00} \left(1 + H_0 \left(\frac{1 - H^M}{1 - H} \right) \right), \end{aligned} \tag{6.16}$$

which upon rearranging gives (6.14).

6.1.3 State dependent rates

The procedure developed in Section 6.1.2 can now be employed to analyse the previous model with state dependent service rates.

A single server utilising a pre-emptive priority discipline operates at node i, $(i = 1, 2)$. The service time distribution is negative exponential with unit mean and the service rates are $\mu_i(n, m)$ (respectively $\nu_1(0, m)$ and $\nu_2(N, m)$) for high (low) priority customers when the state is (n, m) (respectively $(0, m)$ and (N, m)).

Theorem 6.1

The network described above has equilibrium distribution $\{p_{nm}, n = 0, \ldots, N, m =$

$0, \ldots, M\}$ given by

(6.17)

$$p_{0m} = p_{00} \prod_{j=0}^{m-1} K(j) \quad m = 1, \ldots, M,$$

$$p_{nm} = p_{00} \prod_{j=0}^{m-1} K(j) \left[\begin{array}{l} \prod_{i=0}^{n-1} \frac{\mu_2(i,m)}{\mu_1(i+1,m)} + \nu_1(0,m)(1-\delta_{m0})G(n,m) \\ - K(m)\nu_1(0,m+1)(1-\delta_{mM})G(n,m) \end{array} \right], \quad n = 1, \ldots, N,$$

where

$$K(m) = \frac{\prod_{j=0}^{N-1} \frac{\mu_2(j,m)}{\mu_1(j+1,m)} + \nu_1(0,m)(1-\delta_{m0})G(N,m)}{\frac{\nu_1(0,m+1)}{\nu_2(N,m)} + \nu_1(0,m+1)G(N,m)} \tag{6.18}$$

and

$$G(n,m) = \frac{1 + \sum_{i=1}^{n-1} \prod_{j=1}^{i} \frac{\mu_2(n-j,m)}{\mu_1(n-j,m)}}{\mu_1(n,m)}, \tag{6.19}$$

with sums over descending ranges taken to be zero, products over descending ranges taken to be unity and p_{00} is a normalising constant.

Proof : Consider the flux equations produced by cuts of type **1**. This gives

(6.20)

$$\begin{aligned} &p_{nm}\mu_2(n,m) + p_{0m}\nu_1(0,m)(1-\delta_{m0}) \\ &\qquad = p_{n+1m}\mu_1(n+1,m) + p_{0m+1}\nu_1(0,m+1)(1-\delta_{mM}) \end{aligned}$$

for $n = 0, \ldots, N$. Rearranging gives

$$\begin{aligned} p_{n+1m} =& \{p_{nm}\mu_2(n,m) + p_{0m}\nu_1(0,m)(1-\delta_{m0}) \\ &- p_{0m+1}\nu_1(0,m+1)(1-\delta_{mM})\}/\mu_1(n+1,m). \end{aligned} \tag{6.21}$$

So for $n = 1, \ldots, N$,

$$\begin{aligned} p_{nm} =& p_{0m} \left[\prod_{i=0}^{n-1} \frac{\mu_2(i,m)}{\mu_1(i+1,m)} + \nu_1(0,m)(1-\delta_{m0})G(n,m) \right] \\ &- p_{0m+1}\nu_1(0,m+1)(1-\delta_{mM})G(n,m). \end{aligned} \tag{6.22}$$

Using cut **2**,

$$p_{0m+1}\nu_1(0,m+1) = p_{Nm}\nu_2(N,m). \tag{6.23}$$

Substituting from (6.22) gives

$$
\begin{aligned}
p_{0m+1} =& p_{0m} \frac{\prod_{i=0}^{N-1} \frac{\mu_2(i,m)}{\mu_1(i+1,m)} + \nu_1(0,m)(1-\delta_{m0})G(N,m)}{\frac{\nu_1(0,m+1)}{\nu_2(N,m)} + \nu_1(0,m+1)G(N,m)} \\
=& p_{0m} K(m) \qquad (6.24) \\
=& p_{00} \prod_{j=0}^{m} K(j),
\end{aligned}
$$

remembering that $\delta_{mM} = 0$ for all m in the required range. Substituting (6.24) into (6.22) completes the proof and p_{00} can be found using the normalisation equation. ■

Remark : As pointed out in Morris (1981), the system with transition rates independent of state can be generalised to the case where, upon completion of service at one queue, the customer moves with probability ρ_{ij} to the other queue, with i the present queue and j the type of customer. The same can be done in this, more general, process, by replacing $\mu_i(n,m)$ by $\mu_i(n,m)\rho_{i1nm}$ and $\nu_i(n,m)$ by $\nu_i(n,m)\rho_{i2nm}$, simply by realising that as the service times are negative exponentially distributed, such a change is just a modification to the service rate.

Rates independent of the number of low priority customers

In the case where $\mu_i(n,m)$ and $\nu_i(n,m)$ are independent of m, (for notational convenience we suppress m in the following), we obtain $p_{n\cdot} = \sum_{m=0}^{M} p_{nm}$ from equation (6.22), the marginal distribution of high priority customers, and find that it agrees with the result for an ordinary two node closed network, i.e.

$$
p_{n\cdot} = \frac{\prod_{j=0}^{n-1} \frac{\mu_2(j)}{\mu_1(j+1)}}{1 + \sum_{i=1}^{N} \prod_{j=0}^{i-1} \frac{\mu_2(j)}{\mu_1(j+1)}}, \qquad (6.25)
$$

with products over descending ranges taken to be unity.

Noting that $K = K(m)$ is independent of m for $m = 1, \ldots, M$, using the technique

of Section 6.1.2, $\sum_{m=0}^{M} p_{0m} = p_{0\cdot}$ and substituting for p_{0m} from (6.24) we have

$$
\begin{aligned}
p_{0\cdot} =& p_{00} + \sum_{m=0}^{M-1} p_{00} K(0) K^m \\
=& p_{00} \left(1 + K(0) \sum_{m=0}^{M-1} K^m \right) \\
=& p_{00} \left(1 + K(0) \left(\frac{1 - K^M}{1 - K} \right) \right).
\end{aligned} \tag{6.26}
$$

Rearranging gives

$$
p_{00} = \left[\left(1 + K(0) \left(\frac{1 - K^M}{1 - K} \right) \right) \left(1 + \sum_{i=1}^{N} \prod_{j=0}^{i-1} \frac{\mu_2(j)}{\mu_1(j+1)} \right) \right]^{-1}. \tag{6.27}
$$

State dependent service rates may be useful in analysing systems where extra facilities may be called upon when demand for service is high.

Examples : (1) A system with N processors (each with service rate μ_1) dedicated to high priority jobs and a single processor for low priority jobs at queue 1 and only a single server at queue 2 would have service rates

$$
\begin{aligned}
\mu_1(n,m) &= n\mu_1, \\
\mu_2(n,m) &= \mu_2, \\
\nu_1(0,m) &= \nu_1, \\
\nu_2(N,m) &= \nu_2.
\end{aligned}
$$

(2) A modified version of Example (1) with $K < N$ processors for high priority jobs at queue 1 would have service rates

$$
\begin{aligned}
\mu_1(n,m) &= \begin{cases} n\mu_1 & n = 1, \ldots, K \\ K\mu_1 & n = K, \ldots, N, \end{cases} \\
\mu_2(n,m) &= \mu_2, \\
\nu_1(0,m) &= \nu_1, \\
\nu_2(N,m) &= \nu_2.
\end{aligned}
$$

(3) In a similar fashion, systems with multiple servers at both facilities for high priority jobs may be handled by simply modifying the high priority service rate.

The throughputs of high and low priority customers (see Schmitt (1983)), T_H and T_L respectively, at the left node may be computed from

$$T_H = \sum_{n=1}^{N} \sum_{m=0}^{M} \mu_1(n,m) p_{nm} \tag{6.28}$$

and

$$T_L = \sum_{m=1}^{M} \nu_1(0,m) p_{0m}. \tag{6.29}$$

The mean delays may be computed using Little's formula

$$L = \lambda W \tag{6.30}$$

where L is the mean queue length, λ the mean arrival rate and W the mean waiting time.

With appropriate modification we can use this technique to find the equilibrium distribution for a system with more than two priority classes and state dependent service rates.

The systems described above can be analysed using the quasi birth and death structure of Gaver, Jacobs and Latouche (1984). However, the use of closed form solutions is much more efficient, computationally, than the matrix techniques that they use and hence should be preferred for such processes.

6.2 Two nodes with non-pre-emptive priority at the left node

The previous model may be modified such that the priority at queue 1 is non pre-emptive, i.e. the arrival of a high priority customer does not affect the customer in service. The state is defined as (n, m, S) when there are n high and m low priority customers and there is a type S ($S = L$ or H) customer in service at the left node.

For notational convenience we shall denote the empty state ($n = m = 0$) by $(0, 0, L)$. Let the service time distributions be negative exponential with parameters μ_i and ν_i for high and low priority customers respectively at queue i, $(i = 1, 2)$.

A method for determining the equilibrium distribution for the above system is now given. For notational ease, define

$$
\begin{aligned}
a &= \mu_2/\mu_1, \\
b &= \mu_2/(\mu_2 + \nu_1), \\
c &= (\mu_2 + \nu_1)/\mu_1, \\
d &= \nu_2/(\nu_1 + \nu_2), \\
e &= \mu_2/(\nu_1 + \nu_2),
\end{aligned} \tag{6.31}
$$

$$
f_N = (\mu_1 + \nu_2)c \sum_{i=0}^{N-1} a^i - \mu_2 c \sum_{i=0}^{N-2} a^i - \nu_1 b^{N-1} ed + \nu_2 H(N)
$$

and

$$
H(n) = \frac{\mu_2 + \nu_1}{\mu_1} \sum_{i=0}^{n-1} \left(\frac{\mu_2}{\mu_1}\right)^i \left(1 - \left(\frac{\mu_2}{\mu_2 + \nu_1}\right)^{n-i}\right). \tag{6.32}
$$

$H(n)$ may be simplified by expanding the sum. This gives

(6.33)

$$
\begin{aligned}
H(n) &= \left(\frac{\mu_2 + \nu_1}{\mu_1}\right)\left(\frac{1 - \left(\frac{\mu_2}{\mu_1}\right)^n}{1 - \frac{\mu_2}{\mu_1}}\right) - \left(\frac{\mu_2 + \nu_1}{\mu_1}\right)\sum_{i=0}^{n-1}\left(\frac{\mu_2}{\mu_2 + \nu_1}\right)^{n-i}\left(\frac{\mu_2}{\mu_1}\right)^i \\
&= (\mu_2 + \nu_1)\left(\frac{1 - \left(\frac{\mu_2}{\mu_1}\right)^n}{\mu_1 - \mu_2}\right) - \left(\frac{\mu_2 + \nu_1}{\mu_1}\right)\left(\frac{\mu_2}{\mu_2 + \nu_1}\right)^n \sum_{i=0}^{n-1}\left(\frac{\mu_2}{\mu_1}\right)^i \left(\frac{\mu_2 + \nu_1}{\mu_2}\right)^i \\
&= (\mu_2 + \nu_1)\left(\frac{1 - \left(\frac{\mu_2}{\mu_1}\right)^n}{\mu_1 - \mu_2}\right) - \left(\frac{\mu_2 + \nu_1}{\mu_1}\right)\left(\frac{\mu_2}{\mu_2 + \nu_1}\right)^n \left(\frac{1 - \left(\frac{\mu_2 + \nu_1}{\mu_1}\right)^n}{1 - \left(\frac{\mu_2 + \nu_1}{\mu_1}\right)}\right) \\
&= (\mu_2 + \nu_1)\left[\left(\frac{1 - \left(\frac{\mu_2}{\mu_1}\right)^n}{\mu_1 - \mu_2}\right) - \left(\frac{\mu_2}{\mu_2 + \nu_1}\right)^n \left(\frac{1 - \left(\frac{\mu_2 + \nu_1}{\mu_1}\right)^n}{\mu_1 - \mu_2 - \nu_1}\right)\right] \\
&= (\mu_2 + \nu_1)\left[\left(\frac{1 - \left(\frac{\mu_2}{\mu_1}\right)^n}{\mu_1 - \mu_2}\right) - \frac{\left(\frac{\mu_2}{\mu_2 + \nu_1}\right)^n - \left(\frac{\mu_2}{\mu_1}\right)^n}{\mu_1 - \mu_2 - \nu_1}\right].
\end{aligned}
$$

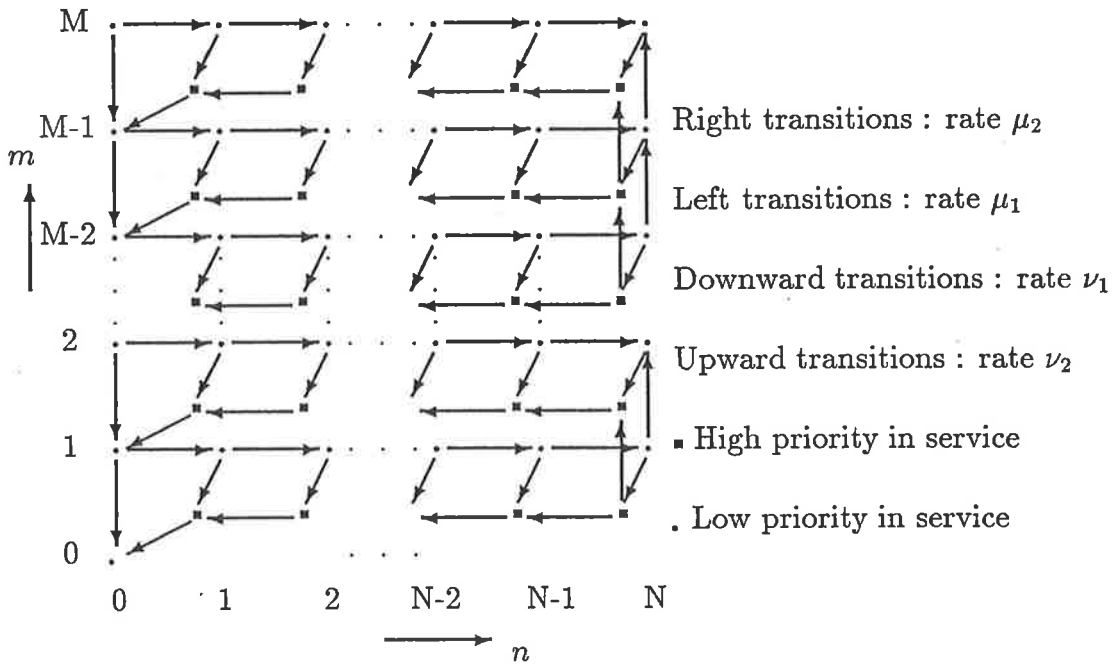


Figure 6.4. State transition diagram for two node system, nonpre-emptive priority at left node, pre-emptive priority at right node.

Theorem 6.2

For the network described above, the equilibrium distribution may be found from

$$p_{01L} = p_{00L}\frac{a^N(\mu_1+\nu_2)-\mu_2 a^{N-1}}{H(N)(\mu_1+\nu_2)-H(N-1)\mu_2+\nu_1 b^{N-1}e}, \tag{6.34}$$

$$p_{02L} = \frac{-p_{00L}\nu_2 a^N + p_{01L}f_N}{H(N)(\mu_1+\nu_2)-H(N-1)\mu_2+\nu_1 b^{N-1}e} \tag{6.35}$$

and for $m = 3,\ldots,M-1$,

$$p_{0mL} = \frac{-p_{0m-2L}\nu_2 c\sum_{i=0}^{N-1}a^i - \nu_1 b^{N-1}e\sum_{i=1}^{m-1}d^{m-i}p_{0iL} + p_{0m-1L}f_N}{H(N)(\mu_1+\nu_2)-H(N-1)\mu_2+\nu_1 b^{N-1}e}, \tag{6.36}$$

$$p_{0ML} = \frac{-p_{0M-2L}\nu_2 c\sum_{i=0}^{N-1}a^i - \nu_2 b^{N-1}e\sum_{i=1}^{M-1}d^{M-i-1}p_{0iL} + p_{0M-1L}f_N}{H(N)(\mu_1+\nu_2)-H(N-1)\mu_2+\mu_2 b^{N-1}}, \tag{6.37}$$

$$p_{nmL} = \left(\frac{\mu_2}{\mu_2+\nu_1}\right)^n p_{0mL}, \quad n = 0,\ldots,N-1, m = 1,\ldots,M, \tag{6.38}$$

(6.39)

$$p_{NmL} = \left(\frac{\mu_2}{\mu_2+\nu_1}\right)^{N-1}\left(\frac{\mu_2}{\nu_1+\nu_2}\right)\sum_{i=1}^{m}\left(\frac{\nu_2}{\nu_1+\nu_2}\right)^{m-i}p_{0iL}, \quad m = 1,\ldots,M-1,$$

(6.40)

$$\begin{aligned} p_{NML} =& \frac{\nu_2}{\nu_1}\left(\frac{\mu_2}{\mu_2+\nu_1}\right)^{N-1}\left(\frac{\mu_2}{\nu_1+\nu_2}\right)\sum_{i=1}^{M-1}\left(\frac{\nu_2}{\nu_1+\nu_2}\right)^{M-i-1}p_{0iL} \\ &+ \frac{\mu_2}{\nu_1}\left(\frac{\mu_2}{\mu_2+\nu_1}\right)^{N-1}p_{0ML}. \end{aligned}$$

This gives all of the probabilities when a low priority customer is in service. The relationships for a high priority customer being in service are given by

(6.41)

$$p_{nmH} = p_{0mL}\frac{\mu_2+\nu_1}{\mu_1}\sum_{i=0}^{n-1}\left(\frac{\mu_2}{\mu_1}\right)^i - p_{0m+1L}H(n), \quad m = 1,\ldots,M-1, n = 1,\ldots,N,$$

$$p_{n0H} = p_{00L}\left(\frac{\mu_2}{\mu_1}\right)^n - p_{01L}H(n), \quad n = 1,\ldots,N \tag{6.42}$$

and

$$p_{nMH} = p_{0ML}\left(\frac{\mu_2+\nu_1}{\mu_1}\right)\left(\frac{1-\left(\frac{\mu_2}{\mu_1}\right)^n}{1-\frac{\mu_2}{\mu_1}}\right), \quad n = 1,\ldots,N, \tag{6.43}$$

where p_{00L} is a normalising constant.

Proof : Consider the global balance equations for states with low priority customers in service. Then

$$p_{nmL}(\mu_2+\nu_1) = p_{n-1mL}\mu_2 \quad n = 1,\ldots,N-1, m = 1,\ldots,M, \tag{6.44}$$

which has solution

$$p_{nmL} = \left(\frac{\mu_2}{\mu_2+\nu_1}\right)^n p_{0mL}, \quad n = 0,\ldots,N-1, m = 1,\ldots,M. \tag{6.45}$$

This is equation (6.38). Also,

$$p_{N1L}(\nu_1+\nu_2) = p_{N-11L}\mu_2 \tag{6.46}$$

and

$$p_{NmL}(\nu_1+\nu_2) = p_{N-1mL}\mu_2 + p_{Nm-1L}\nu_2, \quad m = 2,\ldots,M-1. \tag{6.47}$$

Solving for p_{NmL} and repeated substitution of equations (6.45) and (6.47) gives

$$\begin{aligned} p_{NmL} =& \frac{\mu_2}{\nu_1+\nu_2}p_{N-1mL} + \frac{\nu_2}{\nu_1+\nu_2}p_{Nm-1L} \\ =& \frac{\mu_2}{\nu_1+\nu_2}\left(\frac{\mu_2}{\mu_2+\nu_1}\right)^{N-1} p_{0mL} \\ &+ \frac{\nu_2}{\nu_1+\nu_2}\left[\frac{\mu_2}{\nu_1+\nu_2}p_{N-1m-1L} + \frac{\nu_2}{\nu_1+\nu_2}p_{Nm-2L}\right] \\ =& \left(\frac{\mu_2}{\mu_2+\nu_1}\right)^{N-1}\left(\frac{\mu_2}{\nu_1+\nu_2}\right)\sum_{i=0}^{m-2}\left(\frac{\nu_2}{\nu_1+\nu_2}\right)^i p_{0m-iL} \\ &+ \left(\frac{\nu_2}{\nu_1+\nu_2}\right)^{m-1} p_{N1L}. \end{aligned} \tag{6.48}$$

Substituting for p_{N1L} from equation (6.46) and $p_{N-1,1,L}$ from equation (6.45) gives

(6.49)
$$p_{NmL} = \left(\frac{\mu_2}{\mu_2+\nu_1}\right)^{N-1}\left(\frac{\mu_2}{\nu_1+\nu_2}\right)\sum_{i=1}^{m}\left(\frac{\nu_2}{\nu_1+\nu_2}\right)^{m-i} p_{0iL}, \quad m=1,\ldots,M-1,$$

which is equation (6.39).

We find p_{NML} from

(6.50)
$$p_{NML}\nu_1 = p_{N-1ML}\mu_2 + p_{NM-1L}\nu_2.$$

This gives

(6.51)
$$\begin{aligned} p_{NML} =& \frac{\nu_2}{\nu_1}\left(\frac{\mu_2}{\mu_2+\nu_1}\right)^{N-1}\left(\frac{\mu_2}{\nu_1+\nu_2}\right)\sum_{i=1}^{M-1}\left(\frac{\nu_2}{\nu_1+\nu_2}\right)^{M-i-1} p_{0iL} \\ &+ \frac{\mu_2}{\nu_1}\left(\frac{\mu_2}{\mu_2+\nu_1}\right)^{N-1} p_{0ML}, \end{aligned}$$

which is equation (6.40). This gives all states with low priority customers in service in terms of the p_{0mL}, $m = 1,\ldots,M$. Now consider the flux into and out of $\mathcal{A} = \{(i,m,H), i=1,\ldots,n\} \cup \{(0,m,L)\}, m=1,\ldots,M-1$. Then,

(6.52)
$$p_{nmH}\mu_2 + p_{0mL}(\mu_2+\nu_1) = \sum_{i=0}^{n} p_{im+1L}\nu_1 + p_{n+1mH}\mu_1, \quad n=1\ldots N-1.$$

Rearranging and substituting for $p_{im+1L}, i=0,\ldots,n$,

(6.53)
$$\mu_1 p_{n+1mH} = p_{nmH}\mu_2 + (\mu_2+\nu_1)p_{0mL} - (\mu_2+\nu_1)\left[1-\left(\frac{\mu_2}{\mu_2+\nu_1}\right)^{n+1}\right] p_{0m+1L}$$

and repeated substitution for p_{nmH} gives

(6.54)
$$\begin{aligned} p_{n+1mH} =& p_{0mL}\frac{\mu_2+\nu_1}{\mu_1}\sum_{i=0}^{n-1}\left(\frac{\mu_2}{\mu_1}\right)^i + p_{1mH}\left(\frac{\mu_2}{\mu_1}\right)^n \\ &- p_{0m+1L}\frac{\mu_2+\nu_1}{\mu_1}\sum_{i=0}^{n-1}\left(1-\left(\frac{\mu_2}{\mu_2+\nu_1}\right)^{n-i+1}\right)\left(\frac{\mu_2}{\mu_1}\right)^i. \end{aligned}$$

To find p_{1mH} use the global balance equation centred on (0,m,L).

(6.55)
$$p_{0mL}(\mu_2+\nu_1) = p_{1mH}\mu_1 + p_{0m+1L}\nu_1.$$

Substituting for p_{1mH} in (6.55) gives

(6.56)

$$p_{nmH} = p_{0mL}\frac{\mu_2+\nu_1}{\mu_1}\sum_{i=0}^{n-1}\left(\frac{\mu_2}{\mu_1}\right)^i - p_{0m+1L}H(n), \quad m=1,\ldots,M-1, n=1,\ldots,N,$$

thus deriving equation (6.41). Now let $\mathcal{A}=\{(i,0,H), i=1,\ldots,n\}\cup\{(0,0,L)\}$. Then

$$p_{n0H}\mu_2 = p_{n+10H}\mu_1 + \sum_{i=0}^{n} p_{i1L}\nu_1, \quad n=1,\ldots,N-1, \tag{6.57}$$

which on substituting for p_{i1L} from equation (6.45) and repeatedly substituting for equation (6.57) gives

$$p_{n0H} = p_{10H}\left(\frac{\mu_2}{\mu_1}\right)^{n-1} - p_{01L}\frac{\mu_2+\nu_1}{\mu_1}\sum_{i=0}^{n-2}\left(1-\left(\frac{\mu_2}{\mu_2+\nu_1}\right)^{n-i}\right)\left(\frac{\mu_2}{\mu_1}\right)^i. \tag{6.58}$$

To find p_{10H} we use the global balance equation centred on (0,0,L),

$$p_{00L}\mu_2 = p_{10H}\mu_1 + p_{01L}\nu_1. \tag{6.59}$$

Substituting for p_{10H} from equation (6.59) into equation (6.58) gives

$$p_{n0H} = p_{00L}\left(\frac{\mu_2}{\mu_1}\right)^n - p_{01L}H(n), \tag{6.60}$$

which is equation (6.42).

Now use $\mathcal{A}=\{(i,M,H), i=1,\ldots,n\}\cup\{(0,M,L)\}$. This gives,

$$p_{n+1MH}\mu_1 = p_{nMH}\mu_2 + p_{0ML}(\mu_2+\nu_1), \quad n=1,\ldots,N-1, \tag{6.61}$$

which is a first order nonhomogeneous recurrence relation with solution

$$p_{nMH} = p_{1MH}\left(\frac{\mu_2}{\mu_1}\right)^{n-1} + p_{0ML}\left(\frac{\mu_2+\nu_1}{\mu_1}\right)\sum_{i=0}^{n-2}\left(\frac{\mu_2}{\mu_1}\right)^i. \tag{6.62}$$

Also

$$p_{0ML}(\mu_2+\nu_1) = p_{1MH}\mu_1. \tag{6.63}$$

Substituting for p_{1MH} from equation (6.63) into equation (6.62) gives

$$p_{nMH} = p_{0ML}\left(\frac{\mu_2+\nu_1}{\mu_1}\right)\left(\frac{1-\left(\frac{\mu_2}{\mu_1}\right)^n}{1-\frac{\mu_2}{\mu_1}}\right) \quad n=1,\ldots,N, \tag{6.64}$$

which is equation (6.43). So all of the probabilities of states with a high priority customer in service are in terms of the p_{0mL} probabilities.

A relationship is required between the p_{0mL} probabilities. This is obtained from the global balance equations centred on (N,m,H), $m=0,\ldots,M-1$. For $m=0$,

$$p_{N0H}(\mu_1+\nu_2) = p_{N-10H}\mu_2 + p_{N1L}\nu_1. \tag{6.65}$$

Substituting for these three terms from equations (6.39), (6.41) and (6.42), gives

$$\begin{aligned}(\mu_1+\nu_2)&\left(p_{00L}\left(\frac{\mu_2}{\mu_1}\right)^N - p_{01L}H(N)\right)\\ =&\mu_2\left(p_{00L}\left(\frac{\mu_2}{\mu_1}\right)^{N-1} - p_{01L}H(N-1)\right)\\ &+\nu_1\left(\frac{\mu_2}{\mu_2+\nu_1}\right)^{N-1}\left(\frac{\mu_2}{\nu_1+\nu_2}\right)p_{01L}.\end{aligned} \tag{6.66}$$

Rearranging, we obtain

$$p_{01L} = p_{00L}\frac{a^N(\mu_1+\nu_2)-\mu_2 a^{N-1}}{H(N)(\mu_1+\nu_2)-H(N-1)\mu_2+\nu_1 b^{N-1}e}, \tag{6.67}$$

thus establishing (6.34).

Using the global balance equation centred on (N,m,H), for $m=1,\ldots,M-1$,

$$p_{NmH}(\nu_2+\mu_1) = p_{N-1mH}\mu_2 + p_{Nm+1L}\nu_1 + p_{Nm-1H}\nu_2. \tag{6.68}$$

For $m=1$, substituting for each term from equations (6.39), (6.41) and (6.42)

and rearranging gives

$$
\begin{aligned}
&(\mu_1+\nu_2)\left(p_{01L}\left(\frac{\mu_2+\nu_1}{\mu_1}\right)\sum_{i=0}^{N-1}\left(\frac{\mu_2}{\mu_1}\right)^i - p_{02L}H(N)\right)\\
&\quad=\mu_2\left(p_{01L}\left(\frac{\mu_2+\nu_1}{\mu_1}\right)\sum_{i=0}^{N-2}\left(\frac{\mu_2}{\mu_1}\right)^i - p_{02L}H(N-1)\right)\\
&\quad+\nu_1\left(\frac{\mu_2}{\mu_2+\nu_1}\right)^{N-1}\left(\frac{\mu_2}{\nu_1+\nu_2}\right)\sum_{i=1}^{2}\left(\frac{\nu_2}{\nu_1+\nu_2}\right)^{2-i}p_{0iL}\\
&\quad+\nu_2 p_{00L}\left(\frac{\mu_2}{\mu_1}\right)^N - p_{01L}H(N).
\end{aligned}
\tag{6.69}
$$

Rearranging gives,

$$
p_{02L}=\frac{-p_{00L}\nu_2 a^N + p_{01L}f_N}{H(N)(\mu_1+\nu_2)-H(N-1)\mu_2+\nu_1 b^{N-1}e},
\tag{6.70}
$$

which is equation (6.35).

For $m=3,\ldots,M-1$, substituting from equations (6.39) and (6.41) into equation (6.68) and rearranging gives

$$
p_{0mL}=\frac{-p_{0m-2L}\nu_2 c\sum_{i=0}^{N-1}a^i-\nu_1 b^{N-1}e\sum_{i=1}^{m-1}d^{m-i}p_{0iL}+p_{0m-1L}f_N}{H(N)(\mu_1+\nu_2)-H(N-1)\mu_2+\nu_1 b^{N-1}e},
\tag{6.71}
$$

which is equation (6.36), while for $m=M$ we obtain

$$
p_{0ML}=\frac{-p_{0M-2L}\nu_2 c\sum_{i=0}^{N-1}a^i-\nu_2 b^{N-1}e\sum_{i=1}^{M-1}d^{M-i-1}p_{0iL}+p_{0M-1L}f_N}{H(N)(\mu_1+\nu_2)-H(N-1)\mu_2+\mu_2 b^{N-1}}.
\tag{6.72}
$$

which is equation (6.37). Setting $p_{00L}=1$ we are able find the relative values of all the probabilities. Normalising then gives the equilibrium distribution. ■

We have not derived a closed form expression for the equilibrium distribution because of the excessive complexity of such a solution. The above form, however, is amenable to computation, aspects of which are discussed in Section 6.3.

These results may be generalised to the cases where service rates are state dependent or customers may recommence service. Exactly the same method is employed,

but note that the degree of complexity increases as in the case with pre-emptive priorities. Hence it may be more economical in this case to use an approximation method to obtain performance measures of the system.

The throughputs may be computed according to

$$T_H = \mu_1 \sum_{n=1}^{N} \sum_{m=0}^{M} p_{nmH},$$
$$T_L = \nu_1 \sum_{n=0}^{N} \sum_{m=1}^{M} p_{nmL} \tag{6.73}$$

Again, the mean delays may be computed using Little's result (6.30).

6.3 Comparison with an approximate solution

Morris (1981) proposed an approximate solution to the system considered in Section 6.2. This approximate solution involves denial of service to low priority customers at queue 2 when $n = N$ and a low priority customer was in service at queue 1, noting that this would underestimate low priority throughput and overestimate the high priority throughput. Based on the work in Section 6.2, numerical results indicate that this is not always the case even though this seems plausible intuitively. Apart from a small range of values for ν_1 (in particular, $\nu_1 \ll 1$) the relative error of the approximate method for low priority throughput is very small. Thus the basis of Morris' approximate technique (namely, the probability is small that N high priority customers and one low priority customer can be served at queue 2 in less time than it takes to serve one low priority customer at queue 1) appears well justified. Comparison of utilisations given by both methods support this viewpoint. Figures 6.5, 6.6 and 6.7 give comparisons of some exact throughputs with the approximate model.

An interesting feature to note is that in some cases it is possible to significantly increase the low priority throughput by decreasing the service rate of low priority customers. This may be useful in some design problems where changes may be made

to alter the service rate of low priority customers. In Figure 6.8 a plot is given of low priority throughput as a function of both μ_1 and ν_1 for $\mu_2 = \nu_2 = 1$. Note that in this plot the low priority throughput is, except for ν_1 small, maximised near $\mu_1 = \mu_2 = 1$. For $\mu_1 \ll \mu_2$, high priority customers are bottlenecked at queue 1, thereby lowering the throughput of low priority customers. A similar effect is observed at queue 2 if $\mu_2 \ll \mu_1$. The exception for ν_1 small is because high priority customers are unable to interrupt a low priority customer in service at the queue 1.

It should be noted, however, that while the throughput may be increased by lowering the service rate, this leads to a drastic increase in the mean queue lengths of both types of customer at the queue 1. Table 6.1 gives a comparison of low priority throughputs and mean queue lengths for some typical parameters. This indicates that in certain circumstances throughput is not a very good parameter for describing system behaviour. Similarly, mean queue lengths are not always good indicators because they do not give any indication of the degree of variability of queue length.

Morris introduced the approximate solution in an effort to lower the degree of complexity of the nonpre-emptive system. However, testing has shown little difference in the amount of CPU time required for the approximate and exact methods (using a VAX 11/780 computer). Table 6.2 gives a comparison of CPU time requirements for a variety of different parameters. Examination of Table 6.2 also reveals that the CPU time requirement of the approximate solution is strongly dependent on the number of low priority customers.

On the other hand, for the case of state dependent service rates it would seem that use of an approximation is well justified. However, the complexity in obtaining the exact solution would hardly appear worthwhile in view of the simplicity of Morris' approach.

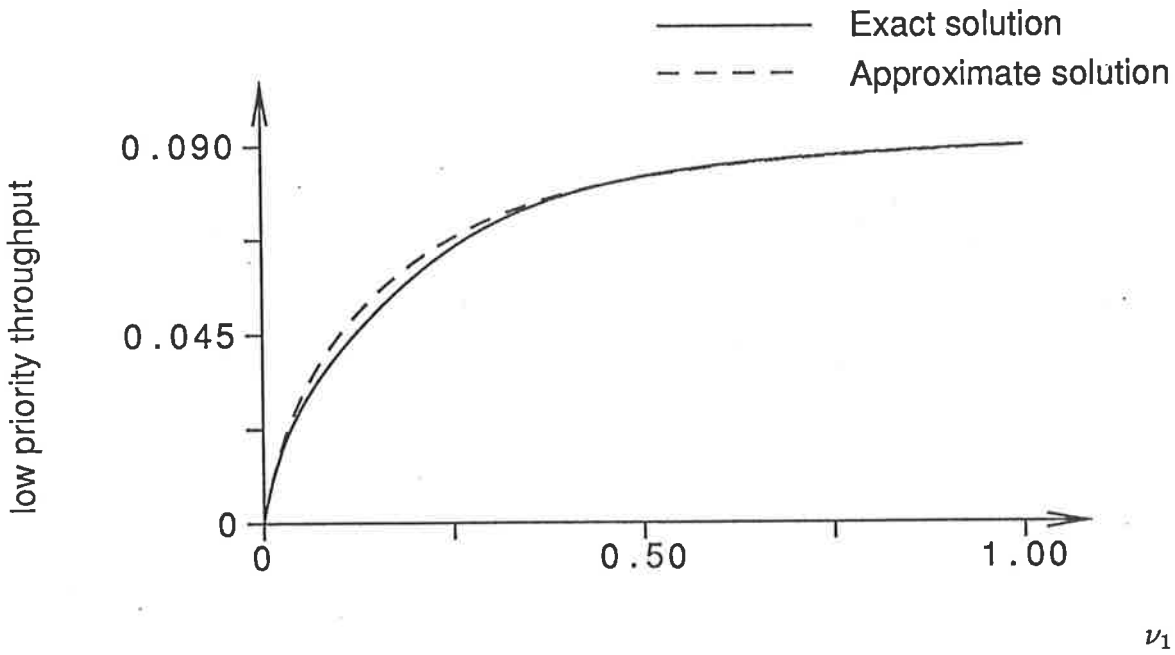


Figure 6.5. Low priority throughput v. ν_1 with 5 high and 5 low priority customers, $\mu_1 = 1$, $\mu_2 = 1$ and $\nu_2 = 1$.

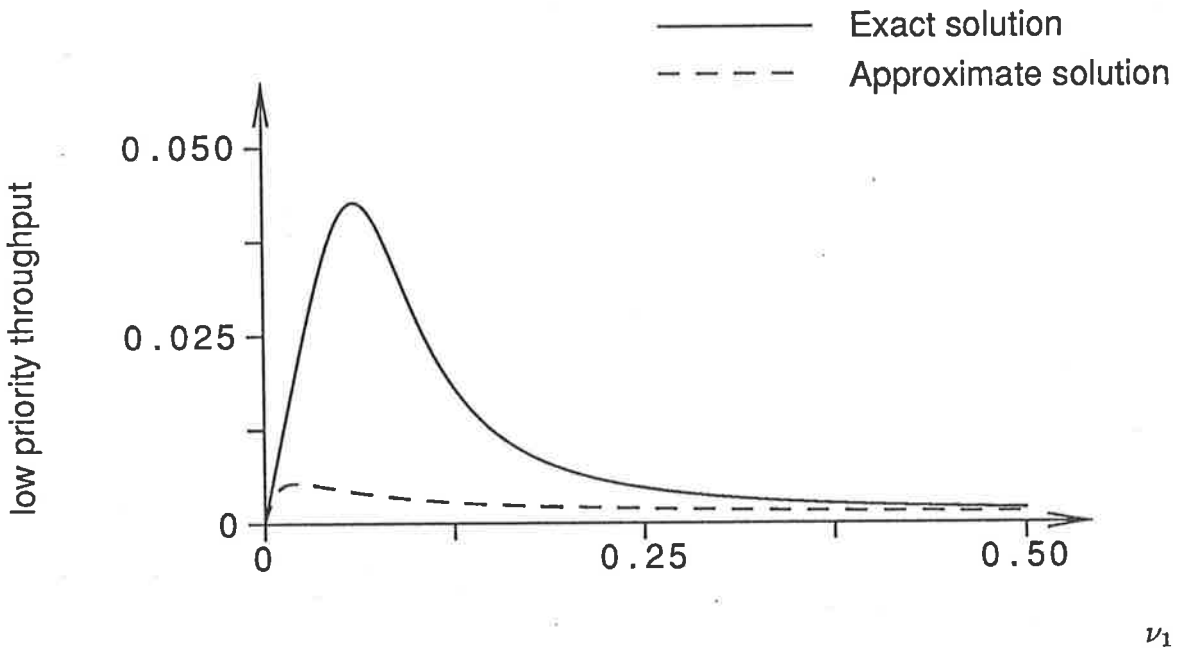


Figure 6.6. Low priority throughput v. ν_1 with 5 high and 5 low priority customers, $\mu_1 = 5$, $\mu_2 = 1$ and $\nu_2 = 5$.

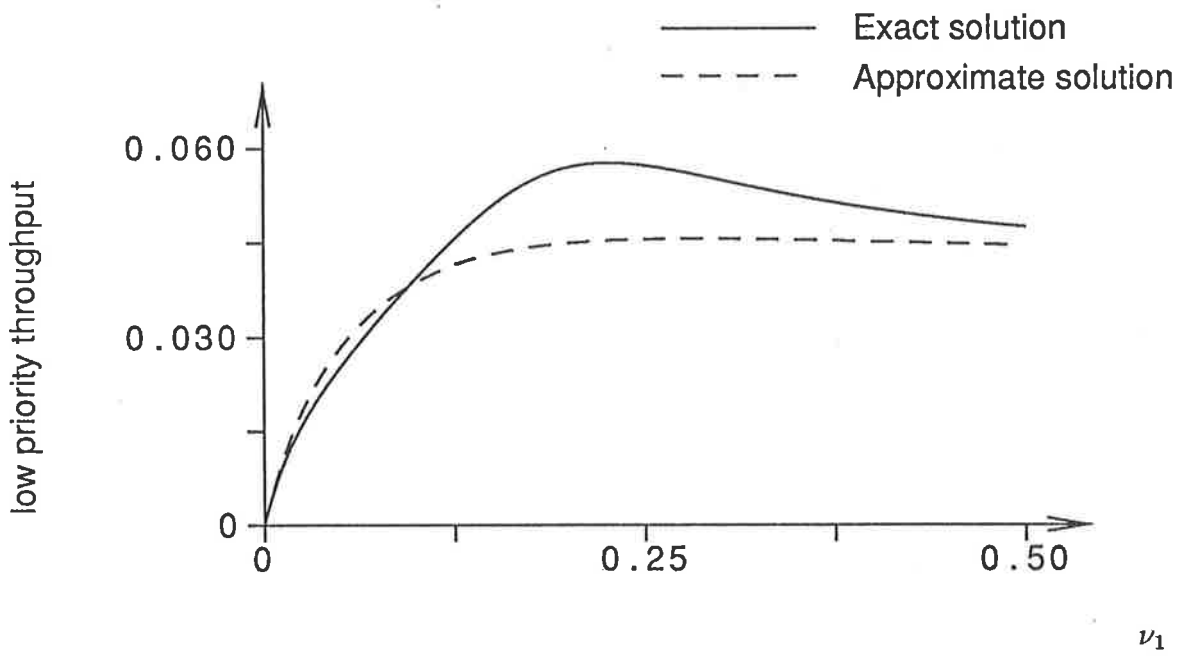


Figure 6.7. Low priority throughput v. ν_1 with 5 high and 10 low priority customers, $\mu_1 = 1$, $\mu_2 = 1$ and $\nu_2 = 0.25$.

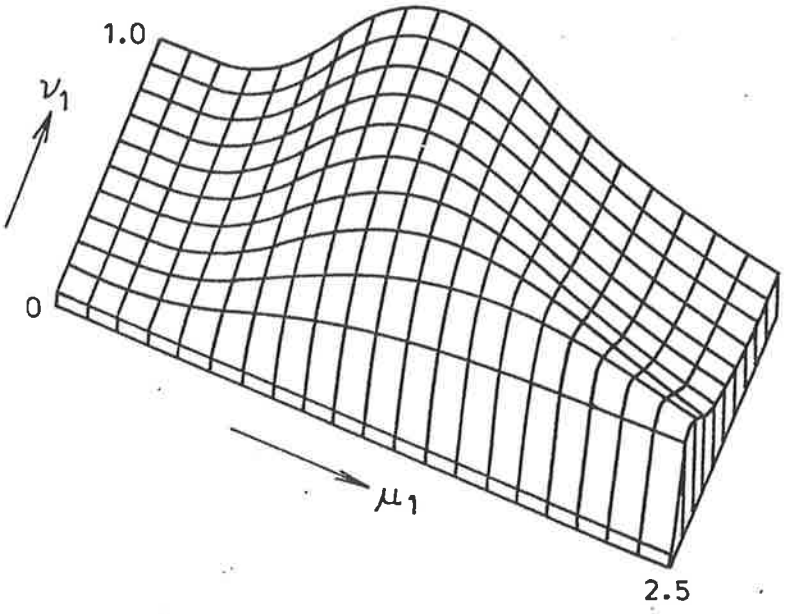


Figure 6.8. Low priority throughput as a function of ν_1 and μ_1.

ν_1	Throughput (×100)	Mean queue length
1.0	1.647	0.1116
0.5	2.094	0.2230
0.25	3.820	0.8532
0.10	6.016	3.683

Table 6.1 Low priority throughput and mean queue length, $\mu_1 = 2, \mu_1 = \nu_2 = 1$.

(N,M)	*CPU time used (seconds)*	
	Exact method	Approximation
(5,5)	0.01	0.01
(5,20)	0.04	0.06
(10,5)	0.02	0.02
(10,20)	0.06	0.07
(20,5)	0.06	0.03
(20,20)	0.10	0.12
(40,5)	0.22	0.06
(40,20)	0.25	0.19

Table 6.2 CPU time usages of exact and approximate solutions.

6.4 Two nodes with priorities reversed at each node

Consider a closed two node network with pre-emptive priorities reversed at the two nodes, that is, type 1 customers are of high priority at queue 1 and low priority at queue two. Let the state of the system be (n, m) when there are n type 1 customers and m type 2 customers at queue 1. Upon completion of service a customer moves immediately to the other queue. There are N type 1 and M type 2 customers. The state transition diagram is given in Figure 6.9.

Morris (1981) considered the case where all service times were distributed negative exponentially with parameters μ_i and ν_i for high and low priority customers respectively at queue i, $(i = 1, 2)$, and the service rates were independent of the state of the system (without loss of generality, unit rate). Morris observed that the only recurrent states of this process are $(0, m)$, $m = 0, \ldots, M$, and (n, M), $n = 0, \ldots, N$, and that

the equilibrium distribution is given by

$$p_{nm} = \begin{cases} C(\frac{\nu_2}{\nu_1})^m & n = 0, m = 0, \ldots, M \\ C(\frac{\nu_2}{\nu_1})^M(\frac{\mu_2}{\mu_1})^n & n = 0, \ldots, N, m = M \\ 0 & \text{otherwise} \end{cases} \tag{6.74}$$

where C is a normalising constant.

In equilibrium, the flow of customers around the network has an interesting feature. Suppose the system is in state (N, M), that is, all customers are at queue 1. A type 1 customer is then in service and upon completing service moves to queue 2. As they are the only customer there, they move directly into service, eventually completing the cycle and returning to queue 1. Thus, the type 2 customers only get into service at queue 1 when all the type 1 customers are at queue 2 (that is, when the process reaches state $(0, M)$). If a type 2 customer then completes service before the next type 1 arrival, they move directly into service at queue 2 (as they are now high priority customers). Hence the system alternates amongst three phases;

(1) Type 1 customers being served at both queues with all type 2 customers at queue 1.

(2) All type 1 customers are at queue 2 while all type 2 customers are at queue 1.

(3) Type 2 customers being served at both queues with all type 1 customers at queue 2.

These phases are given schematically in Figure 6.10. As a result, we may completely describe the state of the system (in equilibrium) by the total number of customers at queue 1. We shall adopt this convention.

Let the process begin in one of the recurrent states. Looking at Figure 6.10, it is easy to see that a general queueing discipline may be constructed in which type 2 customers only occupy positions $1, \ldots, M$ at queue 1 and $N+1, \ldots, M+N$ at queue 2, with the convention that only the highest priority customers present are served. Similarly, type 1 customers only occupy positions $1, \ldots, N$ at queue 2 and $M+1, \ldots, M+N$ at queue 1.

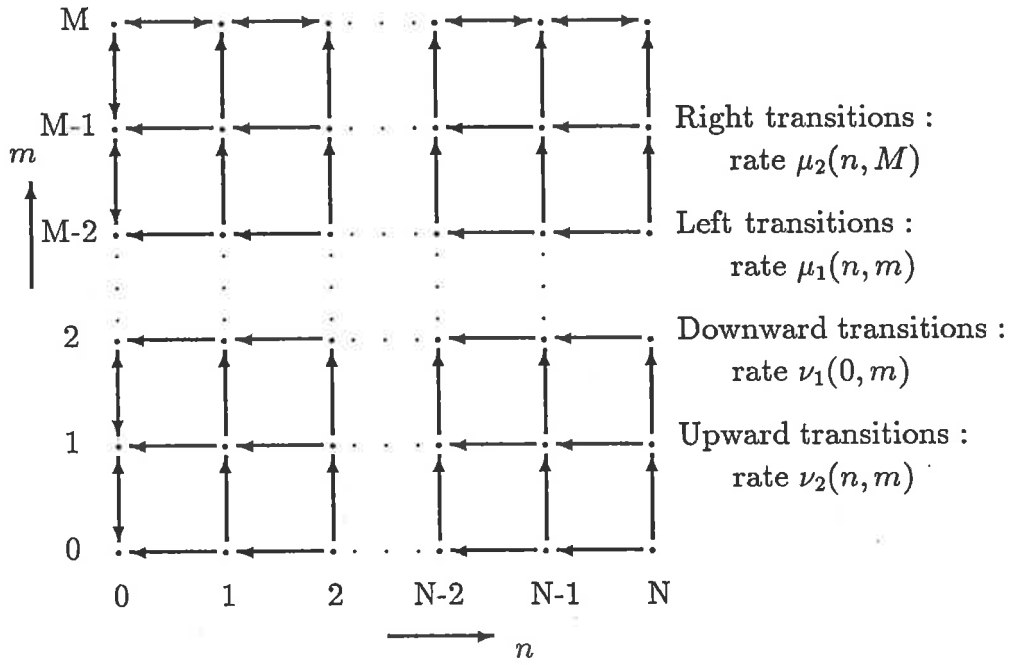


Figure 6.9. State transition diagram for two node pre-emptive priority system with priorities reversed.

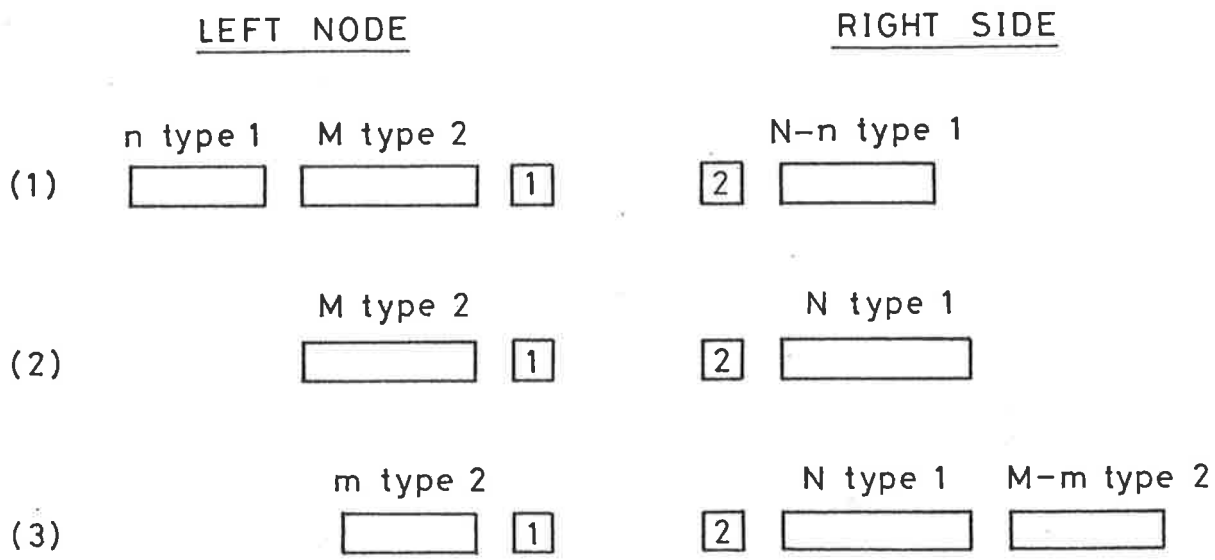


Figure 6.10. Schematic representation of the possible configurations for customers in the two node network with reversed priorities.

Keeping this in mind, we assume that a proportion $\gamma_i(l,k)$ of the total service effort of queue i is dedicated to the customer in position l when the k customers are present. Remember that when the process is in equilibrium the high priority customers at queue i only occupy certain positions. Hence the function $\gamma_i(l,k)$ implicitly carries information on the type of customer in position l.

If, after the arrival of a type t customer to queue i the system is in state k, the customer moves into position l with probability $\delta_{it}(l,k)$. In a fashion similar to that used above, the subscript t may be dropped as the number of customers present provides enough information to determine the type of an arriving customer. However, it will prove convenient later to leave in the subscript.

When a customer leaves position l in queue i the customers in positions $l+1,\ldots,n(i)$ move to positions $l,\ldots,n(i)-1$ respectively, where $n(i)$ is the number of customers at queue i. Similarly, when a customer moves into position l the customers in positions $l,\ldots,n(i)-1$ move into positions $l+1,\ldots,n(i)$.

Definition : Queue i is symmetric with respect to type t customers if

$$\delta_{it}(l,k) = \gamma_i(l,k) \quad \forall l,k. \tag{6.75}$$

Note that in Chapter 3, Section 4 the definition of symmetric was independent of the type of the arriving customer. For example, queue i may behave as a last come-first served pre-emptive resume queue for high priority customers and as a first come first served queue for low priority customers until a high priority customer arrives.

We shall consider a general version of the model considered by Morris. If node i is symmetric (respectively not symmetric) with respect to type t customers then their service times are generally distributed (respectively negative exponentially distributed) with unit mean. Pre-emptive priorities are employed, that is, low priority customers are only served when no high priority customers are present. The service facility at

queue i works at rate $\mu_i(k)$ and $\nu_i(k)$ for type 1 and 2 customers respectively when k customers are present at queue 1.

Lemma 6.3

For the purely Markov process, the network described above has equilibrium distribution given by

$$\pi_k = C \prod_{j=0}^{k-1} \frac{\lambda_j}{\alpha_{j+1}} \quad k = 0, \ldots, M+N, \tag{6.76}$$

where $\pi_k = p_{nm}$ if $k = n + m$ and (n, m) is a recurrent state and

$$\alpha_j = \begin{cases} \nu_1(j) & j = 1, \ldots, M \\ \mu_1(j) & j = M+1, \ldots, M+N, \end{cases} \tag{6.77}$$

$$\lambda_j = \begin{cases} \nu_2(j) & j = 0, \ldots, M-1 \\ \mu_2(j) & j = M, \ldots, M+N-1. \end{cases} \tag{6.78}$$

Proof : The detailed balance equations for the process with all service times negative exponentially distributed are

$$\pi_k \alpha_k = \pi_{k-1} \lambda_{k-1}, \quad k = 1, \ldots, M+N \tag{6.79}$$

where α_k and λ_{k-1} are given by equations (6.77) and (6.78) respectively. Equation (6.76) is the unique solution to the detailed balance equations. ■

For the process with generally distributed service times, we shall construct the network as a Generalised Semi Markov Process in order to derive an insensitivity result regarding the equilibrium distribution. The network we are considering does not fall into the category of networks that are examined by Chandy and Martin (1983). When considered in isolation with Poisson arrivals the queues are not symmetric and hence the analysis of Chandy and Martin may not be utilised.

We employ a construction similar to that used in Barbour and Schassberger (1981).

Let $s_1, s_2, \ldots, s_M$ (respectively $s_{M+1}, \ldots, s_{M+N}$) be labels for lifetimes of type 2 customers (respectively type 1 customers) at queue 1 and let $t_1, \ldots, t_N$ (respectively $t_{N+1}, \ldots, t_{N+M}$) be labels for lifetimes of type 1 (respectively type 2) at queue 2.

Customers of type i arriving at queue j choose randomly from the labels available for their type (they may not select a label already in use). Upon completing service a customer returns its label to the available pool and moves to the other queue. The state of the system is defined by the labels in use and by their position in the queue (that is, the position occupied by the customer with that label).

Let l_i (respectively r_i) denote the label of the customer in position i of queue 1 (respectively queue 2). When k customers are present at queue 1 a complete description of state is then given by the vector $\mathbf{c} = (l_1, \ldots, l_k, r_1, \ldots, r_{M+N-k})$.

Lemma 6.4

The equilibrium distribution for the purely Markov process is

$$p(\mathbf{c}) = \begin{cases} C\left(\prod_{j=0}^{k-1} \frac{\lambda_j}{\alpha_{j+1}}\right) \frac{(M-k)!}{M!} \frac{k!}{M!} \frac{1}{N!} & \text{for } k = 0, \ldots, M \\ C\left(\prod_{j=0}^{k-1} \frac{\lambda_j}{\alpha_{j+1}}\right) \frac{(N+M-k)!}{M!} \frac{(k-M)!}{N!} \frac{1}{N!} & \text{for } k = M+1, \ldots, M+N, \end{cases} \tag{6.80}$$

with products over descending ranges taken to be unity.

Proof : This is most easily proved by considering the possible permutations of labels in use in any state.

Suppose k customers are present at queue 1. Since labels are chosen randomly from the set available to the type of an arriving customer, the probability that labels $l_1, \ldots, l_k$ are in use at queue 1 is given by

$$\begin{cases} \binom{M}{k}^{-1} & \text{for } k = 0, \ldots, M \\ \binom{N}{k-M}^{-1} & \text{for } k = M+1, \ldots, M+N. \end{cases}$$

For the first expression, none of the labels $s_{M+1}, \ldots, s_{M+N}$ are in use (which occurs in precisely one way), while k of the labels $s_1, \ldots, s_M$ are employed (which happens

in $\binom{M}{k}$ ways. Hence the probability that labels $l_1, \ldots, l_k$ are in use is $\binom{M}{k}^{-1}$. For the second expression, all of the labels $s_1, \ldots, s_M$ are in use (which occurs in precisely one way), while $k - M$ of the labels $s_{M+1}, \ldots, s_{M+N}$ are employed (which happens in $\binom{N}{k-M}$ ways. Hence the probability that labels $l_1, \ldots, l_k$ are in use is $\binom{N}{k-M}^{-1}$.

Given that the labels $l_1, \ldots, l_k$ are in use, the probability that they are in the given permutation is given by

$$\begin{cases} \frac{1}{k!} & \text{for } k = 0, \ldots, M \\ \frac{1}{M!}\frac{1}{(k-M)!} & \text{for } k = M+1, \ldots, M+N. \end{cases}$$

In the first expression, there are $k!$ permutations of the labels. In the second expression, all of the labels $s_1, \ldots, s_M$ are in use and can only occupy positions $1, \ldots, M$, while the rest of the labels must occupy positions $M+1$ to k. There are $M!(k-M)!$ such permutations.

Thus the probability that labels $l_1, \ldots, l_k$ are in use and in the given permutation is given by

$$\begin{cases} \binom{M}{k}^{-1}\frac{1}{k!} = \frac{(M-k)!}{M!} & \text{for } k = 0, \ldots, M \\ \binom{N}{k-M}^{-1}\frac{1}{M!}\frac{1}{(k-M)!} = \frac{(N+M-k)!}{N!M!} & \text{for } k = M+1, \ldots, M+N. \end{cases}$$

In a similar fashion when k customers are present at queue 1, (that is, $N+M-k$ are present at queue 2), the probability that labels $r_1, \ldots, r_{M+N-k}$ are being used is given by

$$\begin{cases} \binom{M}{M-k}^{-1}\frac{1}{(M-k)!} = \frac{k!}{N!M!} & \text{for } k = 0, \ldots, M \\ \binom{N}{M+N-k}^{-1}\frac{1}{(M+N-k)!} = \frac{(k-M)!}{N!} & \text{for } k = 0, \ldots, M+N. \end{cases}$$

Therefore, the probability that labels $(l_1, \ldots, l_k, r_1, \ldots, r_{M+N-k})$ are in use in the given permutation, conditional upon exactly k customers being present at queue 1, is

$$\begin{cases} \frac{(M-k)!}{M!}\frac{k!}{M!}\frac{1}{N!} & k = 0 \ldots M \\ \frac{(N+M-k)!}{N!}\frac{(k-M)!}{N!}\frac{1}{M!} & k = M+1 \ldots M+N. \end{cases}$$

To obtain the probability of being in state $(l_1, \ldots, l_k, r_1, \ldots, r_{M+N-k})$, we then simply multiply by the probability of having k customers at queue 1, which is given by equation (6.76). We have thus derived equation (6.80). ■

Theorem 6.5

The equilibrium distribution for the network described earlier (that is, with generally distributed service times for type i customers at queue j when queue j is symmetric with respect to type i) is given by equation (6.76).

Proof : We will use the GSMP construction given above and show that the partial balance equations (see König and Jansen (1974) or Section 2.3) are satisfied for the purely Markov process when queue j is symmetric with respect to type i. When the partial balance equations are satisfied, general distributions may be used in place of negative exponential distributions without affecting the equilibrium distribution.

Suppose k customers are present at queue 1 in state $\mathbf{c}$. For $k = 0, \ldots, M$, let $R(\mathbf{c})$ be the subset of $\{t_{N+1}, \ldots, t_{N+M}\}$ not in use at queue 2 in state $\mathbf{c}$. For $k = M+1, \ldots, M+N$, let $R(\mathbf{c})$ be the subset of $\{t_1, \ldots, t_N\}$ not in use at queue 2 in state $\mathbf{c}$. Let $\mathbf{c}(l, r^*, i)$ denote the state formed from $\mathbf{c}$ by the removal of label l from queue 1 and the insertion of label $r^* \in R(\mathbf{c})$ into position i of queue 2.

For $k \leq M$, the flux out of state $\mathbf{c}$ due to the death of the label, l_j, in position j of queue 1 is $p(\mathbf{c})\gamma_1(j,k)\alpha_k$.

The flux of that label being born in position j of queue 1 to create state $\mathbf{c}$ is,

$$\sum_{r^* \in R(\mathbf{c})} \sum_{i=1}^{M+N-k+1} p(\mathbf{c}(l_j, r^*, i))\gamma_2(i, M+N-k+1)\lambda_{k-1}\delta_{12}(j,k)/(M-k+1). \tag{6.81}$$

Note that

$$\begin{aligned} p(\mathbf{c}(l_j, r^*, i)) &= C \prod_{j=0}^{k-1} \frac{\lambda_j}{\alpha_{j+1}} \frac{(M-k)!}{M!} \frac{k!}{M!} \frac{1}{N!} \\ &= p(\mathbf{c}) \frac{\alpha_k}{\lambda_{k-1}} \frac{M-k+1}{k}. \end{aligned} \tag{6.82}$$

Substituting from expression (6.82), expression (6.81) then becomes

$$
\begin{aligned}
& p(\mathbf{c})\frac{\alpha_k}{\lambda_{k-1}}\frac{M-k+1}{k}\sum_{r^*\in R(\mathbf{c})}\sum_{i=1}^{M+N-k+1}\gamma_2(i,M+N-k+1)\lambda_{k-1}\frac{\delta_{12}(j,k)}{M-k+1} \\
&\quad = p(\mathbf{c})\frac{\alpha_k\delta_{12}(j,k)}{k}\sum_{r^*\in R(\mathbf{c})}1 \\
&\quad = p(\mathbf{c})\alpha_k\delta_{12}(j,k),
\end{aligned}
\tag{6.83}
$$

where the first equality holds because $\sum_{i=1}^{j}\gamma_2(i,j)=1$ and the second equality holds because there are k labels in the set $R(\mathbf{c})$.

So, for $k=1\ldots M$, the partial balance equations hold when

$$
p(\mathbf{c})\alpha_k\gamma_{12}(j,k)=p(\mathbf{c})\alpha_k\delta_{12}(j,k). \tag{6.84}
$$

This is true if $\delta_{12}(j,k)=\gamma_1(j,k)$, that is, when queue 1 is symmetric with respect to type 2 customers. As a consequence, general distributions may then be used for type 2 customers without affecting the equilibrium distribution.

Using a similar procedure at queue 1 for $k=M+1,\ldots,M+N$ shows that general distributions may be used for type 1 customers at queue 1 when it is symmetric with respect to type 1, without affecting the equilibrium distribution.

Repeating the analysis for queue 2 shows that the equilibrium distribution is insensitive to changes in the service time distributions (provided the mean is fixed) when queue 2 is symmetric with respect to type i.

To find the equilibrium distribution for the number of customers at each queue, we simply sum over the appropriate permutations of labels of customers at each queue. This gives expression (6.76). ■

Morris noted that this system may be used to model a full-duplex data link which is used for transmission of messages under a window flow control protocol when there are two grades of messages.

6.6 Two node closed networks, pre-emptive priority at the left node and non-batch servicing at the right

The analysis is now extended to two node closed networks where pre-emptive priority discipline is used at the left node but at the right node the discipline is allowed to be any state dependent discipline except for batch servicing. All service times will now be assumed to be negative exponentially distributed. Let the set of states $\{(n,m); 0 \leq n \leq N\}$ be denoted by the level m. This allows the system to be considered as a quasi birth and death process (see Neuts (1981)), whereby transitions with positive rate cause the level to change by at most one up or down. This is the reason we do not allow more than one customer at a time to complete service.

The service times are negative exponentially distributed with parameters $\mu_i(n,m)$ and $\nu_i(n,m)$ at node i for high and low priority customers respectively when the state is (n,m). The state transition diagram is given in Figure 6.11.

It is clear, since the process at each transition may only move from level m to one of $m-1$ (for $m>0$), m and $m+1$ (for $m<M$), that the process has a block tri-diagonal Q-matrix,

$$Q = \begin{bmatrix} A_1^{(0)} & A_0^{(0)} & & & \\ A_2^{(1)} & A_1^{(1)} & A_0^{(1)} & & \\ & \ddots & \ddots & \ddots & \\ & & A_2^{(M-1)} & A_1^{(M-1)} & A_0^{(M-1)} \\ & & & A_2^{(M)} & A_1^{(M)} \end{bmatrix} \tag{6.85}$$

with submatrices of order $(N+1)\times(N+1)$, where $[A_0^{(m)}]_{ij}$ is the rate of moving from state (i,m) to state $(j,m+1)$. $[A_0^{(m)}]_{ij} = 0$ for $i \neq j$ and $[A_0^{(m)}]_{ii} = \nu_2(i,m)$.

$[A_2^{(m)}]_{ij}$ is the rate of moving from state (i,m) to state $(j,m-1)$. Thus $[A_2^{(m)}]_{ij} = \nu_1(0,m)\delta_{i0}\delta_{j0}$. Hence $A_2^{(m)}$ may be written in the form $A_2^{(m)} = \nu_1(0,m)\beta^t\beta$, where $\beta = (1,0,\ldots,0)$. In the following we shall adopt this convention.

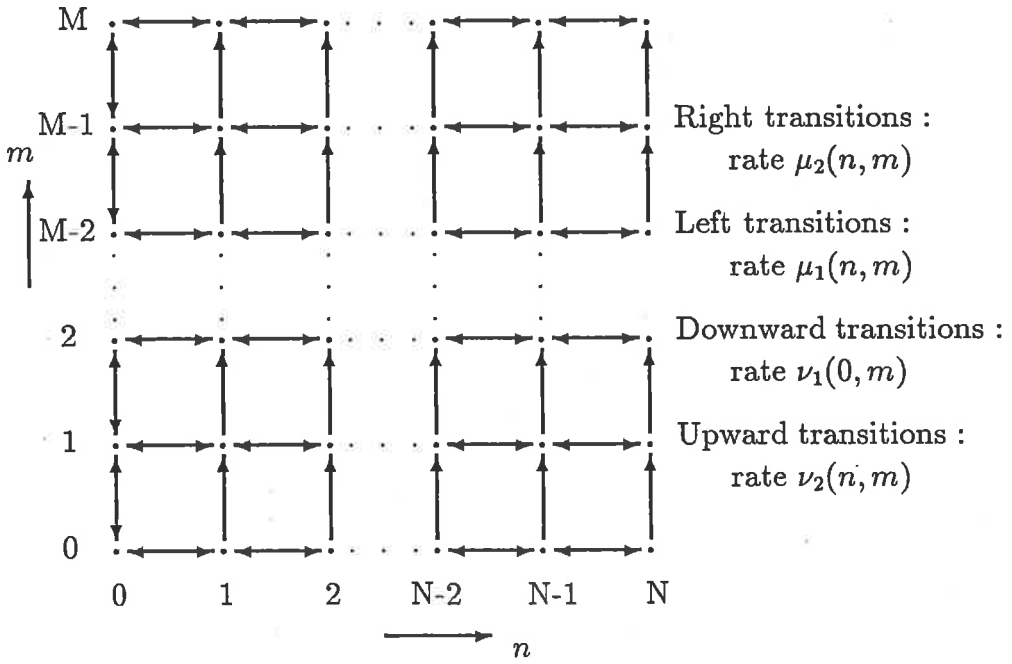


Figure 6.11. State transition diagram for two node network with pre-emptive priority at queue 1 and non-batch servicing at queue 2.

$[A_1^{(m)}]_{ij}$ is the rate of moving from state (i,m) to state (j,m) for $i \neq j$, and is the negative of the total rate of moving out of state (i,m) when $i = j$. Thus $[A_1^{(m)}]_{ij} = 0$ for $j \neq i-1, i$ or $i+1$, $[A_1^{(m)}]_{i,i-1} = \mu_1(i,m)$ and $[A_1^{(m)}]_{i,i+1} = \mu_2(i,m)$.

Thus, the global balance equations may be written as

$$\begin{aligned} \mathbf{p}_0 A_1^{(0)} + \mathbf{p}_1 A_2^{(1)} &= 0, \\ \mathbf{p}_{m-1} A_0^{(m-1)} + \mathbf{p}_m A_1^{(m)} + \mathbf{p}_{m+1} A_2^{(m+1)} &= 0, \quad m = 1 \ldots M-1 \\ \mathbf{p}_{M-1} A_0^{(M-1)} + \mathbf{p}_M A_1^{(M)} &= 0, \end{aligned} \tag{6.86}$$

where $\mathbf{p}_m = (p_{0m}, p_{1m}, \ldots, p_{N,m})$.

Systems with the structure of (6.86) have been studied in general by Gaver, *et al*, (1984). However, in the system being examined here, the matrices $A_2^{(m)}$ are of rank one which allows a closed form of solution to be found, rather than an iterative solution.

Theorem 6.6

If the service rates in the above system are independent of m, the number of low priority customers at the left node, then

(1) $A_1^{(1)} = A_1^{(2)} = \ldots = A_1^{(M-1)} = A_1$, say, $A_0^{(0)} = A_0^{(1)} = \ldots = A_0^{(M-1)} = A_0$, $A_2^{(1)} = A_2^{(2)} = \ldots = A_2^{(M)} = A_1$ $A_1^{(0)} = A_1 + A_2$ and $A_1^{(M)} = A_1 + A_0$.

(2) The equilibrium distribution is given by

$$\begin{aligned} \mathbf{p}_m &= \mathbf{p}_{m-1}(-A_0)(A_1 + A_0 \mathbf{e}\beta)^{-1} \\ &= \mathbf{p}_0 R^m, \qquad 0 \leq m \leq M-1, \\ \mathbf{p}_M &= \mathbf{p}_0 R^{M-1}(-A_0 A_1^{(M)-1}), \end{aligned} \tag{6.87}$$

where $R = (-A_0)(A_1 + A_0\mathbf{e}\beta)^{-1}$, $\mathbf{e} = (1,1,\ldots,1)^t$ and $\mathbf{p}_0$ is the unique positive solution to the system

$$\mathbf{p}_0(A_1 + A_2 + A_0\mathbf{e}\beta) = 0 \tag{6.88}$$

and

$$\mathbf{p}_0 \sum_{m=0}^{M-1} R^m \mathbf{e} + \mathbf{p}_0 R^{M-1}(-A_0(A_1+A_0)^{-1})\mathbf{e} = 1. \tag{6.89}$$

Proof : (1) As the rates are now independent of the level, it is immediate that $A_1^{(1)} = A_1^{(2)} = \ldots = A_1^{(M-1)}$. Similarly for A_0 and A_2. No downward transitions are allowed from level zero, and since this is the only difference between level zero and levels 1, 2,... ,M-1, adding the row sums of A_2 to the diagonal elements of A_1 gives $A_1^{(0)}$. As A_2 is diagonal, this gives $A_1^{(0)} = A_1 + A_2$. In a similar fashion, $A_1^{(M)} = A_1 + A_0$.

(2) This could be proved by application of the results of Gaver, *et al* (1984), but we give the outline of a proof which exploits the particular structure of the process.

Cut equations between levels m and $m+1$ yield :

$$\mathbf{p}_m A_0 \mathbf{e} = \mathbf{p}_{m+1} \nu_1(0) \beta^t. \tag{6.90}$$

Multiplying on the right by β gives

$$\mathbf{p}_m A_0 \mathbf{e}\beta = \mathbf{p}_{m+1} \nu_1(0) \beta^t \beta = \mathbf{p}_{m+1} A_2. \tag{6.91}$$

Substituting from expression (6.91) into (6.86b) gives

$$\mathbf{p}_{m-1} A_0 + \mathbf{p}_m (A_1 + A_0 \mathbf{e}\beta) = \mathbf{0} \tag{6.92}$$

Rearranging gives equation (6.87a) and using induction gives (6.87b) ($A_1 + A_0\mathbf{e}\beta$ is non-singular by application of the Perron-Fröbenius Theorem). Rearranging (6.86c) and substituting from (6.87b) yields (6.87c). The solution is completed by finding $\mathbf{p}_0$. Equations (6.86a), (6.91) and $A_1^{(0)} = A_1 + A_2$ give (6.88) and the normalisation equation is (6.89). ■

A similar result may be shown if the service rates are made dependent upon (n, m).

Theorem 6.7

For the system described above with state dependent service rates, $\mu_i(n,m)$ and $\nu_i(n,m)$, $(i=1,2)$, the equilibrium distribution is given by

$$\begin{aligned} \mathbf{p}_m &= \mathbf{p}_{m-1}(-A_0^{(m-1)})(A_1^{(m)} + A_0^{(m)}\mathbf{e}\beta)^{-1}, \quad m = 1 \ldots M-1, \\ \mathbf{p}_M &= \mathbf{p}_{M-1}(-A_0^{(M-1)}A_1^{(M)-1}), \end{aligned} \tag{6.93}$$

with $\mathbf{p}_0$ the unique positive solution to

$$\begin{aligned} \mathbf{p}_0(A_1^{(0)} + A_0^{(0)}\mathbf{e}\beta) &= 0, \\ \sum_{m=0}^{M} \mathbf{p}_m\mathbf{e} &= 1. \end{aligned} \tag{6.94}$$

Proof : The proof follows along lines similar to those used in Theorem 6.6 and so is not given here. ■

The low priority throughput, T_L, and utilisation, U_L, at the left node are given by

$$T_L = \sum_{n=0}^{N}\sum_{m=1}^{M} p_{nm}\nu_1(n,m) \tag{6.95}$$

and

$$U_L = \sum_{n=0}^{N}\sum_{m=1}^{M} p_{nm}. \tag{6.96}$$

Examples : (1) A simplified model of an interactive computer system with N interactive users (high priority), M batch jobs (low priority) and one CPU would have rates $\mu_1(n,m) = \mu_1$, $\mu_2(n,m) = n\mu_2$, $\nu_1(n,m) = \nu_1\delta_{n0}$ and $\nu_2(n,m) = m\nu_2$. Table 6.3 gives low priority throughputs and utilisations at the left node for this model and the system described in Section 6.1 where only one low priority server is present at the right node.

Service rates				Low priority throughput (x10^{-3}) *Number of servers at right node*	
μ_1	μ_2	ν_1	ν_2	1 server	10 servers
1	1	1	1	3.066	3.067
0.5	1	1	1	0.1579	0.1580
2	1	1	1	35.54	37.00
1	2	1	1	0.1580	0.1580
1	1	2	1	6.092	6.131
1	1	1	0.5	3.065	3.067
1	1	1	2	3.066	3.067

Table 6.3 Low priority throughputs for a system with 5 high and 5 low priority customers and one server at the left node.

(2) A system with $K < M + N$ servers at the right node, one server and preemptive priority at the left node would have rates

$$\mu_1(n,m) = \mu_1,$$

$$\mu_2(n,m) = min(n,K)\mu_2,$$

$$\nu_1(n,m) = \nu_1\delta_{n0},$$

$$\nu_2(n,m) = min(m,K-n)\nu_2.$$

CHAPTER 7 : CONCLUSION

7.1 Conclusions and suggestions for further research

In part one of this thesis we considered insensitivity in Generalised Semi-Markov Processes. In Chapter 3 it was shown that age dependent GSMPs have a product form supplemented equilibrium distribution if and only if a related purely Markov process satisfies a certain set of partial balance equations. New forms of the supplemented global balance equations were employed which incorporated information on both the spent and residual lifetimes. In this fashion, new results on insensitivity in GSMPs were derived.

Using this framework, we showed that the results on insensitivity in queueing networks may be extended to networks where the routing, from symmteric nodes, is dependent upon the customers' service requirement. It was also shown that results on semi-Markov Processes (Çinlar (1975)) and interruption processes (Henderson and Taylor (1987) and Taylor (1987)) fall out as special cases of the theory developed here. It would be interesting to see if, for age dependent processes, an insensitivity implies partial balance result exists (along the lines of Henderson and Taylor (1987)), but this has not been attempted here. Another possibility would be to consider some form of anticipating routing. That is, processes where the next service requirement is chosen in advance, and the routing is made dependent upon this requirement.

We also considered the use of age dependent speeds, in which the speed that a lifetime is worked on is allowed to depend on the amount of service it has received. While the assumptions made may be somewhat restrictive in this case, it might be possible to derive an approximation method, using the results presented here, for such processes.

In Chapter 4 an alternative and simpler proof of a result of Taylor (1987) on processes with zero speeds was given. By using an expanded state space to model

such systems, we then extended the results of Henderson (1983b) on processes with instantaneous attention.

In part two of this thesis we examined closed, two node priority queueing networks. When both nodes utilise a pre-emptive priority discipline, results of Morris (1981) were extended to allow state dependent service rates. We then found the equilibrium distribution for a closed network with one node using a nonpre-emptive priority discipline. Using results discussed in Chapter 2, the equilibrium distribution for a network with reversed priorities, under certain conditions, was found to be insensitive. The Chapter was concluded by a general formulation of two node networks in terms of the quasi-birth and death processes of Gaver *et al* (1984).

The complexity of the results in Chapter 6 indicate that exact solutions for networks of queues with priorities will be impractical and hence future work should be directed towards finding efficient approximation techniques.

References

BARBOUR, A. (1976) Networks of queues and the method of stages. *Adv. Appl. Prob.*, **8**, 584-591.

BARBOUR, A. (1982) Generalised semi-Markov schemes and open queueing networks. *J. Appl. Prob.*, **19**, 469-474.

BARBOUR, A. AND SCHASSBERGER, R. (1981) Insensitive average residence times in generalised semi-Markov processes. *Adv. Appl. Prob.*, **13**, 720-735.

BASKETT, F., CHANDY, K.M., MUNTZ, R.R. AND PALACIOS, F.G. (1975) Open, closed and mixed networks of queues with different classes of customers. *J. Assoc. Comp. Mach.*, **22**, 248-260.

BRANDWAJN, A. (1982) A finite difference equations approach to a priority queue. *Oper. Res.*, **30**, 74-81.

BRUMELLE, S.L. (1978) A generalization of Erlang's loss system to state dependent arrival and service rates. *Math. Op. Res.*, **3**, 10-16.

BUNDAY, B.D. AND SCRATON, R.E. (1980) The $G/M/r$ machine interference model. *Euro. J. Oper. Res.*, 4, 399-402.

CHAIKEN, J. AND IGNALL, E. (1972) An extension of Erlang's formulas which distinguishes individual servers. *J. Appl. Prob.*, **9**, 192-197.

CHANDY, K.M., HOWARD, J.H. AND TOWSLEY, D.F. (1977) Product form and local balance in queueing networks. *J. Assoc. Comp. Mach.*, **24**, 250-263.

CHANDY, K.M. AND MARTIN, J. (1983) A characterization of product form queueing networks. *J. Assoc. Comp. Mach.*, **30**, 250-263.

COHEN, J. (1957) The generalised Engset formulae, *Philips Telecomm. Review*, **18**, 158-170.

ÇINLAR, E. (1975) *Introduction to stochastic processes.* Prentice-Hall, New Jersey.

ERLANG, A. (1917) Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. *Post Office Electrical Engineer's Journal*,

10, 826-834.

FORTET, R. (1950) Calcul des probabilités. *Centre National de la Recherche Scientifique*, Paris.

FRANKEN, P., ARNDT, U., KÖNIG, D. AND SCHMIDT, V. (1982) *Queues and point processes.* Wiley, London.

GAVER, D.P. (1962) A waiting line with interrupted service including priorities. *J. Royal Stat. Soc. B*, **24**, 73-90.

GAVER, D.P., JACOBS, P.A. AND LATOUCHE, G., (1984) Finite birth-and-death models in randomly changing environments. *Adv. Appl. Prob.*, **16**, 715-731.

GORDON, W. AND NEWELL, G. (1967) Closed queueing systems with exponential servers. *Oper. Res.*, **15**, 254-265.

HENDERSON, W. (1969) GI/M/1 Priority queue. *Oper. Res.* **17**, 907-910.

HENDERSON, W. (1983a) Insensitivity and reversed Markov processes. *Adv. Appl. Prob.*, **15**, 752-768.

HENDERSON, W. (1983b) Non standard insensitivity. *J. Appl. Prob.* **20**, 288-296.

HENDERSON, W. AND TAYLOR, P. (1987) Insensitivity with interruptions. Submitted.

HOKSTAD, P., (1978) A M/G/1 priority queue. *INFOR* **16**, No. 2, 158-170.

HORDIJK, A. AND VAN DIJK, N. (1981) Networks of queues with blocking. *Performance '81*, (ed. K. Kylstra), North Holland, 51-65.

HORDIJK, A. AND VAN DIJK, N. (1983a) Networks of queues. *Proc. International Seminar on Modelling and Performance Evaluation*, INRIA, **1**, 79-135.

HORDIJK, A. AND VAN DIJK, N. (1983b) Adjoint processes, job local balance and insensitivity for stochastic networks. *Bull.* 44^{th} *Session Int. Stat. Inst.*, **50**, 776-788.

IKEHARA, S. AND MIYAZAKI, M., (1985) Approximate analysis of queueing networks with nonpre-emptive scheduling. *International Teletraffic Congress 11*, 3.4A-2-1 to 3.4A-3-7.

JAISWAL, N.K., (1968) *Priority Queues*, Academic Press, New York.

JACKSON, J. (1957) Networks of waiting lines. *Oper. Res.*, **5**, 518-521.

JACOBI, H. (1965) Eine unempfindlichkeitseigenschaft für geordnete bündel ungeordnete teilbündel. *Wiss. Z. Friedrich-Schiller-Universität Jena. Math.-Nat.*, **14**, 251-260.

JANSEN, U. (1980) Unempfindlichkeitseigenschaft für verschiedene auswahlregeln ohne vorliegen produktform der stationären verteilung. *Elek. Info. Kyb.*, **16**, 443-448.

JANSEN, U. (1984) Conditional expected sojourn times in insensitive queueing systems and networks. *Adv. Appl. Prob.*, **15**, 752-768.

JANSEN, K., KÖNIG, D. AND NAWROTZKI, K. (1979) A criterion of insensitivity for a class of queueing systems with random marked point processes. *Math. Operationsforsch u. Statist., Ser. Optimization*, **10**, 379-403.

KAUFMAN, J.S., (1982) Approximate analysis of priority scheduling disciplines in queueing network models of computer systems, *6th International conference on computer communication, London*, 955-961.

KEILSON, J. (1962) Queues subject to service interruptions. *Ann. Math. Stat.*, **33**, 1314-1322.

KELLY, F.P. (1976) Networks of queues. *Adv. Appl. Prob.* **8**, 416-432.

KELLY, F.P. (1979) *Reversibility and stochastic processes.* Wiley, London.

KÖNIG, D. Verallgemeinerungen der Engsetschen formeln. *Math. Nachr.*, **28**, 145-155.

KÖNIG, D. AND JANSEN, U. (1974) Stochastic processes and properties of invariance for queueing systems with speeds and interruptions. *Trans 7th Prague Conference on Information Theory* Reidel, Dordrecht, 335-343.

KÖNIG, D. AND MATTHES, K. (1963) Verallgemeinerungen der Erlangschen formeln I. *Math. Nachr.*, **26**, 45-56.

KÖNIG, D., MATTHES, K. AND NAWROTZKI, K. (1967) Verallgemeinerungen

der Erlangschen und Ensetschen formeln (eine methode in der Bedienungtheorie). Akademie-Verlag, Berlin.

KUEHN, P.J. AND SCHMITT W., (1985) Transit delay distributions in priority queueing networks. *International Teletraffic Congress 11*, 3.4A-3-1 to 3.4A-3-6.

MATTHES, K. (1962) Zur Theorie der Bedienungprozesse. *Trans 3rd Prague Conf Information Theory.*

MILLER, R.G., (1959) Priority queues. *Ann. Math. Stats.* **31**, 86-103.

MORRIS, R.J.T., (1981) Priority queueing networks, *The Bell System Technical Journal* **60**, No. 8, 1745-1769.

NEUTS, M.F., (1981) *Matrix geometric solutions in stochastic models : An algorithmic approach.* The Johns Hopkins University Press, Baltimore.

NOETZEL, A. (1979) A generalized queueing discipline for product form queueing networks. *J. Assoc. Comp. Mach.*, **26**, 779-793.

SCHASSBERGER, R. (1977) Insensitivity of steady state distributions of generalised semi-Markov processes, Part I. *Ann. Prob.*, **5**, 87-99.

SCHASSBERGER, R. (1978a) Insensitivity of steady state distributions of generalised semi-Markov processes, Part II. *Ann. Prob.*, **6**, 85-93.

SCHASSBERGER, R. (1978b) Insensitivity of steady state distributions of generalised semi-Markov processes with speeds. *Adv. Appl. Prob.* **10**, 836-851.

SCHASSBERGER, R. (1986) Two remarks on insensitive stochastic models. Adv. Appl. Prob. **18**, 791-814.

SCHMITT W., (1983) Approximate analysis of Markovian queueing networks with priorities. *International Teletraffic Congress 10*, Paper 3, Session 1.3.

SEVAST'YANOV, B. (1957) An ergodic theorem for Markov processes and its application to telephone systems with refusals. *Theor. Prob. Appl.*, **2**, 104-112.

SEVCIK, K.C. (1977) Priority scheduling disciplines in queueing network models of computer systems. *Information processing '77*, B. Gilchrist (Ed.), IFIP, Amsterdam

: North-Holland Publishing Co.

STEPHAN, S.S., (1958) Two queues under pre-emptive priority with Poisson arrival and service rates. *Oper. Res.* **6**, 399-418.

TAKACS, L. (1969) On Erlang's formula. *Ann. Math. Stat.*, **40**, 71-78.

TAYLOR, P. (1987) *Aspects of insensitivity in stochastic processes.* Ph.D. Thesis, University of Adelaide.

WHITE, H. AND CHRISTIE, L.S., (1958) Queueing with preemptive priorities and breakdowns. *Oper. Res.* **6**, 79-96.

WHITT, W. (1980) Continuity of generalised semi-Markov processes. *Math. Oper. Res.*, **5**, 494-501.

WHITTLE, P. (1985) Partial balance and insensitivity. *J. Appl. Prob.* **22**, 168-176.

WHITTLE, P. (1986) *Systems in stochastic equilibrium*, Wiley, London.

WOLFF, R. AND WRIGHTSON, C. (1976) An extension of Erlang's loss formula, *J. Appl. Prob.*, **13**, 628-632.