Priority queues

R. J. Reed

P R I O R I T Y Q U E U E S

by

R. J. Reed, A.R.C.S., B.Sc.

A thesis presented for the degree of Doctor of Philosophy

at the University of Durham.

September, 1971

Mathematics Department,
University of Durham.

C O N T E N T S

ACKNOWLEDGEMENTS

My thanks go to my supervisor Dr. A. G. Hawkes and to the Science Research Council for financial support in the form of an S.R.C. Research Studentship.

ABSTRACT

Extensive research has been carried out in the subject of Priority Queues over the past ten years, culminating in the book by Jaiswal [8]. In this thesis, certain isolated problems which appear to have been omitted from the consideration of other authors are discussed.

The first two chapters are concerned with the question of how priorities should be allocated to customers (or 'units') arriving at a queue so as to minimize the overall mean waiting time [it is perhaps worth mentioning at the outset that following current usage, the terms 'queueing time' and 'waiting time' will be used synonymously throughout; both refer to the time a unit waits before commencing service]. In previous treatments of this 'allocation of priorities' problem it has always been assumed that on arrival, the service time requirement of a unit could be predicted exactly; the effect of having only imperfect information in the form of an estimated service time is considered here. Chapter 1 deals with the non-preemptive discipline; Chapter 2 with discretionary disciplines.

Priority queues in which the arrival epochs of different types of units form independent renewal processes have only been solved under the assumption of random arrivals. However, if the following modified arrival scheme is considered:

> arrival epochs form an ordinary renewal process, and at any arrival epoch, independently of what happened at all previous epochs, with probability q_1 the arrival is a priority unit and with probability q_2 a non-priority unit (where $q_1+q_2=1$)

then the priority analogues of the ordinary single-server queues $E_b/G/1$ and GI/M/1 can be solved (Chapters 3 and 4 respectively).

In conclusion, Chapter 5 is concerned with approximate methods:

section 1 is a review of previous work on deriving bounds for the mean waiting time in a GI/G/1 queue, section 2 extends this work to the GI/G/1 priority queue.

September, 1971

CHAPTER 1

Non-Preemptive Priority Classification in the M/G/1 Queue

1.1 Introduction

In this section the Laplace-Stieltjes transform of the distribution function of the waiting time of customers or 'units' in a non-preemptive priority queue is derived; this has been obtained before by many authors - the derivation below is an adaptation of that given by Miller [14] and Takacs [18].

The basic model to be considered is as follows:- units arriving for service from a single server are of two different types with units of type 1 (or '1-units') having non-preemptive priority over units of type 2. It is further assumed that k-units arrive at random and independently of arrivals of other units with rate λ_k, and have service times which are independently and identically distributed random variables with mean $1/\mu_k$ and distribution function $S_k(t)$, $t \geq 0$ $(k=1,2)$. Let $\rho_k = \lambda_k/\mu_k$ and $\lambda = \lambda_1 + \lambda_2$ where necessarily $\rho_1 + \rho_2 < 1$ for stability.

The departure epochs $\tau_1', \tau_2', \tau_3', \ldots$ form a set of regeneration points and therefore an imbedded Markov chain can be defined.

Let $\tau_0' = 0$,

and for $n \geq 0$

$$\alpha_o^n = P[\text{no 1-units and a non-zero number of 2-units waiting at time } \tau_n' + 0].$$

$$\alpha_k^n = P[k \text{ 1-units waiting at time } \tau_n' + 0] \qquad k \geq 1$$

$$\beta_o^n = P[\text{system empty at time } \tau_n' + 0].$$

It follows that if

$$a_k^i = P[k \text{ 1-units arrive during the service time of an i-unit}].$$

$$= \int_0^\infty \frac{e^{-\lambda_1 x}(\lambda_1 x)^k}{k!} \, d\, S_i(x) \qquad i = 1,2;\ k \geq 0$$

and $q_i = \lambda_i/(\lambda_1+\lambda_2) \qquad i = 1,2$

then

$$\alpha_o^{n+1} + \beta_o^{n+1} = P\left[\text{no 1-units waiting at } \tau'_{n+1} + 0\right]$$

$$= \alpha_1{}^n \, a_o{}^1 + \alpha_o{}^n \, a_o{}^2 + \beta_o{}^n \left[q_1 a_o{}^1 + q_2 a_o{}^2\right] \qquad (1)$$

and for $k \geq 1$

$$\alpha_k^{n+1} = \sum_{i=0}^{\infty} \alpha_{k+1-i}^n \, a_i{}^n + \alpha_o{}^n \, a_k{}^2 + \beta_o{}^n \left[q_1 a_k{}^1 + q_2 a_k{}^2\right] \qquad (2)$$

For $|z| \leq 1$, $|w| < 1$ define

$$\alpha^n(z) = \sum_{i=0}^{\infty} \alpha_i{}^n \, z^i \qquad n \geq 0$$

$$\alpha(z,w) = \sum_{n=0}^{\infty} \alpha^n(z) \, w^n \ ; \quad \beta_o(w) = \sum_{n=0}^{\infty} \beta_o{}^n \, w^n$$

and

$$\alpha_o(w) = \sum_{n=0}^{\infty} \alpha_o{}^n \, w^n$$

then equations (1) and (2) give

$$\alpha^{n+1}(z) + \beta_o^{n+1} = \bar{S}_1(\lambda_1-\lambda_1 z) \, \frac{\alpha^n(z)-\alpha_o{}^n}{z} + \alpha_o{}^n \, \bar{S}_2(\lambda_1-\lambda_1 z)$$

$$+ \beta_o{}^n\left[q_1 \bar{S}_1(\lambda_1-\lambda_1 z) + q_2 \bar{S}_2(\lambda_1-\lambda_1 z)\right] \qquad (3)$$

where

$$\bar{S}_i(\lambda_1-\lambda_1 z) = \int_0^\infty \exp\left[-x(\lambda_1-\lambda_1 z)\right] d\, S_i(x)$$

$$= \sum_{k=0}^{\infty} a_k{}^i \, z^k \qquad i = 1,2$$

Assuming the queue to be initially empty gives

$$\alpha(z,w) = \{z + w\,\alpha_o(w)\,[z\,\bar{S}_2(\lambda_1-\lambda_1 z) - \bar{S}_1(\lambda_1-\lambda_1 z)] + z\,\beta_o(w)\,[wq_1\,\bar{S}_1(\lambda_1-\lambda_1 z) + wq_2\,\bar{S}_2(\lambda_1-\lambda_1 z) - 1]\}/\{z - w\,\bar{S}_1(\lambda_1-\lambda_1 z)\} \qquad (4)$$

Clearly

$$\beta_o^{\,n} = P[\text{no customers waiting after the } n^{th} \text{ departure in an ordinary } M/G/1 \text{ queue, arrival rate } \lambda, \text{ service time d.f. } q_1 S_1(x) + q_2 S_2(x)].$$

Therefore, by page 71 of Takacs [17]

$$\beta_o(w) = \frac{1}{1-g(w)} \qquad (5)$$

where $g(w)$ is the unique root in z within the unit circle of the equation

$$z = w[q_1\,\bar{S}_1(\lambda-\lambda z) + q_2\,\bar{S}_2(\lambda-\lambda z)]$$

Also

$$g(1) = 1,\ g'(1) = 1/(1-\rho_1-\rho_2)$$

The Lemma on page 47 of [17] shows that the denominator of the right hand side of equation (4) has exactly one root $z = h(w)$ say in the unit circle, and this must be a root of the numerator also. Therefore

$$\alpha_o(w) = \frac{h(w)[1+\beta_o(w)(wq_1\bar{S}_1(\lambda_1-\lambda_1 h(w))+wq_2\bar{S}_2(\lambda_1-\lambda_1 h(w))-1)]}{w[\bar{S}_1(\lambda_1-\lambda_1 h(w))-h(w)\,\bar{S}_2(\lambda_1-\lambda_1 h(w))]} \qquad (6)$$

where

$$h(1) = 1 \text{ and } h'(1) = 1/(1-\rho_1)$$

Equations (4), (5) and (6) determine $\alpha(z,w)$. For $\rho_1 + \rho_2 < 1$ the Markov chain is irreducible and aperiodic and so the limiting probabilities

$$\alpha_k = \lim_{n\to\infty} \alpha_k^{\,n}, \quad \beta_o = \lim_{n\to\infty} \beta_o^{\,n}$$

always exist and are independent of the initial distribution. Using Abel's theorem gives

$$\alpha(z) = \sum_{k=0}^{\infty} \alpha_k z^k = \lim_{w\to 1} (1-w)\, \alpha(z,w)$$

$$= \frac{\alpha_o\left[z\bar{S}_2(\lambda_1-\lambda_1 z)-\bar{S}_1(\lambda_1-\lambda_1 z)\right]+\beta_o z\left[q_1\bar{S}_1(\lambda_1-\lambda_1 z)+q_2\bar{S}_2(\lambda_1-\lambda_1 z)-1\right]}{z-\bar{S}_1(\lambda_1-\lambda_1 z)}$$

where

$$\beta_o = \lim_{w\to 1} (1-w)\, \beta_o(w) = 1 - \rho_1 - \rho_2$$

and

$$\alpha_o = \lim_{w\to 1} (1-w)\, \alpha_o(w) = \frac{-(1-\rho_1)+\beta_o(1+q_1\rho_2-q_2\rho_1)}{\lambda_1/\mu_1-\lambda_1/\mu_2-1}$$

$$= q_2(\rho_1+\rho_2)$$

$\alpha(z)$ is thus determined; clearly the joint distribution of the number of 1-units and the number of 2-units at the n^{th} departure epoch for any initial condition could be determined in exactly the same manner. If W_1 denotes the waiting time of a 1-unit in the steady state, and $\bar{W}_1(s) = E\left[e^{-sW_1}\right]$, then because $\alpha(s) + \beta_o = E\left[s^N\right]$ where N is the number of 1-units waiting at a departure epoch and $q_i = P\left[\text{last unit served was an } i\text{-unit}\right]$, it follows that

$$\alpha(s) + \beta_o = q_1\, E\left[s^{N_1}\right] + q_2\, E\left[s^{N_2}\right]$$

$$= q_1\, \bar{W}_1(\lambda_1-\lambda_1 s)\, \bar{S}_1(\lambda_1-\lambda_1 s) + q_2\, \bar{S}_2(\lambda_1-\lambda_1 s)$$

where N_1 is the number of 1-units arriving in $W_i + S_1$ and N_2 is the number in S_2. Thus

$$\bar{W}_1(s) = \frac{\alpha(1-s/\lambda_1)+\beta_o-q_2\bar{S}_2(s)}{q_1\bar{S}_1(s)}$$

Substituting for $\alpha(s)$:-

$$\bar{W}_1(s) = \frac{(1-\rho_1-\rho_2)s+\lambda_2\left[1-\bar{S}_2(s)\right]}{s+\lambda_1\left[\bar{S}_1(s)-1\right]} \qquad (7)$$

Therefore

$$E\left[W_1\right] = \frac{\lambda_1 E\left[S_1{}^2\right]+\lambda_2 E\left[S_2{}^2\right]}{2(1-\rho_1)} \qquad (8)$$

More General Arrival Scheme

Suppose each unit arriving for service has a priority number Q: a lower value of Q denoting non-preemptive priority over a higher value. Arrivals occur at random with rate λ and have priority numbers which are i.i.d. positive random variables, independent of the arrival times. Let $Q(y) = P(\text{arrival has priority number} \leq y)$. Service times of arrivals with $Q = y$ are i.i.d. random variables with distribution function $S_y(x)$, mean $1/\mu_y$, Laplace-Stieltjes transform $\bar{S}_y(s)$ and

$$\lambda \int_o^\infty (1/\mu_x)\, dQ(x) < 1 \qquad \text{(for stability)}.$$

Group all units into two classes:

those with $Q \leq y$ called 1-units

those with $Q > y$ called 2-units

Then by equations (7) and (8):-

$$\bar{W}_1(s) = \frac{\left[1-\lambda \int_0^\infty (1/\mu_x)dQ(x)\right]s+\lambda\left[\int_y^\infty dQ(x)-\int_y^\infty \bar{S}_x(s)dQ(x)\right]}{s-\lambda\left[\int_0^y dQ(x)-\int_0^y \bar{S}_x(s)dQ(x)\right]} \tag{9}$$

and

$$E[W_1] = \frac{\lambda \int_0^\infty E[S_x^2]dQ(x)}{2\left[1-\lambda \int_0^y E[S_x]dQ(x)\right]} \tag{10}$$

When a unit with Q = y arrives, its waiting time can be decomposed into two parts:

(i) the time to serve any unit already in service and all units already in the system with $Q \leq y$ - this corresponds to W_1, and

(ii) the time to serve all units with $Q < y$ which arrive before its entry into service.

Therefore W(y), the stationary waiting time of a unit with Q = y, has the same distribution as the length of the busy period initiated by a waiting time W_1 in an ordinary queue (i.e. without a priority discipline) composed of units with $Q < y$. Thus

$$E(e^{-sW(y)}) = \bar{W}_1(s+\lambda \int_0^{y-0} dQ(x)\left[1 - \bar{B}_y(s)\right]) \tag{11}$$

where $\bar{B}_y(s)$ is the Laplace-Stieltjes transform of the length of an ordinary busy period in a queue composed of units with $Q < y$

i.e. $\bar{B}_y(s)$ is the root with smallest absolute value in z of the equation

$$z = \frac{\int_{x=0}^{y-0} \bar{S}_x\left[s+\lambda(1-z) \int_0^{y-0} dQ(x)\right] dQ(x)}{\int_{x=0}^{y-0} dQ(x)}$$

- the equation is $z = \bar{S}[s + \lambda(1-z)]$ for an ordinary M/G/1 queue.

Equations (9) and (11) determine $E[\exp(-sW(y))]$; the moments can be obtained by differentiation:

$$E[W(y)] = \frac{E[W_1]}{1-\lambda \int_0^{y-0} E[S_y]\, dQ(y)}$$

$$= \frac{\lambda \int_0^{\infty} E[S_x^2]\, dQ(x)}{2\left[1-\lambda \int_0^{y-0} E[S_x]\, dQ(x)\right]\left[1-\lambda \int_0^{y} E[S_x]\, dQ(x)\right]} \qquad (12)$$

Two Special Cases

(i) Units are of k different priority types, q_j being the probability that an arrival has priority number $j(j=1,2,\ldots,k)$. For this case (12) gives the mean waiting time of a j-unit as

$$E[W_j] = \frac{\lambda \sum_{i=1}^{k} q_i E[S_i^2]}{2\left[1-\lambda \sum_{i=1}^{j-1} q_i E[S_i]\right]\left[1-\lambda \sum_{i=1}^{j} q_i E[S_i]\right]}$$

$$= \frac{\sum_{i=1}^{k} \lambda_i E[S_i^2]}{2\left[1-\sum_{i=1}^{j-1} \rho_i\right]\left[1-\sum_{i=1}^{j} \rho_i\right]} \qquad (13)$$

where $\lambda_i = q_i\lambda$ and $\rho_i = \lambda_i E[S_i]$. $\qquad i = 1,2,\ldots,k$

(ii) Service times of all arrivals are i.i.d. random variables, distribution function $S(x)$, $x \geq 0$. The priority number Q is identical with the service time S: i.e. a customer with shorter service time has non-preemptive priority over a customer with longer service time.

i.e. $Q(y) = S(y)$ and $S_y(x) = \begin{cases} 1 & x \geq y \\ 0 & x < y \end{cases}$

Therefore

$$E[S_y] = y \quad \text{and} \quad E[S_y^2] = y^2$$

Equation (12) gives, provided y is a continuity point of S(x),

$$E[W(y)] = \frac{\lambda \int_0^\infty x^2 \, dS(x)}{2[1-\lambda \int_0^y x \, dS(x)]^2} \tag{14}$$

Thus the mean queueing time taken over all arrivals is:

$$E[W] = \lambda \int_0^\infty x^2 \, dS(x) \int_0^\infty \frac{dS(y)}{2[1-\lambda \int_0^y x \, dS(x)]^2}$$

Both this equation and equation (13) can be found in Cox and Smith [5], pages 83 and 85.

1.2 Optimal Priority Classification in the M/G/1 Queue

Suppose that arrivals requiring service from a single server occur at random with rate λ and have service times, S, which are i.i.d. random variables with distribution function S(x), $x \geq 0$, and mean $1/\mu$ where $\rho = \lambda/\mu < 1$. If every customer is to be assigned to one of two non-preemptive priority classes the problem arises of how to allocate priorities to arrivals on the basis of their service time requirements so as to minimize the overall mean waiting time.

Suppose a rule for allocating priorities is specified so that the priority function

$$P(x) = P[\text{classifying a customer into class 1} \mid S = x]$$

can be defined, where this probability is independent of the classification of all previous arrivals. Let $S_j(x)$ be the distribution function of the service time of a j-unit corresponding to this rule (j=1,2). Then, by Bayes' Theorem for continuous distributions,

$$S_1(x) = \frac{\int_0^x P(u)\, dS(u)}{\int_0^\infty P(u)\, dS(u)} \quad \text{and} \quad S_2(x) = \frac{\int_0^x [1-P(u)]\, dS(u)}{1-\int_0^\infty P(u)\, dS(u)}$$

(where $x \geq 0$)

Also, customers of each class arrive at random with rates

$$\lambda_1 = \lambda \int_0^\infty P(u)\, dS(u) \qquad \text{for class 1}$$

and $$\lambda_2 = \lambda - \lambda \int_0^\infty P(u)\, dS(u) \qquad \text{for class 2}$$

Equation (13) gives the overall mean waiting time

$$E[W] = \frac{E[S^2]}{2}\left[\frac{\lambda_1}{1-\rho_1} + \frac{\lambda_2}{(1-\rho_1)(1-\rho_1-\rho_2)}\right]$$

$$= \frac{\lambda E[S^2]\left[(1-\rho)\int_0^\infty P(u)\, dS(u) + 1 - \int_0^\infty P(u)\, dS(u)\right]}{2(1-\rho)(1-\lambda \int_0^\infty u\, P(u)\, dS(u))}$$

$$= \frac{\lambda E[S^2]}{2(1-\rho)} \cdot R$$

where

$$R = \frac{1-\rho \int_0^\infty P(u)\, dS(u)}{1-\lambda \int_0^\infty u\, P(u)\, dS(u)} \qquad (15)$$

Note that

(i) R is the reduction factor:

if $P(x) \equiv 0$ or $P(x) \equiv 1$ then $R = 1$ and

$$E[W] = \frac{\lambda E[S^2]}{2(1-\rho)}$$

This corresponds to an ordinary M/G/1 queue with all customers in a single class and the first-come first-served discipline.

(ii) The case

$$P(x) = \begin{cases} 1 & 0 \leq x \leq \phi\rho \\ 0 & \text{otherwise} \end{cases}, \quad S(x) = 1 - e^{-\mu x}$$

$x \geq 0$, where ϕ is a non-negative constant, is considered in Cox and Smith [5], page 86, and it will be shown that this priority function (which corresponds to putting all customers with service times less than some constant into the priority class) is the optimum form for $P(x)$ for any distribution of service times.

(iii) Allocating priorities in this way leads to an improvement in the mean waiting time if and only if $R < 1$,

$$\text{i.e. iff} \qquad \frac{1}{\mu} > \frac{\int_0^\infty u\, P(u)\, dS(u)}{\int_0^\infty P(u)\, dS(u)}$$

i.e. iff the overall mean service time > mean service time of 1-units.

The truth of the statement in note (ii) will now be established:- Given any continuous priority function $G(x)$, $(x \geq 0,\ 0 \leq G(x) \leq 1)$, it seems intuitively reasonable to suppose there exists an optimum cutoff c_G such that the priority function

$$P(x) = \begin{cases} G(x) & 0 \leq x \leq c_G \\ 0 & \text{otherwise} \end{cases}$$

minimizes R. For any cutoff c,

$$R = R(c) = \frac{1-\rho \int_0^c G(u)\, dS(u)}{1-\lambda \int_0^c u\, G(u)\, dS(u)} \tag{16}$$

Setting $dR/dc = 0$ gives an equation for c_G:-

$$\left[1 - \lambda \int_0^{c_G} u\, G(u)\, dS(u)\right] \left[-\rho\, G(c_G)\, S'(c_G)\right]$$

$$= \left[1 - \rho \int_0^{c_G} G(u)\, dS(u)\right] \left[-\lambda c_G\, G(c_G)\, S'(c_G)\right]$$

Hence, if $G(c_G)\, S'(c_G) \neq 0$,

$$\frac{\mu c_G - 1}{\lambda} = c_G \int_0^{c_G} G(u)\, dS(u) - \int_0^{c_G} u\, G(u)\, dS(u)$$

$$= c_G \int_0^{c_G} G(u)\, dS(u) - c_G \int_0^{c_G} G(u)\, dS(u)$$

$$+ \int_0^{c_G} \int_0^{x} G(u)\, dS(u)\, dx$$

$$= \int_0^{c_G} \int_0^{x} G(u)\, dS(u)\, dx \qquad (17)$$

It follows that $c_G \geq 1/\mu$.

From (17),

$$\mu c_G \left[1 - \rho \int_0^{c_G} G(u)\, dS(u)\right] = 1 - \lambda \int_0^{c_G} u\, G(u)\, dS(u)$$

and hence, if the optimum R for this priority function G(x) is denoted by R_G,

$$R_G = \frac{1}{\mu c_G} \qquad (18)$$

and this can easily be seen to be a minimum value of R by considering the behaviour of the second derivative. Equation (18) implies that the optimal function for P(x) is one which maximizes c_G, the optimum cutoff.

Consider the function $H(x) \equiv 1$. The optimum cutoff c_H for this function satisfies the equation

$$\frac{c_H \mu - 1}{\lambda} = c_H \int_0^{c_H} dS(u) - \int_0^{c_H} u\, dS(u) \qquad (19)$$

Then if K(x) is any other possible priority function with optimum cutoff c_K,

$$\frac{c_K \mu - 1}{\lambda} = c_K \int_0^{c_K} K(u)\, dS(u) - \int_0^{c_K} u\, K(u)\, dS(u) \qquad (20)$$

Subtracting equation (19) and (20) gives

$$\frac{\mu(c_K-c_H)}{\lambda} = (c_K-c_H)\int_0^{c_K} K(u)\,dS(u) + \int_0^{c_K} (c_H-u)\,K(u)\,dS(u)$$

$$+ \int_0^{c_H} (u-c_H)\,dS(u)$$

Therefore

$$(c_K-c_H)\left(\frac{1}{\rho} - \int_0^{c_K} K(u)\,dS(u)\right) = \int_0^{c_K} (c_H-u)\,K(u)\,dS(u)$$

$$+ \int_0^{c_H} (u-c_H)\,dS(u) \qquad (21)$$

Suppose $c_K > c_H$, then the left-hand side is strictly positive whilst the right-hand side is equal to

$$\int_0^{c_H} (c_H-u)\,\left[K(u) - 1\right]\,dS(u) + \int_{c_H}^{c_K} (c_H-u)\,K(u)\,dS(u)$$

which is negative: a contradiction. It follows that for any other possible priority function $K(x)$, $c_K \leq c_H$. The optimum priority function must be therefore,

$$P(x) = \begin{cases} 1 & 0 \leq x \leq \bar{c} \\ 0 & \text{otherwise} \end{cases}$$

where $\bar{c}$ satisfies the equation

$$\frac{\bar{c}\mu-1}{\lambda} = \bar{c}\int_0^{\bar{c}} dS(u) - \int_0^{\bar{c}} u\,dS(u) \qquad (22)$$

Note: For exponential service times $S(x) = 1 - e^{-\mu x}$, $x \geq 0$ and equation (22) becomes

$$\bar{c}\mu - 1 = \frac{\rho\, e^{-\mu\bar{c}}}{1-\rho} \qquad (23)$$

At the optimum, the mean waiting time equals

$$E[W] = \frac{\lambda E[S^2]}{2(1-\rho)} R_G = \frac{\rho}{(1-\rho)\bar{c}\,\mu^2} \qquad (24)$$

This is the special case considered in Cox and Smith [5].

1.3 Arrivals with Estimated Service Times

In most practical situations it is extremely unlikely that the service time of a customer can be exactly predicted on his arrival, and so the optimum rule of the last section cannot be implemented. However it is quite plausible that on arrival an estimate Y can be made of a customer's service time: the effect of having only this imperfect information will now be considered.

The model is assumed to be the same as before: arrivals at random with rate λ, service time p.d.f. $f(x)$, $x \geq 0$ with mean $1/\mu$ and $\rho = \lambda/\mu < 1$. Suppose that all customers with estimates $Y \leq c$ are given non-preemptive priority over all other customers. Then

$$P(x) = P[Y \leq c \mid x]$$

$$= \int_{-\infty}^{c} f_{Y|S}(y;x)\, dy$$

where $f_{Y|S}(y;x)$ is the conditional distribution of the estimate given the true service time S. Equation (15) gives

$$R = \frac{1-\rho \int_0^\infty f(x)\, P(x)\, dx}{1-\lambda \int_0^\infty x\, f(x)\, P(x)\, dx}$$

$$= \frac{1-\rho \int_{-\infty}^{c} \int_0^\infty f_{SY}(x,y)\, dx\, dy}{1-\lambda \int_{-\infty}^{c} \int_0^\infty x\, f_{SY}(x,y)\, dx\, dy} \qquad (25)$$

i.e. $$R(c) = \frac{1-\rho \int_{-\infty}^{c} f_Y(y)\, dy}{1-\lambda \int_{-\infty}^{c} E[S|y]\, f_Y(y)\, dy} \qquad (26)$$

By differentiation, the optimum cutoff $\bar{c}$ satisfies the equation

$$\left[1 - \lambda \int_{-\infty}^{\bar{c}} \int_{0}^{\infty} x \, f_{SY}(x,y) \, dx \, dy\right] \left[-\rho \int_{0}^{\infty} f_{SY}(x,\bar{c}) \, dx\right]$$

$$= \left[1 - \rho \int_{-\infty}^{\bar{c}} \int_{0}^{\infty} f_{SY}(x,y) \, dx \, dy\right] \left[-\lambda \int_{0}^{\infty} x \, f_{SY}(x,\bar{c}) \, dx\right]$$

and therefore the minimum value of R(c) is given by

$$\bar{R} = \frac{\rho \int_{0}^{\infty} f_{SY}(x,\bar{c}) \, dx}{\lambda \int_{0}^{\infty} x \, f_{SY}(x,\bar{c}) \, dx}$$

$$= \frac{E[S]}{E[S|Y=\bar{c}]} \qquad \text{c.f. equation (18)}$$

A sufficient condition for a reduction in the mean waiting time by this method of giving non-preemptive priority to certain customers on the basis of their estimated service time requirements can easily be found:

<u>If $E[S|y]$ increases monotonically with y then</u>

(a) <u>$R \leq 1$ for any c</u>

(b) <u>R has a unique minimum provided $E[S|y] \neq 1/\mu$</u>

<u>Proof of (a):-</u>

As $E[S|y]$ increases monotonically with y, for any cutoff c

$$\int_{-\infty}^{+\infty} E[S|y] \, f_Y(y) \, dy - \frac{\int_{-\infty}^{c} E[S|y] \, f_Y(y) \, dy}{\int_{-\infty}^{c} f_Y(y) \, dy}$$

$$\geq E[S|c] \, P[Y > c] + \left[1 - \frac{1}{\int_{-\infty}^{c} f_Y(y) \, dy}\right] \int_{-\infty}^{c} E[S|y] \, f_Y(y) \, dy$$

$$= E[S|c] \, P[Y > c] - \frac{P[Y>c]}{P[Y \leq c]} \int_{-\infty}^{c} E[S|y] \, f_Y(y) \, dy$$

$$= P[Y > c] \left\{ E[S|c] - \frac{E[S|y] \, F(y)\Big|_{-\infty}^{c}}{P[Y \leq c]} + \right.$$

$$\left. + \frac{1}{P[Y \leq c]} \int_{-\infty}^{c} \frac{d}{dy} E[S|y] \, F(y) \, dy \right\}$$

$$= \frac{P[Y>c]}{P[Y\leqslant c]} \int_{-\infty}^{c} \frac{d}{dy} E[S|y]\, F(y)\, dy > 0$$

where $F(y) = \int_{-\infty}^{y} f_Y(u)\, du$

As $\frac{1}{\mu} = \int_{-\infty}^{+\infty} E[S|y]\, f_Y(y)\, dy$, it follows that

$$R(c) = \frac{1-\rho \int_{-\infty}^{c} f_Y(y)\, dy}{1-\lambda \int_{-\infty}^{c} E[S|y]\, f_Y(y)\, dy} \leqslant 1$$

Proof of (b):-

From equation (26)

$$\frac{dR}{dc} = \frac{f_Y(c)}{\{1-\lambda \int_{-\infty}^{c} f_Y(y)\, E[S|y]dy\}^2} \{-\rho + \lambda\rho \int_{-\infty}^{c} f_Y(y)\, E[S|y]dy$$

$$+ \lambda E[S|c] - \lambda\rho E[S|c]\, P[Y \leqslant c]\}$$

$$= \frac{f_Y(c)}{\{1-\lambda \int_{-\infty}^{c} f_Y(y)\, E[S|y]dy\}^2} \cdot H(c) \qquad \text{say.}$$

At any root of dR/dc, $H(c) = 0$.

Also for any c,

$$\frac{dH(c)}{dc} = \lambda \{1 - \rho\, P[Y \leqslant c]\} \frac{d}{dc} E[S|c] > 0 \quad \text{as } \rho < 1$$

Therefore dR/dc has at most one root. Either $E[S|c] = 1/\mu$ for all c, in which case $R(c) \equiv 1$ and $H(c) \equiv 0$, or for c sufficiently large negative $E[S|c] < 1/\mu$; but

$$\int_{-\infty}^{c} f_Y(y)\, E[S|y]dy < E[S|c]\, P[Y \leqslant c]$$

and therefore

$$dR/dc < 0$$

Now

$$H(c) = \lambda E[S|c] - \rho + \rho^2 - \lambda\rho \{E[S|c] P[Y \leq c] + \int_c^\infty f_Y(y) E[S|y]dy\}$$

$$\geq \{\lambda E[S|c] - \rho\} \{1 - \rho\} - \lambda\rho \int_c^\infty f_Y(y) E[S|y]dy \qquad (27)$$

For $E[S|y] \neq 1/\mu$, there exists a c' sufficiently large such that $E[S|c'] > \frac{1}{\mu}$ and then

$$\{\lambda E[S|c'] - \rho\} \{1 - \rho\} > 0$$

The last term on the right-hand side of (27) can be made as small as desired. Thus, for sufficiently large c, $H(c) > 0$ and hence $dR/dc > 0$ (provided $E[S|y] \neq 1/\mu$). The function $R(c)$ has therefore exactly one turning point which must be a point of minimum value.

1.4 Examples

a) Exponential Service Times Estimated with a Normally Distributed Error

Suppose that service times are independently and identically distributed random variables with distribution function $S(x) = 1 - e^{-\mu x}$, $x \geq 0$ and that the estimate Y of the service time requirement of a unit made on its arrival is equal to $S + Z$, where

S is the true service time of the unit

S and Z are independent

and Z is distributed as $N(0,\sigma^2)$

Then

$$f_{Y|S}(y;x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{(y-x)^2}{2\sigma^2}\right\}$$

and

$$f_{SY}(x,y) = \frac{\mu}{\sqrt{2\pi}\,\sigma} \exp(-\mu x) \exp\left\{-\frac{(y-x)^2}{2\sigma^2}\right\}$$

Using this distribution gives

$$\int_{-\infty}^{c}\int_{0}^{\infty} f_{SY}(x,y)\,dx\,dy = \int_{0}^{\infty} \mu \exp(-\mu x) \int_{-\infty}^{c} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{(y-x)^2}{2\sigma^2}\right\} dy\,dx$$

$$= \int_{0}^{\infty} \mu \exp(-\mu x) \int_{-\infty}^{(c-x)/\sigma} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{y^2}{2}\right\} dy\,dx$$

$$= \int_{-\infty}^{c/\sigma} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{y^2}{2}\right\} \int_{0}^{c-\sigma y} \mu \exp(-\mu x)\,dx\,dy$$

$$= \int_{-\infty}^{c/\sigma} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{y^2}{2}\right\} \{1 - \exp(-\mu c+\mu\sigma y)\}\,dy$$

$$= \Phi(c/\sigma) - \exp\left\{\frac{\mu^2\sigma^2}{2} - \mu c\right\} \Phi(c/\sigma-\mu\sigma) \qquad (28)$$

where

$$\Phi(y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{t^2}{2}\right\} dt$$

Similarly

$$\int_{-\infty}^{c}\int_{0}^{\infty} x\, f_{SY}(x,y)\,dx\,dy = \frac{1}{\mu}\,\Phi\left(\frac{c}{\sigma}\right) + \left(\mu\sigma^2-c-\frac{1}{\mu}\right) \exp\left\{\frac{\mu^2\sigma^2}{2} - \mu c\right\} \Phi\left(\frac{c}{\sigma}-\mu\sigma\right)$$

$$- \frac{\sigma}{\sqrt{2\pi}} \exp\left\{-\frac{c^2}{2\sigma^2}\right\} \qquad (29)$$

Substituting in (25) gives

$$R(c) = \frac{\mu-\lambda\,\Phi\left(\frac{c}{\sigma}\right)+\lambda \exp\left\{\frac{\mu^2\sigma^2}{2}-\mu c\right\} \Phi\left(\frac{c}{\sigma}-\mu\sigma\right)}{\mu-\lambda\,\Phi\left(\frac{c}{\sigma}\right)-\lambda(\mu^2\sigma^2-\mu c-1)\exp\left\{\frac{\mu^2\sigma^2}{2}-\mu c\right\} \Phi\left(\frac{c}{\sigma}-\mu\sigma\right) + \frac{\lambda\mu\sigma}{\sqrt{2\pi}} \exp\left\{-\frac{c^2}{2\sigma^2}\right\}}$$

Setting $dR/dc = 0$ gives an equation for the optimum cutoff $\bar{c}$:-

$$0 = \lambda\{\Phi(\bar{c}/\sigma-\mu\sigma)\}^2 + \{\mu - \lambda\Phi(\bar{c}/\sigma)\} \{(1-\mu\bar{c}+\mu^2\sigma^2)\,\Phi(\bar{c}/\sigma-\mu\sigma)$$

$$- \frac{\mu\sigma}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}(\bar{c}/\sigma-\mu\sigma)^2)\} \exp(\mu c-\mu^2\sigma^2/2)$$

Without loss of generality, λ can be taken to be unity. Numerical values of $\bar{c}$ and the minimum mean waiting time $\bar{W}$ for various values of $1/\mu$ (the traffic intensity) and σ are given in Table 1. (The special case $\sigma = 0$ is covered in Cox and Smith [5]). From Table 1 it can be seen that quite considerable reductions in the mean waiting time are possible when the traffic intensity is large when even very rough estimation is valuable. Clearly for any particular problem, costs could be assigned to the method of estimation and to the waiting time of customers and the optimum σ to minimize costs obtained. It is perhaps worth mentioning that the restriction to unbiased estimates is unnecessary: a monotonic function of Y would yield the same minimum value of R.

TABLE 1

The lower value in each classification is the minimum mean waiting time; the upper value is the optimum cutoff, $\bar{c}$. Analogous tables for the various preemptive disciplines could also be constructed.

ρ	first come first-served	Non-preemptive priority: 2 priority classes					
		$\sigma=0$	$\sigma=0.01$	$\sigma=0.1$	$\sigma=0.3$	$\sigma=0.5$	$\sigma=1.0$
0.3	0.129	0.34 0.113	0.34 0.113	0.37 0.114	0.50 0.117	0.59 0.120	0.71 0.124
0.5	0.5	0.64 0.391	0.64 0.391	0.66 0.392	0.79 0.402	0.93 0.416	1.19 0.443
0.7	1.633	1.06 1.079	1.06 1.079	1.07 1.082	1.17 1.101	1.33 1.139	1.73 1.248
0.9	8.1	1.89 3.855	1.89 3.855	1.90 3.861	1.96 3.909	2.10 4.007	2.61 4.447
0.95	18.05	2.40 7.153	2.40 7.153	2.40 7.162	2.46 7.239	2.58 7.395	3.10 8.160
0.99	98.01	3.59 27.010	3.59 27.010	3.60 27.038	3.65 27.259	3.75 27.712	4.24 30.0

(b) A Bivariate Exponential Distribution for the Service Time and Estimate

In Downton [6] a bivariate exponential distribution (S,Y) which seems suitable for the joint distribution of the service time and estimate is considered. Some of its properties are as follows:

(i)

$$\psi(s_1, s_2) = E[e^{-s_1 S - s_2 Y}] = \frac{\mu \mu'}{(\mu+s_1)(\mu'+s_2)-\alpha s_1 s_2}$$

where μ, μ' are strictly positive and α, the correlation coefficient, is restricted to $0 \leq \alpha \leq 1$.

(ii) the marginal distributions of S and Y are $\mu e^{-\mu x}$ $(x \geq 0)$ and $\mu' e^{-\mu' y}$ $(y \geq 0)$ respectively.

(iii)

$$E[S|y] = \frac{1-\alpha}{\mu} + \frac{\alpha \mu'}{\mu} y \qquad \text{i.e. } E[S|y] \text{ increases}$$

monotonically with y.

(iv)

$$\text{var}[Y|x] = \frac{1-\alpha}{\mu'} \left[\frac{1-\alpha}{\mu'} + \frac{2\alpha\mu x}{\mu'}\right]$$

i.e. var[Y|x] increases with x, and in this respect this distribution is more realistic than that considered in example (a).

Substituting in equation (26) gives

$$R(c) = \frac{1-\rho+\rho \exp(-\mu' c)}{1-\rho+\rho \exp(-\mu' c)+c\rho\alpha \mu' \exp(-\mu' c)}$$

Setting $dR/dc = 0$ gives an equation for $\bar{c}$:

$$\rho \exp(-\mu' \bar{c}) = (1-\rho)(\mu' \bar{c}-1) \qquad (30)$$

This is a natural two dimensional generalisation of the equation in Cox and Smith [5], page 86 (and given as equation (23) above) for the case of perfect information.

It follows from (30) that the optimum cutoff $\bar{c}$ is independent of the correlation α, and the minimum reduction factor is $1/(1-\alpha+\alpha\,\mu'\bar{c})$.
For $\alpha = 1$, this reduces to $1/\mu'c$ c.f. equation (18).
For $\alpha = 0$, it takes the value 1 - the estimate Y then giving no information about the value of S.

1.5 An Infinite Number of Non-Preemptive Priority Classes

Proceeding as in Section 1.3, suppose arrivals occur at random with rate λ and have service times which are independently and identically distributed random variables with p.d.f. $f(x)$, $x \geq 0$ and mean $1/\mu$, where $\rho = \lambda/\mu < 1$. On arrival an estimate Y is made of the service time requirement of a unit, and at every departure epoch the server selects for service that unit waiting in the queue with the lowest estimate. Let $f_{SY}(x,y)$ $(x\geq 0,\ -\infty<y<+\infty)$ denote the joint distribution of the service time S and estimate Y. Then

$$Q(y') = P[\text{arrival has estimate} \leq y'] = \int_{-\infty}^{y'} \int_{0}^{\infty} f_{SY}(x,y)\,dx\,dy = \int_{-\infty}^{y'} f_Y(y)\,dy$$

$$S_y(x') = P[S \leq x' \mid Y = y] = \int_{0}^{x'} f_{S|Y}(x,y)\,dx$$

$$E[S_y^{\,2}] = \int_{0}^{\infty} x^2 f_{S|Y}(x;y)\,dx$$

$$\int_{-\infty}^{+\infty} E[S_y^{\,2}]\,dQ(y) = \int_{-\infty}^{+\infty}\int_{0}^{\infty} x^2 f_{SY}(x,y)\,dx\,dy$$

Thus equation (12) becomes

$$E[W(y')] = \frac{\lambda \int_{-\infty}^{+\infty}\int_{0}^{\infty} x^2 f_{SY}(x,y)\,dx\,dy}{2\left[1-\lambda \int_{-\infty}^{y'-0}\int_{0}^{\infty} x\, f_{SY}(x,y)dxdy\right]\left[1-\lambda \int_{-\infty}^{y'}\int_{0}^{\infty} x\, f_{SY}(x,y)dxdy\right]}$$

Provided y' is a continuity point of $F(y) = \int_{-\infty}^{y} f_Y(u)\,du$,

$$E[W(y')] = \frac{\lambda\, E[S^2]}{2\left[1-\lambda \int_{-\infty}^{y'}\int_{0}^{\infty} x\, f_{SY}(x,y)\,dx\,dy\right]^2} \qquad (31)$$

The overall mean waiting time is, therefore

$$E[W] = \int_{-\infty}^{+\infty} E[W(y')]\, f_Y(y')\,dy$$

$$= \frac{\lambda\, E[S^2]}{2} \int_{-\infty}^{+\infty} \frac{\int_{0}^{\infty} f_{SY}(x,y')\,dx}{\left[1-\lambda \int_{0}^{\infty}\int_{-\infty}^{y'} x\, f_{SY}(x,y)\,dy\,dx\right]^2}\,dy' \qquad (32)$$

For the case of perfect information $f_{SY}(x,y) = \begin{cases} f(x) & y = x \\ 0 & \text{otherwise} \end{cases}$ and (31) reduces to (14).

Numerical results for either of the two joint distributions discussed in the last section could be calculated. In the case of example (a), (32) becomes

$$E[W] = \rho \int_{-\infty}^{+\infty} \frac{\exp\left(\frac{\mu^2\sigma^2}{2} - \mu z\right)\, \Phi\left(\frac{z}{\sigma} - \mu\sigma\right)\, dz}{\left[1-\rho\, \Phi(z/\sigma) - \lambda(\mu\sigma^2 - 1/\mu - z)\exp\left(\frac{\mu^2\sigma^2}{2} - \mu z\right)\Phi\left(\frac{z}{\sigma} - \mu\sigma\right) + \frac{\lambda\sigma}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2\sigma^2}\right)\right]^2}$$

For $\sigma = 0$ this has been tabulated by Schrage and Miller [16]. Values for various non-zero values of σ are given in Table 2. Without loss of generality λ has been set equal to unity, the traffic intensity ρ then being $1/\mu$.

TABLE 2

Values of the mean waiting time for a continuous number of non-preemptive priority classes; priority classification is on the basis of an estimate with normal error. The format is the same as for Table 1 for two priority classes with which it should be compared.

ρ \ σ	0	0.01	0.1	0.3	0.5	1.0
0.3	0.108	0.108	0.109	0.113	0.117	0.121
0.5	0.356	0.356	0.359	0.372	0.388	0.421
0.7	0.919	0.919	0.923	0.951	0.994	1.116
0.9	2.877	2.877	2.886	2.947	3.057	3.482
0.95	5.001	5.001	5.013	5.099	5.259	5.945
0.99	17.276	17.276	17.300	17.483	17.841	19.553

CHAPTER 2

Discretionary Queues

2.1 Basic Results

The discretionary discipline was first introduced by Avi-Itzhak, Brosh and Naor [1] who gave a solution for the case of constant service times; the first published solution for general service times for the special case of early preemption is in Jaiswal [8]. By considering the general stochastic process as a sequence of alternating busy and idle periods, Jaiswal derives all the main properties of the process including the Laplace-Stieltjes transforms of the busy period distribution, the waiting time distribution and the transient generating function of queue length probabilities. Since the work for this chapter was completed, Balachandran [2] has given a method for deriving the mean waiting time of units in more general discretionary queues. A simple method will be given in this section for deriving the Laplace-Stieltjes transform of the waiting time of units in such queues and this is extended in Section 2.4 to derive properties of a discretionary queue based on estimated remaining service time.

Suppose all arrivals belong to one of two types and that type i-units (i=1,2) arrive at random with rate λ_i and have service times which are i.i.d. random variables with distribution function $S_i(x)$, $x \geq 0$, mean $1/\mu_i$. Let $\rho_i = \lambda_i/\mu_i$ where $\rho_1 + \rho_2 < 1$. If $E(S_2-z|S_2>z)$ regarded as a function of z has the shape shown in Figure 1 (which could happen if S_2 is the mixture of two distributions with differing means for example), then the server may conform to the following discretionary rule with regard to a 1-unit arriving

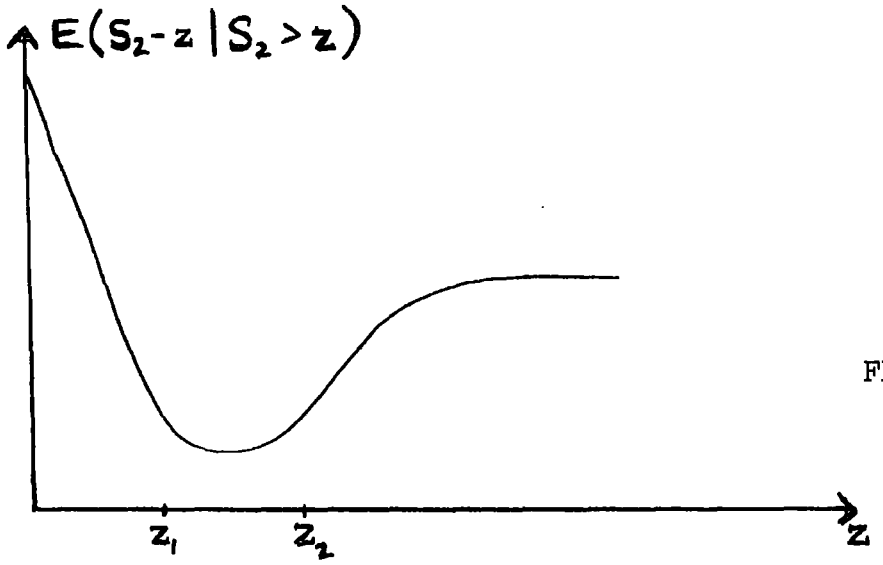


FIGURE 1

to find a 2-unit in service:-

if $S_2' \leq z_1$ or $S_2' > z_2$ the 1-unit preempts the 2-unit from service, the 2-unit resuming its service when there are no 1-units left in the system.

if $z_1 < S_2' \leq z_2$ the non-preemptive rule is followed. (Where S_2' denotes the time the 2-unit has already spent in service when the 1-unit arrives).

The discretionary rule is thus a mixture of the non-preemptive and preemptive resume priority disciplines.

Waiting Time of 1-units

Suppose a 2-unit arrives at time τ and has service time S_2; then without affecting W_1, the waiting time of 1-units,

if $S_2 \leq z_1$ the 2-unit can be ignored (because any 1-unit arriving would preempt it from service).

if $z_1 < S_2 \leq z_2$ the 2-unit can be replaced by an ordinary non-preemptive 2-unit arriving at time $\tau + z_1$ and having service time $S_2 - z_1$.

if $z_2 < S_2$ the 2-unit can be replaced by an ordinary non-

preemptive 2-unit arriving at time $\tau + z_1$ and having service time $z_2 - z_1$.

Thus W_1 can be evaluated from the results for an ordinary two-class non-preemptive queue in which:

1-units arrive at random with rate λ_1 and have service times which are i.i.d. random variables with distribution function $S_i(x)$, $x \geq 0$.

2-units arrive at random with rate $\lambda_2 P[S_2 > z_1]$ and have service times T_2 which are i.i.d. random variables with distribution function $T_2(y)$ given by

$$T_2(y) = \frac{S_2(y+z_1)-S_2(z_1)}{P[S_2>z_1]} \qquad 0 < y \leq z_2 - z_1$$

with the remaining probability $\dfrac{P[S_2>z_2]}{P[S_2>z_1]}$ in a spike at $T_2 = z_2 - z_1$.

Therefore

$$E[e^{-sT_2}] = \frac{e^{sz_1} \int_{z_1}^{z_2} e^{-sx}\, dS_2(x)}{P[S_2>z_1]} + \frac{e^{-s(z_2-z_1)}\, P[S_2>z_2]}{P[S_2>z_1]} \tag{1}$$

and

$$E[T_2] = \frac{\int_{z_1}^{z_2} (x-z_1)\, dS_2(x)+(z_2-z_1)\, P[S_2>z_2]}{P[S_2>z_1]} \tag{2}$$

The traffic intensity for units in this modified queue is

$$\lambda_2 . P[S_2 > z_1] . E[T_2] = \lambda_2 \int_{z_1}^{z_2} (x-z_1)\, dS_2(x) + \lambda_2 (z_2-z_1)\, P[S_2 > z_2]$$

Therefore, by the usual formula for the Laplace-Stieltjes transform of the waiting time of 1-units in a 2-class non-preemptive queue (see equation (7) of Chapter 1)

$$\{s + \lambda_1 \bar{S}_1(s) - \lambda_1\} E[e^{-sW_1}] =$$

$$\{1-\rho_1-\lambda_2 \int_{z_1}^{z_2} (x-z_1)dS_2(x)-\lambda_2(z_2-z_1)P[S_2>z_2]\}s+\lambda_2\{P[S_2>z_1]-e^{sz_1} \int_{z_1}^{z_2} e^{-sx}dS_2(x)$$

$$- e^{-s(z_2-z_1)} P[S_2 > z_2]\} \qquad (4)$$

and

$$E[W_1] = \frac{\lambda_1 E[S_1^2]+\lambda_2 \int_{z_1}^{z_2} (x-z_1)^2 \, dS_2(x)+\lambda_2(z_2-z_1)^2 \, P[S_2>z_2]}{2(1-\rho_1)} \qquad (5)$$

Waiting and Completion Time of 2-units

Clearly the waiting time of a 2-unit W_2 in a discretionary queue is the same as in other priority queues in which the priority discipline does not affect the length of busy periods (i.e. non-preemptive and preemptive resume). Therefore

$$E[e^{-sW_2}] = \frac{(1-\rho_1-\rho_2)(s+\lambda_1-\lambda_1 \bar{B}_1(s))}{s+\lambda_1-\lambda_1 \bar{B}_1(s)-\sum_{i=1}^{2} \lambda_i[1-\bar{S}_i(s+\lambda_1-\lambda_1 \bar{B}_1(s))]} \qquad (6)$$

and

$$E[W_2] = \frac{\lambda_1 E[S_1^2]+\lambda_2 E[S_2^2]}{2(1-\rho_1)(1-\rho_1-\rho_2)} \qquad (7)$$

where B_1 equals the length of a busy period in a queue consisting of 1-units only.

i.e. $E[e^{-sB_1}] = \bar{B}_1(s)$ is the root with smallest absolute value in z of the equation

$$z = \bar{S}_1(s+\lambda_1-\lambda_1 z) \qquad (8)$$

and hence $E[B_1] = (1/\mu_1)/(1-\rho_1)$.

Let C denote the completion time of a 2-unit (this term was introduced by Gaver [7] and denotes the time from the commencement to the termination of the service of a 2-unit including interruptions). The length of an interruption or preemption equals B_1 (the length of a busy period in a queue of 1-units only) and the times from the end of an interruption to the next 1-arrival are i.i.d. random variables, density $\lambda_1 \exp(-\lambda_1 x)$, $x \geq 0$. Thus, if N denotes the number of interruptions during the service time of a 2-unit,

$$E[e^{-tC} \mid N, S_2] = e^{-t S_2} [\tilde{B}_1(t)]^N$$

There are three cases to consider:

if $S_2 \leq z_1$ preemptions can occur throughout S_2.

if $z_1 < S_2 \leq z_2$ preemptions can only occur during the initial period of length z_1 of the 2-unit's service.

if $S_2 > z_2$ any 1-units arriving during the non-preemptive part of a 2-unit's service will queue up and preempt the 2-unit after it has spent a time z_2 at the server. The total completion time for this case is therefore just the same as if the 2-unit behaved as an ordinary preemptive resume unit.

Therefore

$$E[e^{-tC}|S_2] = \begin{cases} e^{-tS_2} \sum_{n=0}^{\infty} [\tilde{B}_1(t)]^n \frac{(\lambda_1 S_2)^n}{n!} e^{-\lambda_1 S_2} & S_2 \leq z_1 \\ e^{-tS_2} \sum_{n=0}^{\infty} [\tilde{B}_1(t)]^n \frac{(\lambda_1 z_1)^n}{n!} e^{-\lambda_1 z_1} & z_1 < S_2 \leq z_2 \\ e^{-tS_2} \sum_{n=0}^{\infty} [\tilde{B}_1(t)]^n \frac{(\lambda_1 S_2)^n}{n!} e^{-\lambda_1 S_2} & z_2 < S_2 \end{cases}$$

$$= \begin{cases} \exp\left[-t\,S_2 - \lambda_1\,S_2(1-\bar{B}_1(t))\right] & S_2 \leq z_1 \\ \exp\left[-t\,S_2 - \lambda_1\,z_1(1-\bar{B}_1(t))\right] & z_1 < S_2 \leq z_2 \\ \exp\left[-t\,S_2 - \lambda_1\,S_2(1-\bar{B}_1(t))\right] & z_2 < S_2 \end{cases}$$

Unconditionally:

$$E\left[e^{-tC}\right] = \int_0^{z_1} \exp\left[-tx - \lambda_1 x(1-\bar{B}_1(t))\right] dS_2(x)$$

$$+ \exp\left[-\lambda_1 z_1(1-\bar{B}_1(t))\right] \int_{z_1}^{z_2} \exp(-tx)\, dS_2(x) \qquad (9)$$

$$+ \int_{z_2}^{\infty} \exp\left[-tx - \lambda_1 x(1-\bar{B}_1(t))\right] dS_2(x)$$

and similarly, or by differentiating (9),

$$E[C] = \int_0^{z_1} \left[x + \frac{x\,\rho_1}{1-\rho_1}\right] dS_2(x) + \int_{z_1}^{z_2} \left[x + \frac{\rho_1 z_1}{1-\rho_1}\right] dS_2(x)$$

$$+ \int_{z_2}^{\infty} \left[x + \frac{x\,\rho_1}{1-\rho_1}\right] dS_2(x)$$

$$= \frac{1}{\mu_2} + \frac{\rho_1}{1-\rho_1}\left[\int_0^{z_1} x\, dS_2(x) + \int_{z_1}^{z_2} z_1\, dS_2(x) + \int_{z_2}^{\infty} x\, dS_2(x)\right]$$

$$= \frac{1/\mu_2}{1-\rho_1} - \frac{\rho_1}{1-\rho_1} \int_{z_1}^{z_2} (x-z_1)\, dS_2(x) \qquad (10)$$

2.2 Early Preemption: the Optimal Policy

Letting $z_2 \to \infty$ gives the particular discipline considered by Jaiswal [8]:-

if the service time already received by a 2-unit is less than or equal to z, a 1-unit arriving preempts the 2-unit from service.

if the service time received by the 2-unit is greater than z, the non-

preemptive rule is followed.

By equations (4) and (5)

$$E[e^{-sW_1}] = \frac{\{1-\rho_1-\lambda_2 \int_z^\infty (x-z)dS_2(x)\}s+\lambda_2\{P[S_2>z]-e^{sz}\int_z^\infty e^{-sx}dS_2(x)\}}{s+\lambda_1 \bar{S}_1(s)-\lambda_1} \tag{11}$$

and

$$E[W_1] = \frac{\lambda_1 E[S_1^2]+\lambda_2 \int_z^\infty (x-z)^2\, dS_2(x)}{2(1-\rho_1)} \tag{12}$$

By equations (7) and (10)

$$E[W_2] = \frac{\lambda_1 E[S_1^2]+\lambda_2 E[S_2^2]}{2(1-\rho_1)(1-\rho_1-\rho_2)} \tag{13}$$

$$E[C] = \frac{1/\mu_2}{1-\rho_1} - \frac{\rho_1}{1-\rho_1}\int_z^\infty (x-z)\, dS_2(x) \tag{14}$$

The overall mean time in the system $F_z(DE)$, (where DE refers to the discipline: discretionary with early preemption) is therefore

$$\begin{aligned} F_z(DE) &= \rho_1 + \frac{\rho_2}{1-\rho_1} + \frac{\lambda_1^2 E[S_1^2]+\lambda_1\lambda_2 \int_z^\infty (x-z)^2\, dS_2(x)}{2(1-\rho_1)} \\ &\quad + \frac{\lambda_1\lambda_2 E[S_1^2]+\lambda_2^2 E[S_2^2]}{2(1-\rho_1)(1-\rho_1-\rho_2)} - \frac{\rho_1\lambda_2}{1-\rho_1}\int_z^\infty (x-z)\, dS_2(x) \\ &= F(PR) + \frac{\lambda_1\lambda_2}{1-\rho_1}\{\tfrac{1}{2}\int_z^\infty (x-z)^2\, dS_2(x) - \frac{1}{\mu_1}\int_z^\infty (x-z)\, dS_2(x)\} \end{aligned} \tag{15}$$

where

$$F(PR) = \rho_1 + \frac{\lambda_1^2 E[S_1^2]}{2(1-\rho_1)} + \frac{\rho_2}{1-\rho_1} + \frac{\lambda_1\lambda_2 E[S_1^2]+\lambda_2^2 E[S_2^2]}{2(1-\rho_1)(1-\rho_1-\rho_2)}$$

is the overall mean time in the system in an ordinary preemptive resume (PR) queue.

Note that $F_0(DE) = F(NP)$ and $\lim_{z\to\infty} F_z(DE) = F(PR)$ where NP is an abbreviation for non-preemptive.

Differentiating (15) with respect to z gives

$$\frac{d\,F_z(DE)}{dz} = \frac{\lambda_1\lambda_2}{1-\rho_1}\left[-\int_z^\infty (x-z)\,dS_2(x) + \frac{1}{\mu_1}\int_z^\infty dS_2(x)\right]$$

$$= \frac{\lambda_1\lambda_2}{1-\rho_1}\,P[S_2 > z]\left\{\frac{1}{\mu_1} - E[S_2 - z \mid S_2 > z]\right\}$$

Therefore, at a stationary point of $F_z(DE)$, z satisfies the equation

$$1/\mu_1 = E[S_2 - z \mid S_2 > z] \qquad (16)$$

Differentiating again

$$\frac{d^2\,F_z(DE)}{dz^2} = \frac{\lambda_1\lambda_2}{1-\rho_1}\left\{\int_z^\infty dS_2(x) - \frac{1}{\mu_1}\,\delta_2(z)\right\}$$

A solution $\bar{z}$ of equation (16) is therefore a point of minimum value iff

$$\frac{1}{\phi_2(\bar{z})} = \frac{\int_{\bar{z}}^\infty dS_2(x)}{\delta_2(\bar{z})} > \frac{1}{\mu_1} = \frac{\int_{\bar{z}}^\infty (x-\bar{z})\,dS_2(x)}{P[S_2>\bar{z}]}$$

(where $\phi_2(\bar{z})$ is the age specific failure rate of the service time of a 2-unit).

$$\text{i.e. iff } \frac{d}{dz}\,E[S_2 - z \mid S_2 > z] < 0 \qquad \text{at } z = \bar{z}$$

Similarly it is a point of maximum value iff

$$\frac{d}{dz}\,E[S_2 - z \mid S_2 > z] > 0 \qquad \text{at } z = \bar{z}$$

The solution of equation (16) for the distributions E_k, D, M and rectangular is discussed in Jaiswal [8]. It is perhaps worth mentioning that for the exponential distribution $S_2(x) = 1 - \exp(-\mu_2 x)$, $x \geq 0$,

$E[S_2 - z | S_2 > z]$ equals $1/\mu_2$ for all z; the optimum discipline in this case is therefore either preemptive resume ($z=\infty$) or non-preemptive ($z=0$) depending on the relative mean service times of the two types of units.

2.3 Late Preemption: the Optimal Policy

If $E[S_2 - z | S_2 > z]$ is a monotonic increasing function of z then equation (16) gives as its only finite solution a point of maximum of $F_z(DE)$. This suggests that for such a situation the optimum discretionary rule is the converse of that discussed in Section 2.2: i.e. the 1-unit preempts the 2-unit only if the service already received by the latter is GREATER than z. The properties of this model can be obtained by setting $z_1 = 0$ in Section 2.1. From equations (4) and (5)

$$\{s + \lambda_1 \bar{S}_1(s) - \lambda_1\} E[e^{-sW_1}]$$

$$= \{1 - \rho_1 - \lambda_2 \int_0^z x \, dS_2(x) - \lambda_2 z \int_z^\infty dS_2(x)\} s$$

$$+ \lambda_2 \{1 - \int_0^z e^{-sx} \, dS_2(x) - e^{-sz} \int_z^\infty dS_2(x)\} \qquad (17)$$

and

$$E[W_1] = \frac{\lambda_1 E[S_1^2] + \lambda_2 \int_0^z x^2 \, dS_2(x) + \lambda_2 z^2 \int_z^\infty dS_2(x)}{2(1-\rho_1)} \qquad (18)$$

From equation (10)

$$E[C] = \frac{1/\mu_2}{1-\rho_1} - \frac{\rho_1}{1-\rho_1} \int_0^z x \, dS_2(x) \qquad (19)$$

combining these equations gives the overall mean residence time in the system

$$F_z(DL) = \rho_1 + \frac{\rho_2}{1-\rho_1} + \frac{\lambda_1^2 E[S_1^2] + \lambda_1\lambda_2 \int_0^z x^2\, dS_2(x) + \lambda_1\lambda_2 z^2 \int_z^\infty dS_2(x)}{2(1-\rho_1)}$$

$$+ \frac{\lambda_1\lambda_2 E[S_1^2] + \lambda_2^2 E[S_2^2]}{2(1-\rho_1)(1-\rho_1-\rho_2)} - \frac{\rho_1\lambda_2}{1-\rho_1} \int_0^z x\, dS_2(x)$$

$$= F(PR) + \frac{\lambda_1\lambda_2}{1-\rho_1} \left\{ \tfrac{1}{2} \int_0^z x^2\, dS_2(x) + \frac{z^2}{2} \int_z^\infty dS_2(x) - \frac{1}{\mu_1} \int_0^z x\, dS_2(x) \right\} \qquad (20)$$

where DL is used as an abbreviation for 'discretionary rule with late preemption'.

Note that $F_0(DL) = F(PR)$ and $\lim_{z\to\infty} F_z(DL) = F(NP)$.

Differentiating (20) gives

$$\frac{d\, F_z(DL)}{dz} = \frac{\lambda_1\lambda_2}{1-\rho_1} \left\{ z \int_z^\infty dS_2(x) - \frac{z\, s_2(z)}{\mu_1} \right\}$$

Therefore $dF_z(DL)/dz = 0$ implies

$$z = 0,\ z = \infty \quad \text{or} \quad \phi_2(z) = \frac{s_2(z)}{P[S_2 > z]} = \mu_1 \qquad (21)$$

($\phi_2(z)$ is the age specific failure rate of S_2. This compares with equation (16). Note that

$$E\left[\frac{1}{\phi_2(S_2)}\right] = \int_0^\infty \int_z^\infty s_2(x)\, dx\, dz = \frac{1}{\mu_1}\Bigg)$$

$$\frac{d^2F_z(DL)}{dz^2} = \frac{\lambda_1\lambda_2}{1-\rho_1} \left\{ \int_z^\infty dS_2(x) - z\, s_2(z) - \frac{1}{\mu_1}\, s_2(z) - \frac{z}{\mu_1}\, \frac{d}{dz}\, s_2(z) \right\}$$

At the third solution, z_0 say, $\phi_2(z_0) = \mu_1$ and

$$\frac{d^2F_z(DL)}{dz^2} = -\frac{\lambda_1\lambda_2 z_0}{1-\rho_1} \left\{ s_2(z_0) + \frac{1}{\mu_1}\, \frac{d}{dz}\, s_2(z_0) \right\}$$

$$= \frac{-\rho_1\lambda_2 z_0 \int_{z_0}^\infty dS_2(x)}{1-\rho_1}\, \frac{d\, \phi_2(z_0)}{dz}$$

because

$$\frac{d\phi_2(z)}{dz} = \frac{ds_2(z)/dz}{P[S_2>z]} + \frac{\{s_2(z)\}^2}{\{P[S_2>z]\}^2}$$

$$= \frac{ds_2(z_o)/dz + \mu_1 s_2(z_o)}{P[S_2>z_o]} \qquad \text{at } z = z_o$$

Therefore, if $\phi_2(z)$ is decreasing at the third solution it gives a point of minimum value of $F_z(DL)$.

2.4 Discretionary Rule Based on the Estimated Remaining Service Time

Consider the alternative discretionary rule: whenever a 1-unit arrives to find a 2-unit in service

if the 2-unit still requires time greater than z to complete its service the 1-unit preempts the 2-unit from service.

if the 2-unit still requires time less than or equal to z to complete service the non-preemptive rule is followed.

Suppose a 2-unit arrives at time τ and has service time S_2; then, without affecting W_1 the waiting time of a 1-unit

if $S_2 > z$, the 2-unit can be replaced by an ordinary non-preemptive 2-unit arriving at time $\tau + S_2 - z$ and having service time z.

if $S_2 \leq z$, the 2-unit behaves as an ordinary non-preemptive unit.

For this situation it is slightly more difficult to check that the modified 2-units still arrive at random: the procedure below utilizes the results of Vere-Jones [22].

Definition: For a stochastic point process π, characterised by the counting measure N(.), the PROBABILITY GENERATING FUNCTIONAL $G[h]$ is defined by the equation

$$G[h] = E\{\exp \int \log h(t)\, dN(t)\} = E\{\prod_i h(t_i)\}$$

where the $\{t_i\}$ are the epochs of events and, to ensure convergence, $1 - h(t) \in L(\pi)$, the class of functions g satisfying $0 \leq g(t) < 1$ for every t and

$$\int g(t)\, M(dt) < \infty$$

($M(\cdot)$ denotes the first moment measure of π, i.e. $M(I)$ is the expected number of events during the interval I). Theorem (Vere-Jones [22], page 327). If π_1, π_2 are the input and output streams for the queue GI/G/∞, the corresponding p.g.fls. are related by the equation

$$G_2[h] = G_1[g] \tag{22}$$

where

$$g(t) = \int_0^\infty h(t+x)\, dF(x) \tag{23}$$

and F(x) is the distribution function of the service time. Both sides of (22) are well defined provided $1 - h(t) \in L(\pi_2)$ and $\int_0^\infty M(I-x)\, dF(x)$ is convergent for all finite intervals I. (I-x denotes the interval I translated to the left by a distance x).

Proof: An arrival at τ_i generates a departure at $\tau_i' = \tau_i + S_i$ where S_i has distribution function F(x). The p.g.fl. of the departure is

$$g(\tau_i) = G_2[h|\tau_i] = \int_0^\infty h(\tau_i+x)\, dF(x) = E_S[h(\tau_i+S_i)]$$

- E_S denoting expectation taken over the variable S_i. Thus

$$\begin{aligned} G_2[h] &= E\{\pi_i\, h(\tau_i')\} \\ &= E\{\pi_i\, h(\tau_i+S_i)\} \\ &= E\{\pi_i\, g(\tau_i)\} \\ &= G_1[g] \end{aligned}$$

It remains to check that $1 - g(t) \in L(\pi_1)$. As

$$1 - g(t) = \int_0^\infty \{1 - h(t+x)\}\, dF(x)$$

this condition is satisfied provided that $1 - h(t)$ is integrable with respect to the measure which assigns to the interval I the value

$$\int_0^\infty M(I-x)\, dF(x)$$

Corollary (first proved by Mirasol [15]). The output of the M/G/∞ queue is Poisson with rate equal to the input rate λ.

Proof:

$$G_1[g] = \exp\{-\int (1-g(t))\, \lambda\, dt\}$$

Therefore

$$G_2[h] = G_1[g]$$

$$= \exp\{-\int_t [1 - \int_0^\infty h(t+x)\, dF(x)]\, \lambda\, dt\}$$

$$= \exp\{-\int_t \int_0^\infty [1 - h(t+x)]\, \lambda\, dF(x)\, dt\}$$

$$= \exp\{-\int_t \int_0^\infty [1 - h(t)]\, \lambda\, dF(x)\, dt\}$$

$$= \exp\{-\int_t [1 - h(t)]\, \lambda\, dt\}.$$

As the p.g.fl. uniquely determines the process, the output is a Poisson process, rate λ. Q.E.D.

Using this corollary it follows directly that the modified 2-units arrive in a Poisson process rate λ and have service times τ_2 which are i.i.d. random variables with distribution function $T_2(y)$ given by

$$T_2(y) = S_2(y) \qquad 0 \leq y < z$$

with the remaining probability $\int_z^\infty dS_2(y)$ in a spike at $T_2 = z$. Therefore

$$E[T_2] = \int_0^z x\, dS_2(x) + z\, P[S_2 > z]$$

From equations (8) and (13) of Chapter 1

$$E[W_1] = \frac{\lambda_1 E[S_1^2] + \lambda_2\{\int_0^z x^2\, dS_2(x) + z^2\, P[S_2 > z]\}}{2(1-\rho_1)}$$

$$E[W_2] = \frac{\lambda_1 E[S_1^2] + \lambda_2 E[S_2^2]}{2(1-\rho_1)(1-\rho_1-\rho_2)}$$

The Laplace-Stieltjes transforms could also be written down. The completion time C can be found in the usual way:-

$$E[e^{-tC}|S_2] = \begin{cases} \exp[-tS_2] & 0 \leq S_2 \leq z \\ \exp[-tS_2 - \lambda_1(S_2-z)(1-\bar{B}_1(t))] & S_2 > z \end{cases}$$

$$E[e^{-tC}] = \int_0^z \exp(-tx)\, dS_2(x) + \int_z^\infty \exp[-tx - \lambda_1(x-z)(1-\bar{B}_1(t))]\, dS_2(x)$$

Also

$$E[C] = \frac{1}{\mu_2} + \frac{\rho_1}{1-\rho_1} \int_z^\infty (x-z)\, dS_2(x)$$

The overall mean time in the system is, therefore (where DR is an abbreviation for 'discretionary rule based on remaining service time'):-

$$F_z(DR) = \rho_1 + \frac{\lambda_1^2 E[S_1^2] + \lambda_1\lambda_2 \{\int_0^z x^2\, dS_2(x) + z^2 \int_z^\infty dS_2(x)\}}{2(1-\rho_1)} + \rho_2$$

$$+ \frac{\rho_1\lambda_2}{1-\rho_1} \int_z^\infty (x-z)\, dS_2(x) + \frac{\lambda_1\lambda_2 E[S_1^2] + \lambda_2^2 E[S_2^2]}{2(1-\rho_1)(1-\rho_1-\rho_2)}$$

$$= F(PR) + \frac{\lambda_1\lambda_2}{1-\rho_1} \left\{ \tfrac{1}{2} \int_0^z x^2\, dS_2(x) + \frac{z^2}{2} \int_z^\infty dS_2(x) - \frac{z}{\mu_1} \int_z^\infty dS_2(x) \right.$$

$$\left. - \frac{1}{\mu_1} \int_0^z x\, dS_2(x) \right\} \qquad (24)$$

Note that $F_o(DR) = F(PR)$ and $\lim_{z\to\infty} F_z(DR) = F(NP)$. The optimum value of z is obtained from

$$\frac{d\,F_z(DR)}{dz} = \frac{\lambda_1\lambda_2}{1-\rho_1}\left\{z - \frac{1}{\mu_1}\right\}\int_z^\infty dS_2(x) = 0$$

i.e. $z = 1/\mu_1$ or $z = \infty$

$$\frac{d^2\,F_z(DR)}{dz^2} = \frac{\lambda_1\lambda_2}{1-\rho_1}\left\{-\left(z-\frac{1}{\mu_1}\right)s_2(x) + \int_z^\infty dS_2(x)\right\}$$

which is positive at $z = 1/\mu_1$.

$z = 1/\mu_1$ is therefore a point of minimum value of $F_z(DR)$.

Consider now the situation of Section 1.3: on arrival an estimate Y, independent of all previous estimates, is made of S_2 the service time of a 2-unit. Let $g(x,y)$ denote the joint density function of (S_2,Y). The discretionary rule discussed above now becomes:

> whenever a 1-unit arrives to find a 2-unit in service, if the time already spent in service by the 2-unit subtracted from its estimated total service time Y is greater than z, then the 1-unit preempts the 2-unit from service; otherwise the non-preemptive rule is followed.

Note: this corresponds to a discretionary rule based on the estimated remaining service time of a 2-unit. In practice it is more likely that as the service of a 2-unit proceeds, the variance of the estimate of the remaining service time will decrease and that for a given 2-unit the estimation procedure will be carried out more than once - in fact, as many times as the number of 1-units which arrive during its service. However, this problem does not appear amenable to solution.

Properties of 1-units

Assuming that after an interruption a 2-unit resumes its service at the

point it was interrupted, the distribution of W_2 is the same as in an ordinary priority queue

$$\text{i.e.} \quad E[W_2] = \frac{\lambda_1 E[S_1^2] + \lambda_2 E[S_2^2]}{2(1-\rho_1)(1-\rho_1-\rho_2)} \tag{25}$$

To evaluate the completion time C of a 2-unit with service time S_2 and estimate Y, note that

if $Y \leqslant z$ $\quad C = S_2$

if $z < Y \leqslant z + S_2$ interruptions occur during $Y - z$

if $z + S_2 < Y$ interruptions occur during the whole service time S_2

Therefore, if as before $\bar{B}_1(t) = E[e^{-tB_1}]$ where B_1 is the length of a busy period in a queue composed of 1-units only,

$$E[e^{-tC}|S_2,Y] = \begin{cases} \exp[-tS_2] & Y \leqslant z \\ \exp[-tS_2 - \lambda_1(Y-z)\,(1-\bar{B}_1(t))] & z < Y \leqslant z + S_2 \\ \exp[-tS_2 - \lambda_1 S_2(1-\bar{B}_1(t))] & z + S_2 < Y \end{cases}$$

Unconditionally:-

$$E[e^{-tC}] = \int_0^\infty \int_{-\infty}^{z} \exp(-tx)\, g(x,y)\, dy\, dx$$

$$+ \int_0^\infty \int_z^{z+x} \exp[-tx - \lambda_1(y-z)\,(1-\bar{B}_1(t))]\, g(x,y)\, dy\, dx$$

$$+ \int_0^\infty \int_{z+x}^{\infty} \exp[-tx - \lambda_1 x\,(1-\bar{B}_1(t))]\, g(x,y)\, dy\, dx$$

and

$$E[C] = \frac{1}{\mu_2} + \frac{\rho_1}{1-\rho_1}\left\{\int_0^\infty \int_z^{z+x} (y-z)\, g(x,y)\, dy\, dx + \int_0^\infty \int_{z+x}^{\infty} x\, g(x,y)\, dy\, dx\right\} \tag{26}$$

Properties of 2-units

Suppose a 2-unit arrives at time τ, then without affecting W_1,

(i) if $Y \leq z$ it can be replaced by an ordinary non-preemptive unit.

(ii) if $z < Y \leq z + S_2$ it can be replaced by a non-preemptive unit with service time $S_2 - Y + z$ and arriving at epoch $\tau + Y - z$.

(iii) if $z + S_2 < Y$ it can be ignored as it behaves like an ordinary preemptive resume unit.

Units in group (i) arrive at random with rate

$$\lambda_2 \int_0^\infty \int_{-\infty}^{z} g(x,y)\, dy\, dx$$

and have service times which are i.i.d. random variables with density

$$s(t) = \frac{\int_{-\infty}^{z} g(t,y)\, dy}{\int_0^\infty \int_{-\infty}^{z} g(x,y)\, dy\, dx} \qquad t \geq 0 \qquad (27)$$

Using the result of Mirasol, units in group (ii) arrive at random with rate

$$\lambda_2 \int_0^\infty \int_{z}^{z+x} g(x,y)\, dy\, dx$$

and have service times which are i.i.d. random variables with density

$$s(t) = \frac{\int_z^\infty g_{S_2|Y}(t+y-z;y)\, g_Y(y)\, dy}{\int_0^\infty \int_z^{z+x} g(x,y)\, dy\, dx} \qquad t \geq 0$$

$$= \frac{\int_z^\infty g(t+y-z,y)\, dy}{\int_0^\infty \int_z^{z+x} g(x,y)\, dy\, dx} \qquad t \geq 0 \qquad (28)$$

Therefore, the distribution of W_1 is the same as in a 2-class non-preemptive queue in which

1-units arrive at random with rate λ_1 and have service times S_1 which are i.i.d. random variables, d.f. $S_1(x)$

2-units arrive at random with rate

$$\lambda_2 \int_0^\infty \int_{-\infty}^{z+x} g(x,y)\, dy\, dx$$

and have service times T_2 which are i.i.d. random variables with density

$$t_2(u) = \frac{\int_{-\infty}^{z} g(u,y)\, dy + \int_z^\infty g(u+y-z,y)\, dy}{\int_0^\infty \int_{-\infty}^{z+x} g(x,y)\, dy\, dx}$$

By equation (8) of Chapter 1

$$\begin{aligned} 2(1-\rho_1)\, E[W_1] &= \lambda_1\, E[S_1^2] + \lambda_2 \{ \int_0^\infty \int_{-\infty}^{z} u^2\, g(u,y)\, dy\, du \\ &\qquad + \int_0^\infty \int_z^\infty u^2\, g(u+y-z,y)\, dy\, du \} \\ &= \lambda_1\, E[S_1^2] + \lambda_2 \int_0^\infty \int_{-\infty}^{z} u^2\, g(u,y)\, dy\, du \\ &\qquad + \lambda_2 \int_0^\infty \int_z^{z+x} (x-y+z)^2\, g(x,y)\, dy\, dx \end{aligned} \tag{29}$$

If the notation DRE is used to refer to this discretionary rule based on estimated remaining service time, then by combining equations (25), (26) and (29), the overall mean time in the system can be written as

$$\begin{aligned} F_z(DRE) &= \rho_1 + \frac{\lambda_1 E[S_1^2] + \lambda_1\lambda_2 \{ \int_0^\infty \int_{-\infty}^{z} x^2 g(x,y)dy\, dx + \int_0^\infty \int_z^{z+x} (x-y+z)^2 g(x,y)dy\, dx \}}{2(1-\rho_1)} \\ &\quad + \rho_2 + \frac{\lambda_1\lambda_2}{1-\rho_1} \left\{ \frac{1}{\mu_1} \int_0^\infty \int_z^{z+x} (y-z) g(x,y)dy\, dx + \frac{1}{\mu_1} \int_0^\infty \int_{z+x}^\infty x\, g(x,y)dy\, dx \right\} \\ &\quad + \frac{\lambda_1\lambda_2\, E[S_1^2] + \lambda_2^2\, E[S_2^2]}{2(1-\rho_1)(1-\rho_1-\rho_2)} \\ &= F(PR) + \frac{\lambda_1\lambda_2}{1-\rho_1} \left\{ -\frac{1}{\mu_1\mu_2} + \tfrac{1}{2} \int_0^\infty \int_{-\infty}^{z} x^2\, g(x,y)\, dy\, dx \right. \end{aligned}$$

$$+ \tfrac{1}{2} \int_0^\infty \int_z^{z+x} (x-y+z)^2 \, g(x,y)dy \, dx + \frac{1}{\mu_1} \int_0^\infty \int_z^{z+x} (y-z) \, g(x,y)dy \, dx$$

$$+ \frac{1}{\mu_1} \int_0^\infty \int_{z+x}^\infty x \, g(x,y) \, dy \, dx \qquad (30)$$

Note that

i) $\lim_{z \to -\infty} F_z(DRE) = F(PR)$; $\lim_{z \to +\infty} F_z(DRE) = F(NP)$

ii) In equation (30), setting $g(x,y) = \delta(x,y) \, g_x(x)$ where

$$\int_0^z \delta(x,y) \, f(y)dy = \begin{cases} f(x) & z \geq x \\ 0 & z < x \end{cases}$$

for any continuous function f, gives equation (24) the case of perfect information.

Differentiating equation (30) leads to

$$\frac{dF_z(DRE)}{dz} = \frac{\lambda_1 \lambda_2}{1-\rho_1} \left\{ \int_0^\infty \int_z^{z+x} (x-y+z) \, g(x,y) \, dy \, dx \right.$$

$$\left. - \frac{1}{\mu_1} \int_0^\infty \int_z^{z+x} g(x,y) \, dy \, dx \right\}$$

which vanishes at $z = \pm\infty$ and

$$\frac{1}{\mu_1} = \frac{\int_0^\infty \int_z^{z+x} (x-y+z) \, g(x,y)dy \, dx}{\int_0^\infty \int_z^{z+x} g(x,y)dy \, dx} \qquad (31)$$

$$= E\left[S_2 - (Y-z) \mid z < Y \leq S_2 + z \right]$$

giving the optimum values for z.

Example

Consider the special case considered in Section 1.4(a):

S_2 has density $\mu_2 \exp(-\mu_2 x)$ $\qquad x \geq 0$

$Y = S_2 + U$

where S_2 and U are independent

and U has the distribution $N(0,\sigma^2)$.

Define V(NP), V_z(DR), V_z(DRE) by the general equation

$$F(\ \cdot\ ) = F(PR) + \frac{\lambda_1\lambda_2}{1-\rho_1}\ V(\ \cdot\ )$$

Then

$$V(NP) = \frac{E[S_2^2]}{2} - \frac{1}{\mu_1\mu_2}$$

$$= \frac{1}{\mu_2}\left(\frac{1}{\mu_2} - \frac{1}{\mu_1}\right) \quad \text{for} \quad S_2(x) = 1 - \exp(-\mu_2 x) \tag{32}$$

$$V_z(DR) = \tfrac{1}{2}\int_0^z x^2\, dS_2(x) + \frac{z^2}{2}\int_z^\infty dS_2(x) - \frac{z}{\mu_1}\int_z^\infty dS_2(x)$$

$$- \frac{1}{\mu_1}\int_0^z x\, dS_2(x)$$

- the optimum z is $1/\mu_1$; denote the corresponding minimum value of V_z(DR) by $\bar{V}$(DR). It follows that

$$\bar{V}(DR) = \tfrac{1}{2}\int_0^{1/\mu_1} x^2\, dS_2(x) - \frac{1}{2\mu_1^2}\int_{1/\mu_1}^\infty dS_2(x) - \frac{1}{\mu_1}\int_0^{1/\mu_1} x\, dS_2(x)$$

$$= \frac{1}{\mu_2}\left(\frac{1}{\mu_2} - \frac{1}{\mu_2}\exp(-\mu_2/\mu_1) - \frac{1}{\mu_1}\right) \tag{33}$$

< 0 for any μ_1, μ_2

Similarly

$$\bar{V}(DRE) = -\frac{1}{\mu_1\mu_2} + \tfrac{1}{2}\int_0^\infty\int_{-\infty}^z x^2\, g(x,y)\, dy\, dx$$

$$+ \tfrac{1}{2}\int_0^\infty\int_z^{z+x} (x-y+z)^2\, g(x,y)\, dy\, dx + \frac{1}{\mu_1}\int_0^\infty\int_z^{z+x} (y-z)\, g(x,y)\, dy\, dx$$

$$+ \frac{1}{\mu_1}\int_0^\infty\int_{z+x}^\infty x\, g(x,y)\, dy\, dx \tag{34}$$

where z satisfies the equation

$$\int_0^\infty \int_z^\infty \left(t - \frac{1}{\mu_1}\right) g(t+y-z,y)\, dy\, dt = 0 \qquad (31)$$

Substituting for $g(t+y-z,y)$ in equation (31) leads to the following equation for the optimum z

$$H(z) = \frac{\sigma}{\sqrt{2\pi}} \exp(-z^2/2\sigma^2) + (z-\mu_2\sigma^2-1/\mu_1)\, B(z) = 0 \qquad (35)$$

where $B(z) = \exp(-\mu_2 z+\mu_2{}^2\sigma^2/2)\ \Phi(z/\sigma-\mu_2\sigma)$.

Also

$$V_z(DRE) = -\frac{H(z)}{\mu_2} - \frac{B(z)}{\mu_2{}^2} + \frac{1}{\mu_2}\left(\frac{1}{\mu_2} - \frac{1}{\mu_1}\right) \Phi\left(\frac{z}{\sigma}\right) \qquad (36)$$

Numerical values of V(NP), $\bar{V}$(DR) and $\bar{V}$(DRE) calculated from these equations are given in Table 3 which shows

(i) if $1/\mu_2 > 1/\mu_1$ the optimum z first increases and then decreases as σ increases
(in this case V(PR) < V(NP)).
if $1/\mu_2 \leqslant 1/\mu_1$ the optimum z increases
(in this case V(PR) $\geqslant$ V(NP)).

(ii) as σ increases, $\bar{V}$ increases.

(iii) if $1/\mu_1$ is large, the reduction obtained in the mean waiting time is less sensitive to changes in σ.

Table 3. Discretionary Rule based on the Remaining Service Time

In each classification the upper figure is the optimum z, the lower figure is the minimum V where

$$F = F(PR) + \frac{\lambda_1 \lambda_2}{1-\rho_1} V$$

(F is the overall mean time in the system).

(a) $1/\mu_1 = 0.5$

$\frac{1}{\mu_2}$	NON PREEMPTIVE	DISCRETIONARY $\sigma=0$	$\sigma=0.1$	$\sigma=0.5$	$\sigma=1.0$
0.5	0.0 -0.092	0.5 -0.092	0.520 -0.090	0.741 -0.064	0.869 -0.042
0.9	0.36 -0.105	0.5 -0.105	0.511 -0.102	0.518 -0.058	-0.020 -0.021
0.95	0.4275 -0.106	0.5 -0.106	0.511 -0.103	0.504 -0.057	-0.079 -0.020

(b) $1/\mu_1 = 0.9$

0.5	-0.2 -0.241	0.9 -0.241	0.920 -0.241	1.351 -0.226	2.260 -0.210
0.9	0.0 -0.298	0.9 -0.298	0.911 -0.296	1.129 -0.258	1.371 -0.197
0.95	0.0475 -0.302	0.9 -0.302	0.911 -0.301	1.114 -0.260	1.312 -0.195

continued ... /

Table 3 continued ...

(c) $1/\mu_1 = 0.95$

0.5	-0.225	0.95 -0.262	0.970 -0.262	1.411 -0.248	2.374 -0.233
0.9	-0.045	0.95 -0.327	0.961 -0.325	1.188 -0.288	1.485 -0.228
0.95	0.0	0.95 -0.332	0.961 -0.330	1.174 -0.291	1.427 -0.226

CHAPTER 3

An $E_b/G/1$ Priority Queue

3.1 Introduction

Apart from a paper by Jaiswal and Thiruvengadam [9], no attempts appear to have been made at solving priority queues in which arrivals do not occur at random. The assumption that non-priority and priority units arrive in two independent renewal processes seems to prohibit any simple analysis - see [9]; however, if the following simplified arrival process is considered:

> arrival epochs form an ordinary renewal process (in the terminology of Cox [4]) and at any arrival epoch, independently of what happened at all previous epochs, with probability q_1 the arrival is a 1-unit and with probability q_2 a 2-unit

then the priority analogues of the ordinary single-server queues $E_b/G/1$ (the subject of this chapter) and GI/M/1 (Chapter 4) can be solved.

3.2 The Ordinary $E_b/G/1$ Queue

Consider an ordinary single-server first-come first-served queue in which:

> interarrival times are i.i.d. random variables with density
>
> $$\frac{\lambda(\lambda x)^{b-1} e^{-\lambda x}}{(b-1)!} \qquad (x \geq 0,\ b \text{ is a positive integer})$$

and service times are i.i.d. positive random variables with distribution function $S(x)$ and independent of the interarrival times.

Erlang's phase device will be used.

Let τ_1', τ_2', ... be the epochs of successive departures.

$\tau_0' = 0$

X^n = queue size at $\tau_n' + 0$

R^n = phase of the arrival mechanism at $\tau_n' + 0$ $(1 \leq R^n \leq b)$

$\frac{1}{\mu} = \int_0^\infty x \, dS(x)$

$\rho = \lambda/b\mu$, the traffic intensity

$A_j = P[j \text{ phases arrive in } S] = \int_0^\infty \frac{e^{-\lambda x}(\lambda x)^j}{j!} \, dS(x)$

$P^n_{mr} = P[X^n = m, R^n = r \mid X^0 = 0, R^0 = 1]$

$\pi_r^n(z) = \sum_{m=0}^{\infty} P^n_{mr} z^{bm} \qquad |z| \leq 1$

(omission of the superscript n will denote the equilibrium distribution).

The basic equations are, for $1 \leq r \leq b$,

$$\pi_r^{n+1}(z) = \{\sum_{k=1}^{b} P^n_{ok}\} \sum_{i=0}^{\infty} z^{bi} A_{bi+r-1} + \sum_{m=1}^{\infty} \sum_{k=1}^{b} P^n_{mk} z^{bm} \sum_{i=0}^{\infty} A_{bi+r-k} z^{b(i-1)} \tag{1}$$

where $$A_\nu = 0 \quad \text{for } \nu < 0 \tag{2}$$

Define

$$P_\nu(z) = \sum_{i=0}^{\infty} z^{bi} A_{bi+\nu} \qquad -(b-1) \leq \nu \leq (b-1) \tag{3}$$

Substituting in (1) gives, for $1 \leq r \leq b$

$$\pi_r^{n+1}(z) = \{\sum_{k=1}^{b} P^n_{ok}\} P_{r-1}(z) + \sum_{m=1}^{\infty} \sum_{k=1}^{b} P^n_{mk} z^{bm} \frac{P_{r-k}(z)}{z^b} \tag{4}$$

Notice that for $-(b-1) \leq \nu \leq -1$

$$z^b P_{b+\nu}(z) = \int_0^\infty \sum_{i=0}^{\infty} \frac{e^{-\lambda x}(\lambda x)^{bi+b+\nu} z^{b(i+1)}}{(bi+b+\nu)!} \, dS(x)$$

$$= \int_0^\infty \sum_{j=1}^{\infty} \frac{e^{-\lambda x}(\lambda x)^{bj+\nu} z^{bj}}{(bj+\nu)!} \, dS(x)$$

$$= \sum_{j=1}^{\infty} z^{bj} A_{bj+\nu}$$

$$= \sum_{j=0}^{\infty} z^{bj} A_{bj+\nu} \quad \text{using (2)}$$

$$= P_{\nu}(z) \tag{5}$$

Also

$$\sum_{r=0}^{b-1} P_r(z)\, z^r = \int_0^{\infty} e^{-\lambda x} \sum_{r=0}^{b-1} \sum_{i=0}^{\infty} \frac{z^{r+bi}(\lambda x)^{bi+r}}{(bi+r)!}\, dS(x)$$

$$= \int_0^{\infty} e^{-\lambda x} \sum_{i=0}^{\infty} \sum_{r=0}^{b-1} \frac{(\lambda x z)^{bi+r}}{(bi+r)!}\, dS(x)$$

$$= \bar{S}(\lambda-\lambda z) \tag{6}$$

where $\bar{S}(s)$ denotes the Laplace-Stieltjes transform of $S(x)$. Finally, for any finite constants $\alpha_1, \ldots, \alpha_k$

$$\sum_{r=1}^{b} \sum_{k=1}^{b} \alpha_k z^{r-1} P_{r-k}(z) = \sum_{k=1}^{b} \sum_{j=1-k}^{b-k} \alpha_k z^{k-1+j} P_j(z)$$

$$= \sum_{k=1}^{b} \alpha_k z^{k-1} \{ \sum_{j=1-k}^{-1} z^j P_j(z) + \sum_{j=0}^{b-k} z^j P_j(z) \}$$

$$= \sum_{k=1}^{b} \alpha_k z^{k-1} \{ \sum_{j=1-k}^{-1} z^{j+b} P_{b+j}(z) + \sum_{j=0}^{b-k} z^j P_j(z) \}$$

using (5)

$$= \sum_{k=1}^{b} \alpha_k z^{k-1} \sum_{j=0}^{b-1} z^j P_j(z)$$

$$= \sum_{k=1}^{b} \alpha_k z^{k-1} \bar{S}(\lambda-\lambda z) \tag{7}$$

Define $\pi^{n+1}(z) = \sum_{r=1}^{b} \pi_r^{n+1}(z)\, z^{r-1}$

Then, from (4) and (6):-

$$\pi^{n+1}(z) = \{ \sum_{k=1}^{b} P_{ok}^n \} \bar{S}(\lambda-\lambda z) + \frac{1}{z^b} \sum_{k=1}^{b} (\pi_k^n(z) - P_{ok}^n) \sum_{r=1}^{b} z^{r-1} P_{r-k}(z)$$

Using (7)

$$\pi^{n+1}(z) = \{\sum_{k=1}^{b} P_{ok}^{n}\}\,\bar{S}(\lambda-\lambda z) + \frac{1}{z^b}\,\bar{S}(\lambda-\lambda z)\sum_{k=1}^{b}(\pi_k{}^{n}(z)-P_{ok}^{n})\,z^{k-1}$$

$$= \frac{1}{z^b}\,\pi^{n}(z)\,\bar{S}(\lambda-\lambda z) + \frac{1}{z^b}\sum_{k=1}^{b} P_{ok}^{n}(z^b-z^{b-1})\,\bar{S}(\lambda-\lambda z)$$

Define $\pi(z,w) = \sum_{n=0}^{\infty} \pi^{n}(z)\,w^{n}$, $|w| < 1$; then as $\pi^{o}(z) = 1$, it follows that

$$\pi(z,w) = \frac{1+wz^{-b}\sum_{k=1}^{b} P_{ok}(w)(z^b-z^{k-1})\,\bar{S}(\lambda-\lambda z)}{1-wz^{-b}\,\bar{S}(\lambda-\lambda z)}$$

$$= \frac{z^b+w\,\bar{S}(\lambda-\lambda z)\sum_{k=1}^{b} P_{ok}(w)(z^b-z^{k-1})}{z^b-w\,\bar{S}(\lambda-\lambda z)} \qquad (8)$$

where $P_{ok}(w) = \sum_{n=0}^{\infty} P_{ok}^{n}\,w^{n}$ $\qquad k = 1,2,\ldots,b.$

<u>LEMMA</u> (Takacs [17], pages 82-83). If $|w| < 1$ or $|w| \leq 1$ and $\lambda > b\mu$ then the equation

$$z^b = w\,\bar{S}(\lambda-\lambda z)$$

has exactly b roots (which are distinct for $w \neq 0$), $z = \gamma_r(w)$ $(r=1,2,\ldots,b)$ in the unit circle $|z| < 1$.

If $\lambda \leq b\mu$, $\gamma_r = \gamma_r(1)$ $(r=1,2,\ldots,b-1)$ are the b-1 roots of the equation

$$z^b = \bar{S}(\lambda-\lambda z)$$

within the unit circle and $\gamma_b = 1$. Also

$$\gamma_b'(1) = \begin{cases} 1/(b-\lambda/\mu) & \lambda < b\mu \\ \infty & \lambda = b\mu \end{cases}$$

If $\lambda > b\mu$ $\gamma_r = \gamma_r(1)$ $(r=1,2,\ldots,b)$ are the b roots of the equation

$$z^b = \tilde{S}(\lambda-\lambda z)$$

within the unit circle.

This lemma is proved by using Rouché's theorem - a generalisation will be given in the next section.

As $\pi(z,w)$ is a regular function of z for $|z| \leqslant 1$, $|w| < 1$ the $\gamma_r(w)$ must be roots of the numerator of (8) also. It follows that

$$\sum_{k=1}^{b} P_{ok}(w)\ (z^{k-1}-z^b)$$

a polynomial of degree b in z is completely determined as it takes the values 0 at $z = 1$ and +1 at $z = \gamma_r(w)$ $(r=1,2,\ldots,b)$. Thus

$$\sum_{k=1}^{b} P_{ok}(w)\ (z^{k-1}-z^b) = 1 - \prod_{r=1}^{b} \frac{[z-\gamma_r(w)]}{[1-\gamma_r(w)]} \qquad (9)$$

Substituting in (8):-

$$\pi(z,w) = 1 + \frac{w\,\tilde{S}(\lambda-\lambda z)\ \prod_{r=1}^{b} \frac{[z-\gamma_r(w)]}{[1-\gamma_r(w)]}}{z^b - w\,\tilde{S}[\lambda-\lambda z]} \qquad (10)$$

- determining the distribution of queue length at epoch n. The Markov chain (Xn,Rn) is irreducible and aperiodic and therefore the limits

$$\lim_{n\to\infty} P_{mr}^{n}$$

always exist.

If $\lambda > b\mu$ $\quad |\gamma_r| < 1 \quad (r=1,2,\ldots,b)$ and therefore

$$\lim_{w\to 1} (1-w)\,\pi(z,w) = 0$$

If $\lambda \leq b\mu$ $\quad |\gamma_r| < 1 \quad (r=1,2,\ldots,b-1)$ but $\gamma_b = 1$; therefore

$$\lim_{w\to 1} (1-w)\,\pi(z,w) = \begin{cases} \dfrac{b(1-\rho)(z-1)\,\bar{S}(\lambda-\lambda z)\prod_{r=1}^{b-1}\dfrac{(z-\gamma_r)}{(1-\gamma_r)}}{z^b-\bar{S}(\lambda-\lambda z)} & \lambda < b\mu \\ 0 & \lambda = b\mu \end{cases} \qquad (11)$$

determining the equilibrium distribution of queue length.

The above approach is an adaptation suitable for generalisation to the corresponding priority queue of that given by Takacs [17] for the M/G/1 queue with batch service.

3.3 An E_b/G/1 Priority Queue: the distribution of queue length

The basic model is as follows:

arrival epochs form an ordinary renewal process with density

$$\frac{\lambda e^{-\lambda x}(\lambda x)^{b-1}}{(b-1)!} \qquad (x\geq 0,\ b \text{ a positive integer});$$

at each arrival epoch the probability that the arrival is an i-unit is q_i $(i=1,2)$ and this probability is independent of events at all previous epochs; all service times are independent (and independent of the inter-arrival times), and there is one server obeying the non-preemptive discipline.

Erlang's phase device will be used.

Let $S_i(x)$ $\quad (x\geq 0,$ mean $1/\mu_i)$ denote the distribution function of the

service time S_i of an i-unit (i=1,2). The departure epochs $\{\tau_n'\}$, $n \geq 1$, form a set of regeneration points.

Let $\tau_o' = 0$, and for $n \geq 0$

X_i^n = number of i-units waiting at $\tau_n' + 0$ (i=1,2)

R^n = phase of the arrival mechanism at $\tau_n' + 0$ $(1 \leq R^n \leq b)$

$A_j = P[j \text{ phases arrive in an } S_1]$

$$= \int_0^\infty \frac{e^{-\lambda x}(\lambda x)^j}{j!}\, dS_1(x) \qquad (j \geq 0)$$

$B_j = P[j \text{ phases arrive in an } S_2]$ $(j \geq 0)$

$$ {}^{i+j}T_i = \binom{i+j}{i} q_1^{\,i}\, q_2^{\,j} \qquad (i \geq 0,\ j \geq 0)$$

$\rho = \dfrac{q_1/\mu_1 + q_2/\mu_2}{b/\lambda}$, the traffic intensity

$$P_{kmr}^n = P[X_1^n = k,\ X_2^n = m,\ R^n = r \mid X_1^o = 0,\ X_2^o = 0,\ R^o = 1]$$

$(k \geq 0,\ m \geq 0,\ 1 \leq r \leq b)$

$$\pi_r^n(s,t) = \sum_{k=0}^{\infty}\sum_{m=0}^{\infty} P_{kmr}^n\, s^k\, t^m \qquad |s| \leq 1,\ |t| \leq 1$$

Then, for $1 \leq r \leq b$,

$$\pi_r^{n+1}(s,t) = \{\sum_{k=1}^{b} P_{ook}^n\} \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} (q_1 A_{bi+bj+r-1} + q_2 B_{bi+bj+r-1})^{i+j}\, {}^{i+j}T_i\, s^i\, t^j$$

$$+ \sum_{m=1}^{\infty}\sum_{k=1}^{b} P_{omk}^n \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} B_{bi+bj+r-k}\, {}^{i+j}T_i\, s^i\, t^{m-1+j}$$

$$+ \sum_{k=1}^{b}\sum_{\ell=1}^{\infty}\sum_{m=0}^{\infty} P_{\ell mk}^n \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} A_{bi+bj+r-k}\, {}^{i+j}T_i\, s^{\ell-1+i}\, t^{m+j} \qquad (12)$$

where $A_\nu = B_\nu = 0$ for $\nu < 0$.

Define

$$P_\nu(s,t) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} s^i\, t^j\, A_{bi+bj+\nu}\, {}^{i+j}T_i \qquad -(b-1) \leq \nu \leq (b-1) \qquad (13)$$

$$Q_\nu(s,t) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} s^i t^j B_{bi+bj+\nu}\, {}^{i+j}T_i \qquad -(b-1) \leq \nu \leq (b-1) \tag{14}$$

$z^b = q_1 s + q_2 t$ i.e. z is a function of s and t.

Using equation (3), for any $\nu = -(b-1), \ldots, b-1$

$$\begin{aligned} P_\nu(z) &= \sum_{k=0}^{\infty} (q_1 s + q_2 t)^k A_{bk+\nu} \\ &= \sum_{k=0}^{\infty} \sum_{i=0}^{k} \frac{q_1{}^i s^i q_2{}^{k-i} t^{k-i} k!}{i!\,(k-i)!} A_{bk+\nu} \\ &= \sum_{i=0}^{\infty} \sum_{k=i}^{\infty} s^i t^{k-i}\, {}^kT_i\, A_{bk+\nu} \\ &= \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} s^i t^j\, {}^{i+j}T_i\, A_{bi+bj+\nu} \\ &= P_\nu(s,t) \end{aligned} \tag{15}$$

It follows from equation (5) that

$$P_{-\nu}(s,t) = z^b P_{b-\nu}(s,t) \qquad 1 \leq \nu \leq b-1 \tag{16}$$

and from (6) that

$$\sum_{r=0}^{b-1} z^r P_r(s,t) = \bar{S}_1(\lambda - \lambda z) \tag{17}$$

with similar relations for the $Q_r(s,t)$. Recall that $\bar{S}_i(s)$ denotes

$$\int_0^\infty e^{-sx}\, dS_i(x) \qquad i = 1,2;\ \mathrm{Re}(s) \geq 0$$

Substituting (13) and (14) in (12) gives

$$\begin{aligned} \pi_r^{n+1}(s,t) = \{ &\sum_{k=1}^{b} P_{ook}^{n} \} \{ q_1 P_{r-1}(z) + q_2 Q_{r-1}(z) \} \\ &+ \sum_{m=1}^{\infty} \sum_{k=1}^{b} P_{omk}^{n}\, t^{m-1} Q_{r-k}(z) \end{aligned}$$

$$+ \sum_{\ell=1}^{\infty} \sum_{m=0}^{\infty} \sum_{k=1}^{b} P_{\ell mk}^{n} \, s^{\ell-1} \, t^{m} \, P_{r-k}(z)$$

$$= \{ \sum_{k=1}^{b} P_{ook}^{n} \} \{ q_1 \, P_{r-1}(z) + q_2 \, Q_{r-1}(z) \}$$

$$+ \frac{1}{t} \sum_{k=1}^{b} \left[\pi_k^{n}(0,t) - P_{ook}^{n} \right] Q_{r-k}(z)$$

$$+ \frac{1}{s} \sum_{k=1}^{b} \left[\pi_k^{n}(s,t) - \pi_k^{n}(0,t) \right] P_{r-k}(z) \qquad (18)$$

Define $\pi^{n+1}(s,t) = \sum_{r=1}^{b} \pi_r^{n+1}(s,t) \, z^{r-1}$. Using (17), (18) and (7) gives

$$\pi^{n+1}(s,t) = \{ \sum_{k=1}^{b} P_{ook}^{n} \} \{ q_1 \, \bar{S}_1(\lambda-\lambda z) + q_2 \, \bar{S}_2(\lambda-\lambda z) \}$$

$$+ \frac{1}{t} \sum_{k=1}^{b} \left[\pi_k^{n}(0,t) - P_{ook}^{n} \right] z^{k-1} \, \bar{S}_2(\lambda-\lambda z)$$

$$+ \frac{1}{s} \sum_{k=1}^{b} \left[\pi_k^{n}(s,t) - \pi_k^{n}(0,t) \right] z^{k-1} \, \bar{S}_1(\lambda-\lambda z) \qquad (19)$$

If $\pi(s,t,w) = \sum_{n=0}^{\infty} \pi^{n}(s,t) \, w^{n}$, $|w| < 1$ then as $\pi^{o}(s,t) = 1$, we have

$$\{ s - w \, \bar{S}_1(\lambda-\lambda z) \} \, \pi(s,t,w)$$

$$= s + sw \{ q_1 \, \bar{S}_1(\lambda-\lambda z) + q_2 \, \bar{S}_2(\lambda-\lambda z) \} \sum_{k=1}^{b} P_{ook}(w)$$

$$- \frac{sw}{t} \sum_{k=1}^{b} P_{ook}(w) \, z^{k-1} \, \bar{S}_2(\lambda-\lambda z)$$

$$+ sw \sum_{k=1}^{b} \pi_k(0,t,w) \, z^{k-1} \left\{ \frac{\bar{S}_2(\lambda-\lambda z)}{t} - \frac{\bar{S}_1(\lambda-\lambda z)}{s} \right\} \qquad (20)$$

The only unknowns in this equation are the $\pi_k(0,t,w)$ $(1 \leq k \leq b)$ because $P_{ook}(w) = P_{ok}(w)$ $(1 \leq k \leq b)$ where the $P_{ok}(w)$, from the results for an ordinary $E_b/G/1$ queue are given by (9). In particular

$$\sum_{k=1}^{b} P_{ook}(w) = \prod_{r=1}^{b} \frac{1}{1-\gamma_r(w)}$$

and

$$\sum_{k=1}^{b} P_{ook}(w)\, z^{k-1} = 1 + \frac{z^b - \prod_{r=1}^{b} (z-\gamma_r(w))}{\prod_{r=1}^{b} (1-\gamma_r(w))}$$

where the $\gamma_r(w)$ are the roots in z within the unit circle of the equation

$$z^b = w\, q_1\, \bar{S}_1(\lambda-\lambda z) + w\, q_2\, \bar{S}_2(\lambda-\lambda z)$$

Consider the denominator of $\pi(s,t,w)$ in (20) - it can be written

$$z^b = q_2 t + q_1 w\, \bar{S}_1(\lambda-\lambda z)$$

LEMMA If $|w| < 1$ or $|w| \leq 1$ and $q_1\lambda > b\mu_1$ then the equation

$$z^b = q_2 t + q_1 w\, \bar{S}_1(\lambda-\lambda z) \qquad |t| \leq 1,\ q_1 + q_2 = 1 \qquad (21)$$

has exactly b roots

$$z = \delta_r(t,w) \qquad 1 \leq r \leq b$$

in the unit circle $|z| < 1$.

If $q_1\lambda \leq b\mu_1$ $\quad \delta_r = \delta_r(1,1)$ $\quad r = 1,2,\ldots,b-1$ are the b-1 roots of the equation

$$z^b = q_2 + q_1\, \bar{S}_1(\lambda-\lambda z)$$

within the unit circle and $\delta_b = \delta_b(1,1) = 1$.

Also

$$\frac{d}{dw}\, \delta_b(1,w)\Big|_{w=1} = q_1/b\left(1 - \frac{q_1\lambda}{b\mu_1}\right)$$

If $q_1\lambda > b\mu_1$ $\quad \delta_r = \delta_r(1,1)$ $\quad r = 1,2,\ldots,b$ are the b roots

of the equation

$$z^b = q_2 + q_1 \bar{S}_1(\lambda - \lambda z)$$

within the unit circle.

PROOF $$|q_2 t + q_1 w \bar{S}_1(\lambda - \lambda z)| \leqslant q_2 + q_1 |w| \, |\bar{S}_1(\lambda - \lambda z)|$$

$$\leqslant q_2 + q_1 |w| \, \bar{S}_1(\lambda - \lambda |z|)$$

$$\leqslant q_2 + q_1 |w| \, \bar{S}_1(\lambda \varepsilon)$$

for $|z| = 1 - \varepsilon$. If $|w| < 1$, then clearly $\varepsilon > 0$ can be chosen sufficiently small so that

$$q_2 + q_1 |w| \, \bar{S}_1(\lambda \varepsilon) < q_2 + q_1 |w| < (1-\varepsilon)^b$$

If $|w| \leqslant 1$ but $q_1 \lambda > b \mu_1$ then for $0 \leqslant \varepsilon \leqslant 1$, $(1-\varepsilon)^b$ and $q_2 + q_1 \bar{S}_1(\lambda \varepsilon)$ are both monotonic decreasing functions of ε which agree at $\varepsilon = 0$ and their derivatives at $\varepsilon = 0$ are $-b$ and $-q_1 \lambda / \mu_1$ respectively. As $q_1 \lambda > b \mu_1$, for ε sufficiently small

$$q_2 + q_1 |w| \, \bar{S}_1(\lambda \varepsilon) \leqslant q_2 + q_1 \bar{S}_1(\lambda \varepsilon) < (1-\varepsilon)^b$$

Thus in both cases

$$|q_2 t + q_1 w \bar{S}_1(\lambda - \lambda z)| < (1-\varepsilon)^b \quad \text{if } |z| = 1 - \varepsilon$$

and ε is sufficiently small. Hence by Rouché's theorem (21) has exactly b roots

$$z = \delta_r(t, w) \qquad 1 \leqslant r \leqslant b$$

in the unit circle $|z| < 1$, and these roots are distinct for $w \neq 0$.

If $q_1\lambda > b\mu_1$, then it follows immediately that

$$\delta_r = \delta_r(1,1) \ , \quad |\delta_r| < 1 \qquad r = 1,2,\ldots,b$$

are the b roots of the equation

$$z^b = q_2 + q_1 \bar{S}_1(\lambda-\lambda z)$$

within the unit circle.

Suppose $q_1\lambda \leqslant b\mu_1$, then the functions z^b and $q_2 + q_1 \bar{S}_1(\lambda-\lambda z)$ coincide at $z = 1$ and have derivatives b and $q_1\lambda/\mu_1$ respectively at that point. If b is even then $z^b = +1$ at $z = -1$ and hence only an odd number of roots can occur in $(-1,+1)$. Similarly if b is odd then $z^b = -1$ whilst $q_2 + q_1 \bar{S}_1(\lambda-\lambda z) > -1$ at $z = -1$ and hence only an even number of roots can occur in $(-1,+1)$. It follows that $\delta_r = \delta_r(1,1)$, $|\delta_r| < 1$ $r = 1,2,\ldots,b-1$ are the b-1 roots of the equation

$$z^b = q_2 + q_1 \bar{S}_1(\lambda-\lambda z)$$

within the unit circle and $\delta_b = 1$.

Finally

$$\frac{d}{dw} \ \delta_b(1,w)\Big|_{w=1} = q_1/b \ (1-q_1\lambda/b\mu_1)$$

As $\pi(s,t,w)$ is a regular function for $|s| \leqslant 1$, $|t| \leqslant 1$, $|w| < 1$ the $\delta_r(t,w)$ must be roots of the numerator of the expression (20) for $\pi(s,t,w)$. It follows that

$$\sum_{k=1}^{b} \pi_k(0,t,w) \ z^{k-1}$$

a polynomial of degree b-1 in z is completely determined as it takes the

values

$$\Delta_r(t,w) = \{w\,\bar{S}_2(\lambda-\lambda\delta_r(t,w)) \sum_{k=1}^{b} P_{ook}(w)\,(\delta_r(t,w))^{k-1} - t - t\lceil(\delta_r(t,w))^b - q_2 t + wq_2\,\bar{S}_2(\lambda-\lambda\delta_r(t,w))\rceil \sum_{k=1}^{b} P_{ook}(w)\}/\{w\,\bar{S}_2(\lambda-\lambda\delta_r(t,w)) - t\} \tag{22}$$

at $z = \delta_r(t,w) \qquad 1 \leqslant r \leqslant b$

Thus

$$\sum_{k=1}^{b} \pi_k(0,t,w)\,z^{k-1} = \sum_{r=1}^{b} \Delta_r(t,w) \prod_{v\neq r} \frac{z-\delta_v(t,w)}{\delta_r(t,w)-\delta_v(t,w)} \tag{23}$$

and hence $\pi(s,t,w)$ has been determined.

Special Case (A) $\bar{S}_1 \equiv \bar{S}_2 \equiv S$. Equation (20) then simplifies to

$$t\{s - w\,\bar{S}(\lambda-\lambda z)\}\,\pi(s,t,w) = ts + sw\,\bar{S}(\lambda-\lambda z) \sum_{k=1}^{b} P_{ook}(w)\,(t-z^{k-1}) + (s-t)w\,\bar{S}(\lambda-\lambda z) \sum_{k=1}^{b} \pi_k(0,t,w)\,z^{k-1}$$

and (22) becomes

$$\Delta_r(t,w) = \{w\,\bar{S}(\lambda-\lambda\,\delta_r(t,w)) \sum_{k=1}^{b} P_{ook}(w)\,\lceil(\delta_r(t,w))^{k-1} - t\rceil - t\}/\{w\,\bar{S}(\lambda-\lambda\,\delta_r(t,w)) - t\}$$

STATIONARY DISTRIBUTION

If $\rho = \frac{\lambda}{b}\left(\frac{q_1}{\mu_1} + \frac{q_2}{\mu_2}\right) < 1$ then

$$\sum_{k=1}^{b} P_{ook} = \lim_{w\to 1}(1-w) \sum_{k=1}^{b} P_{ook}(w) = \frac{1}{\gamma_b'(1)} \prod_{k=1}^{b-1} \frac{1}{(1-\gamma_k)}$$

$$= \frac{b(1-\rho)}{\prod_{k=1}^{b-1}(1-\gamma_k)} \tag{24}$$

and

$$\sum_{k=1}^{b} P_{ook}\, z^{b-1} = b(1-\rho)\ \frac{z^b - \prod_{k=1}^{b} (z-\gamma_k)}{\prod_{k=1}^{b-1} (1-\gamma_k)} \qquad (25)$$

where γ_k $(k=1,2,\ldots,b-1)$ are the b-1 roots within the unit circle of the equation

$$z^b = q_1\, \bar{S}_1(\lambda-\lambda z) + q_2\, \bar{S}_2(\lambda-\lambda z)$$

and $\gamma_b = 1$.

From equation (22)

$$\Delta_r(1,w) = \{w\, \bar{S}_2(\lambda-\lambda\, \delta_r(w)) \sum_{k=1}^{b} P_{ook}(w)\, (\delta_r(w))^{k-1} - 1 - [(\delta_r(w))^b - q_2$$

$$+ wq_2\, \bar{S}_2(\lambda-\lambda\, \delta_r(w))] \sum_{k=1}^{b} P_{ook}(w)\}/\{w\, \bar{S}_2(\lambda-\lambda\, \delta_r(w)) - 1\}$$

where $\delta_r(w) = \delta_r(1,w)$, $1 \leq r \leq b$, are the b roots within the unit circle of the equation

$$z^b = q_2 + q_1 w\, \bar{S}_1(\lambda-\lambda z)$$

As $q_1\lambda/b\mu_1 < 1$, $|\delta_r| < 1$ for $r = 1,2,\ldots,b-1$ and $\delta_b = 1$ (where $\delta_r = \delta_r(1)$, $1 \leq r \leq b$).

Also $\delta_b'(1) = q_1/b(1-q_1\lambda/b\mu_1)$.

Therefore, for $r = 1,2,\ldots,b-1$

$$\Delta_r = \lim_{w\to 1} (1-w)\, \Delta_r(1,w)$$

$$= \frac{\bar{S}_2(\lambda-\lambda\, \delta_r) \sum_{k=1}^{b} P_{ook}\, \delta_r^{k-1} - \{\delta_r^{\,b} - q_2 + q_2\, \bar{S}_2'(\lambda-\lambda\, \delta_r)\} \sum_{k=1}^{b} P_{ook}}{\bar{S}_2(\lambda-\lambda\, \delta_r) - 1} \qquad (26)$$

For the special case $r = b$, note that

$$\Delta_b(1,w) = \frac{\prod_{k=1}^{b}(1-\gamma_k(w))+(\delta_b(w))^b}{\prod_{k=1}^{b}(1-\gamma_k(w))} + \frac{q_2-w\,\bar{S}_2(\lambda-\lambda\,\delta_b(w))\{q_2+\prod_{k=1}^{b}(\delta_b(w)-\gamma_k(w))\}}{\{w\,\bar{S}_2(\lambda-\lambda\,\delta_b(w))-1\}\prod_{k=1}^{b}(1-\gamma_k(w))}$$

and hence

$$\lim_{w\to 1}\Delta_b(1,w) = \frac{1}{\gamma_b'(1)\prod_{k=1}^{b-1}(1-\gamma_k)} + \frac{-q_2-q_2\lambda\,\delta_b'(1)/\mu_2-(\delta_b'(1)-\gamma_b'(1))\prod_{k=1}^{b-1}(1-\gamma_k)}{\gamma_b'(1)\prod_{k=1}^{b-1}(1-\gamma_k)\{1+\lambda\delta_b'(1)/\mu_2\}}$$

i.e.

$$\Delta_b = \frac{q_1+q_1\lambda\,\delta_b'(1)/\mu_2-\{\delta_b'(1)-\gamma_b'(1)\}\prod_{k=1}^{b-1}(1-\gamma_k)}{\gamma_b'(1)\prod_{k=1}^{b-1}(1-\gamma_k)(1+\lambda\delta_b'(1)/\mu_2)}$$

$$= q_1\sum_{k=1}^{b}P_{ook} + q_2 \qquad (27)$$

using the relations

$$\gamma_b'(1) = 1/b(1-\rho)$$

and

$$\delta_b'(1) = q_1/b\,(1-q_1\lambda/b\mu_1)$$

From equation (23)

$$\sum_{k=1}^{b}\pi_k(0,1)\,z^{k-1} = \lim_{w\to 1}(1-w)\sum_{k=1}^{b}\pi_k(0,1,w)\,z^{k-1}$$

$$= \sum_{r=1}^{b}\Delta_r\prod_{\nu\neq r}\frac{(z-\delta_\nu)}{(\delta_r-\delta_\nu)}$$

- note that $\sum_{k=1}^{b}\pi_k(0,1) = \Delta_b$

Therefore, from equation (20),

$$\{s - \bar{S}_1(\lambda-\lambda z)\}\,\pi(s,1) = \frac{sb(1-\rho)\{q_1\,\bar{S}_1(\lambda-\lambda z)+q_2\,\bar{S}_2(\lambda-\lambda z)\}}{\prod_{r=1}^{b-1}(1-\gamma_r)}$$

$$- s\,\bar{S}_2(\lambda-\lambda z)\,b(1-\rho)\,\frac{z^b-\prod_{r=1}^{b}(z-\gamma_r)}{\prod_{r=1}^{b-1}(1-\gamma_r)}$$

$$+ \{s\,\bar{S}_2(\lambda-\lambda z) - \bar{S}_1(\lambda-\lambda z)\}\sum_{r=1}^{b}\Delta_r\prod_{\nu\neq r}\frac{(z-\delta_\nu)}{(\delta_r-\delta_\nu)} \qquad (28)$$

where z^b now equals $q_1 s + q_2$.

Special Case (A) $S_1 \equiv S_2 \equiv S$

If γ_k $(k=1,2,\ldots,b-1)$ are the b-1 roots within the unit circle of the equation $z^b = \bar{S}(\lambda-\lambda z)$ and $\gamma_b = 1$; also, if δ_k $(k=1,2,\ldots,b-1)$ are the b-1 roots within the unit circle of the equation

$$z^b = q_2 + q_1\,\bar{S}(\lambda-\lambda z)$$

and $\delta_b = 1$, then for $1 \leq r \leq b-1$

$$\Delta_r = \lim_{w\to 1}(1-w)\,\Delta_r(1,w) = \frac{\bar{S}(\lambda-\lambda\,\delta_r)\sum_{k=1}^{b}P_{ook}\,(\delta_r^{k-1}-1)}{\bar{S}(\lambda-\lambda\,\delta_r)-1}$$

$$\text{i.e.}\quad \Delta_r = \frac{(\delta_r{}^b-q_2)\sum_{k=2}^{b}P_{ook}\,(1-\delta_r^{k-1})}{1-\delta_r{}^b}$$

and

$$\Delta_b = \lim_{w\to 1}(1-w)\,\Delta_b(1,w) = q_1\sum_{k=1}^{b}P_{ook} + q_2$$

Finally, equation (28) becomes

$$\{s - \bar{S}(\lambda-\lambda z)\}\ \pi(s,1) = \frac{s\ \bar{S}(\lambda-\lambda z)b(1-\rho)\{q_1(1-s)+ \prod_{r=1}^{b}(z-\gamma_r)\}}{\prod_{r=1}^{b-1}(1-\gamma_r)}$$

$$+ (s-1)\ \bar{S}(\lambda-\lambda z) \sum_{k=1}^{b} \pi_k(0,1)\ z^{k-1}$$

where $z^b \equiv q_1 s + q_2$.

Special Case (B) $b = 1$, i.e. random arrivals

$$\prod_{r=1}^{b-1}(1-\gamma_r) = 1 \qquad \text{and } \gamma_1 = 1$$

$$\sum_{k=1}^{b} P_{ook} = P_{001} = 1 - \rho$$

$$\Delta_b = \Delta_1 = q_1 P_{001} + q_2 = 1 - q_1\rho$$

$$\sum_{k=1}^{b} \pi_k(0,1) = \pi_1(0,1) = \Delta_b = 1 - q_1\rho$$

Equation (28) becomes

$$\{s - \bar{S}_1(q_1\lambda(1-s))\}\ \pi(s,1)$$

$$= s(1-\rho)\ \{q_1\ \bar{S}_1(q_1\lambda(1-s)) + q_2\ \bar{S}_2(q_1\lambda(1-s))\}$$

$$- (1-\rho)s\ \bar{S}_2(q_1\lambda(1-s)) + \{s\ \bar{S}_2(q_1\lambda(1-s)) - \bar{S}_1(q_1\lambda(1-s))\}\ (1-q_1\rho)$$

$$= sq_1\ \bar{S}_1(q_1\lambda(1-s))\ (1-\rho) + q_2 s\ \bar{S}_2(q_1\lambda(1-s))$$

$$- (1-q_1\rho)\ \bar{S}_1(q_1\lambda(1-s))$$

and $\pi(s,1)$ then agrees with the expression for $\alpha(s) + \beta_o$ found in Chapter 1.

3.4 The $E_b/G/1$ Priority Queue: the Waiting Time Distributions

Stationary Waiting Time Distribution of a 1-unit

Let W_1 denote the waiting time of a 1-unit in the stationary state. Then for $1 \leq r \leq b$ (recall R denotes the phase of the arrival mechanism).

$$P[n \text{ 1-units waiting and } R = r \text{ at a departure epoch}]$$

$$= P[\text{departure is a 1-unit}] \sum_{j=0}^{\infty} P[nb + jb + r - 1 \text{ phases and } n \text{ 1-units arrive in } W_1 + S_1]$$

$$+ \sum_{k=1}^{b} P[\text{departure is a 2-unit and at last epoch } R = k] \sum_{j=0}^{\infty} B_{nb+jb+r-k}\, {}^{n+j}T_n$$

Therefore, if the symbol * denotes convolution,

$$\pi_r(s,1) = q_1 \sum_{n=0}^{\infty}\sum_{j=0}^{\infty} {}^{n+j}T_n\, s^n \int_0^{\infty} \frac{e^{-\lambda x}(\lambda x)^{nb+jb+r-1}}{(nb+jb+r-1)!}\, d(W_1(x)*S_1(x))$$

$$+ \sum_{k=1}^{b} [\pi_k(0,1) - q_1 P_{ook}] \sum_{n=0}^{\infty}\sum_{j=0}^{\infty} B_{nb+jb+r-k}\, {}^{n+j}T_n\, s^n$$

and therefore,

$$\pi(s,1) = q_1 \sum_{r=1}^{b} (q_1 s+q_2)^{\frac{r-1}{b}} \sum_{n=0}^{\infty}\sum_{j=0}^{\infty} {}^{n+j}T_n\, s^n \int_0^{\infty} \frac{e^{-\lambda x}(\lambda x)^{nb+jb+r-1}}{(nb+jb+r-1)!}\, d(W_1(x)*S_1(x))$$

$$+ \sum_{k=1}^{b} [\pi_k(0,1) - q_1 P_{ook}] \sum_{r=1}^{b} Q_{r-k}(s,1)\, (q_1 s+q_2)^{\frac{r-1}{b}}$$

Then using equation (7) and an analogue of (6) for $W_1 * S_1$ gives

$$\pi(s,1) = q_1 \bar{S}_1[\lambda - \lambda(q_1 s+q_2)^{1/b}]\, \bar{W}_1[\lambda - \lambda(q_1 s+q_2)^{1/b}]$$

$$+ \sum_{k=1}^{b} [\pi_k(0,1) - q_1 P_{ook}]\, (q_1 s+q_2)^{(k-1)/b}\, \bar{S}_2[\lambda - \lambda(q_1 s+q_2)^{1/b}]$$

If $z = (q_1 s+q_2)^{1/b}$

$$\bar{W}_1(\lambda-\lambda z) = \frac{\pi((z^b-q_2)/q_1,1) - \sum_{k=1}^{b} [\pi_k(0,1) - q_1 P_{ook}] z^{k-1} \bar{S}_2(\lambda-\lambda z)}{q_1 \bar{S}_1(\lambda-\lambda z)}$$

In particular, by differentiating

$$\frac{q_1\lambda}{\mu_1} + q_1\lambda E[W_1] = \frac{b}{q_1} \frac{d}{ds} \pi(s,1)\Big|_{s=1} - \sum_{k=1}^{b} [\pi_k(0,1) - q_1 P_{ook}] \left[k - 1 + \frac{\lambda}{\mu_2}\right]$$

i.e.

$$q_1\lambda E[W_1] = \frac{b}{q_1} \frac{d}{ds} \pi(s,1)\Big|_{s=1} - \sum_{k=1}^{b} [\pi_k(0,1) - q_1 P_{ook}] \left[k - 1 + \frac{\lambda}{\mu_2}\right] - \frac{q_1\lambda}{\mu_1}$$

(29)

where $\pi(s,1)$ given by (28) is probably best differentiated numerically, and

$$\sum_{k=1}^{b} \pi_k(0,1)\, z^{k-1} = \sum_{r=1}^{b} \Delta_r \prod_{\nu \neq r} \frac{(z-\delta_\nu)}{(\delta_r-\delta_\nu)}$$

$$\sum_{k=1}^{b} P_{ook}\, s^{(k-1)/b} = b(1-\rho) \frac{s - \prod_{r=1}^{b} (s^{1/b}-\gamma_r)}{\prod_{r=1}^{b-1} (1-\gamma_r)}$$

Although the results are complicated, numerical values for $E[W_1]$ could easily be obtained with the aid of a computer.

Mean Waiting Time of a 2-unit

For a GI/G/1 priority queue of the type considered in the last section consider the corresponding pooled queue: i.e. an ordinary GI/G/1 queue in which service times are i.i.d. with distribution function $S(x) = q_1 S_1(x) + q_2 S_2(x)$ and independent of the interarrival times.

Let V(t) denote the virtual waiting time at time t in this pooled queue, and

$$E[V] = \lim_{T\to\infty} \frac{1}{T} \int_0^T E[V(t)]dt = \text{time average of } E[V(t)]$$

Then

$$E[V] = E[V_q] + E[V_s]$$

where the suffix s denotes the contribution due to the unit already in service and q denotes the contribution due to remaining units in the queue.

Now $E[V_q] = (1/\mu)\, E[N_q]$

where $E[N_q]$ is the time average of the number of units in the queue.

Using Little's "L = λW" result

$$E[V_q] = \rho\, E[W]$$

where $\rho = (q_1\lambda)/\mu_1 + (q_2\lambda)/\mu_2$ and $E[W]$ is the mean waiting time in this pooled GI/G/1 queue.

The GI/G/1 priority queue clearly has exactly the same value for $E[V_q]$, and for this queue

$$E[V_q] = (1/\mu_1)\, E[N_q^{1}] + (1/\mu_2)\, E[N_q^{2}]$$

$$= \rho_1\, E[W_1] + \rho_2\, E[W_2]$$

where $\rho_i = (q_i\lambda)/\mu_i \qquad i = 1,2$

and $E[N_q^{i}]$ is the time average of the number of i-units in the queue excluding the one being serviced. Thus

$$\rho\, E[W] = \rho_1\, E[W_1] + \rho_2\, E[W_2] \qquad (30)$$

and $E[W_2]$ can be found from a knowledge of $E[W]$ and $E[W_1]$.

Distribution of the Stationary Waiting Time of a 2-unit

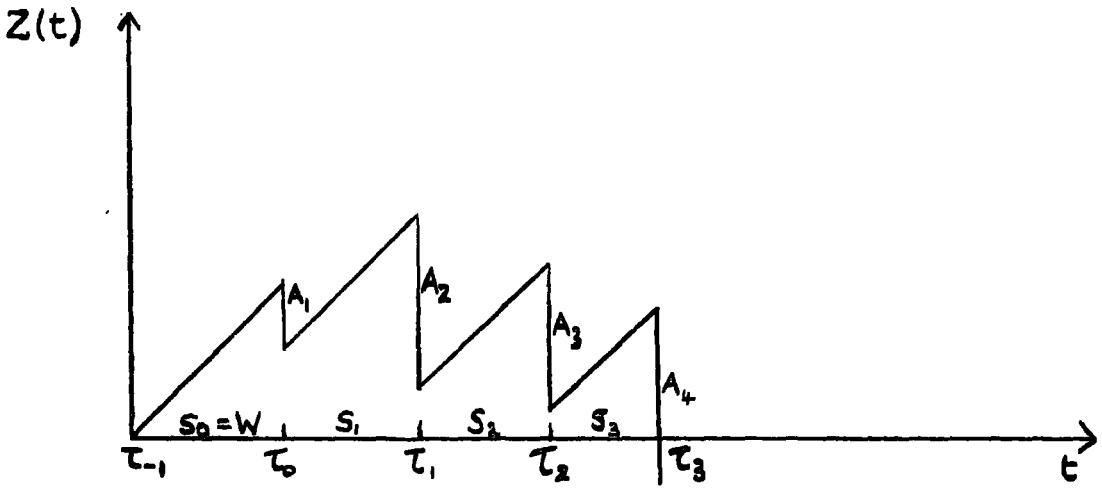
W_2, the waiting time of a 2-unit in the stationary state can be expressed as the sum of two random variables:-

W, the waiting time of a unit in the pooled queue in the stationary state, and

T(W), the time to serve all 1-units arriving in the queue after the arrival of the 2-unit in question but before its entry into service.

The problem is therefore to find T(w), which corresponds to the length of the initial busy period initiated by a waiting time w in a queue composed of 1-units only. The determination of the distribution of the supremum of a compound recurrent process has been considered by Takacs [20] and his method can be readily adapted to give a solution for the particular problem treated here.

Define Z(t), $t \geq 0$, to be the following stochastic process:-



$$Z(t) = t - \sum_{0<\tau_i \leq t} A_i$$

where A_i $(i \geq 1)$ has distribution function $A_1(x)$ with Laplace-Stieltjes transform

$$\bar{A}_1(s) = \frac{q_1 \bar{A}(s)}{1-q_2 \bar{A}(s)} \qquad \text{and} \qquad \bar{A}(s) = \left(\frac{\lambda}{\lambda+s}\right)^b$$

(i.e. A_i is an interarrival time between two 1-arrivals).

τ_0 has d.f. W(x)

and $\tau_{n+1} - \tau_n = S_{n+1}$ $(n \geq 0)$ has d.f. $S_1(x)$.

Moreover all these random variables are independent.

If $I(t) = \inf_{0<u\leq t} Z(u)$, then as the time at which $Z(t)$ first becomes negative corresponds with the length of W_2, it follows that

$$\begin{cases} P[W_2 \leq t] = P[I(t) < 0], \ t > 0 \\ P[W_2 = 0] = P[W = 0] \end{cases} \tag{31}$$

Define

$$\tau_{-1} = 0, \ S_0 = W, \ I_{-1} = 0, \ Z_{-1} = 0$$

$$C_n = S_n - A_{n+1} \qquad n = 0,1,2,\ldots$$

$$Z_n = C_0 + C_1 + \ldots + C_n \qquad n = 0,1,2,\ldots$$

$$I_n = \inf \{Z_0, Z_1, \ldots, Z_n\} \qquad n = 0,1,2,\ldots$$

Then for $t \in [\tau_n, \tau_{n+1})$ $(n \geq -1)$

$$I(t) = I_n$$

$$Z(t) = Z_n + (t-\tau_n)$$

It follows that for $n \geq -1$, $\mathrm{Re}(q) > 0$, $\mathrm{Re}(s) \leq 0$

$$q \int_{\tau_n}^{\tau_{n+1}} e^{-qt-s\,I(t)}\, dt = q \int_{\tau_n}^{\tau_{n+1}} e^{-qt}\, e^{-s\,I_n}\, dt$$

$$= q\, e^{-s\,I_n} \int_{\tau_n}^{\tau_{n+1}} e^{-qt}\, dt$$

$$= e^{-q\,\tau_n - s\,I_n} \left[1 - e^{-q(\tau_{n+1}-\tau_n)}\right]$$

Therefore

$$q \sum_{n=-1}^{\infty} \int_{\tau_n}^{\tau_{n+1}} e^{-qt-s\,I(t)}\, dt = \sum_{n=-1}^{\infty} e^{-q\,\tau_n - s\,I_n} \left[1 - e^{-q(\tau_{n+1}-\tau_n)}\right]$$

i.e.

$$q \int_0^\infty e^{-qt-s\,I(t)}\,dt = \sum_{n=-1}^{\infty} e^{-q\,\tau_n - s\,I_n}\left[1 - e^{-q(\tau_{n+1}-\tau_n)}\right]$$

Taking expectations,

$$q \int_0^\infty e^{-qt}\,E\left[e^{-s\,I(t)}\right]dt = \{1 - E\left[e^{-qW}\right]\} + \{1 - \tilde{S}_1(q)\} \sum_{n=0}^{\infty} E\left[e^{-q\,\tau_n - s\,I_n}\right]$$

$$= \{1 - E\left[e^{-qW}\right]\} + \{1 - \tilde{S}_1(q)\}\,E\left[e^{-q\,\tau_0 - s\,I_0}\right] \sum_{n=0}^{\infty} U_n(s,q) \qquad (32)$$

where $U_n(s,q) = E\left[\exp(-q(\tau_n-\tau_0) - s(I_n-I_0))\right]$.

Suppose the sequence of random variables $\{C_1,C_2,\ldots,C_n\}$ is replaced by the sequence $\{C_n,C_{n-1},\ldots,C_1\}$ then $U_n(s,q)$ is unchanged. For this new sequence define

$$I_0^* = 0$$

$$I_{k+1}^* = \text{Min}\,\{0,\ I_k^* + C_{k+1}\} \qquad 0 \leqslant k \leqslant n-1$$

Then $$I_n^* = \text{Min}\,\{0,\ C_n,\ \ldots,\ C_n + \ldots + C_1\}$$

$$= I_n - I_0$$

Therefore

$$\sum_{n=0}^{\infty} U_n(s,q) = \sum_{n=0}^{\infty} E\left[\exp(-q(\tau_n-\tau_0) - s\,I_n^*)\right]$$

where $U_0(s,q) = 1$.

The U_n can, in principle, be found recursively. The procedure below however utilizes a technique developed by Takacs [19,21]:-

Definition. If L_ε is the path of integration from $-i\infty$ to $-i\varepsilon$ and again from $+i\varepsilon$ to $+i\infty$ ($\varepsilon>0$) then the operator $\underset{\sim}{B}$ is defined by

$$\underset{\sim}{B}\,\phi(s) = \tfrac{1}{2}\,\phi(0) + \lim_{\varepsilon\to 0} \frac{s}{2\pi i} \int_{L_\varepsilon} \frac{\phi(z)}{z(z-s)}\,dz \qquad (33)$$

where $\mathrm{Re}(s) < 0$.

Suppose X,Y are two random variables such that X is real, $E\{|Y|\} < \infty$ and $E[Y\,e^{-sX}]$ exists for $\mathrm{Re}(s) = 0$, then if $\mathrm{Re}(s) < -\varepsilon < 0$

$$\frac{s}{2\pi i} \int_{C_\varepsilon^-} \frac{E[Y \exp(-z\,\mathrm{Min}(X,0))]}{z(z-s)}\,dz = E[Y\,e^{sX^-}]$$

where C_ε^- is the path consisting of L_ε and the semicircle $-\varepsilon\,e^{i\alpha}$ $(-\pi/2 \leq \alpha \leq \pi/2)$, and

$$X^- = -\mathrm{Min}\,(X,0)$$

Also, if $\mathrm{Re}(s) < 0$

$$\frac{s}{2\pi i} \int_{C_\varepsilon^+} \frac{E[Y\,e^{-zX}] - E[Y\,e^{zX^-}]}{z(z-s)}\,dz = 0$$

where C_ε^+ is the same path as C_ε^- but with the semicircle $\varepsilon\,e^{i\alpha}$
Taking the limit in these two equations and adding gives

$$E[Y\,e^{sX^-}] = \lim_{\varepsilon\to 0} \frac{s}{2\pi i} \int_{L_\varepsilon} \frac{E[Y\,e^{-zX}]}{z(z-s)}\,dz + \frac{s\pi i\,E[Y]}{2\pi i s}$$

$$= \tfrac{1}{2}\,E[Y] + \lim_{\varepsilon\to 0} \frac{s}{2\pi i} \int_{L_\varepsilon} \frac{E[Y\,e^{-zX}]\,dz}{z(z-s)}$$

i.e.

$$E[Y\,e^{-s\,\mathrm{Min}(X,0)}] = E[Y\,e^{sX^-}]$$

$$= \underset{\sim}{B}\,E[Y\,e^{-sX}]$$

Clearly the operator $\underset{\sim}{B}$ is linear and $\underset{\sim}{B}\,\underset{\sim}{B} = \underset{\sim}{B}$. For $n \geq 0$

$$\underset{\sim}{B}\{\bar{A}_1(-s)\ \bar{S}_1(q+s)\ U_n(q,s)\}$$

$$= \underset{\sim}{B}\ E\left[\exp(s\ A_{n+2} - q\ S_{n+1} - s\ S_{n+1} - q\ \tau_n + q\ \tau_o - s\ I_n^*)\right]$$

$$= \underset{\sim}{B}\ E\left[\exp(-s\ C_{n+1} - q\ \tau_n + q\ \tau_o - s\ I_n^* - q\ S_{n+1})\right]$$

$$= \underset{\sim}{B}\ E\left[\exp(-s\ C_{n+1} - s\ I_n^* - q\ \tau_{n+1} + q\ \tau_o)\right]$$

$$= \underset{\sim}{B}\ E\left[\exp(-s(C_{n+1} + I_n^*) - q(\tau_{n+1} - \tau_o))\right]$$

$$= U_{n+1}(q,s)$$

i.e. $U_{n+1}(q,s) = \underset{\sim}{B}\{\phi(s)\ U_n(q,s)\}$

where $\phi(s) = \bar{A}_1(-s)\ \bar{S}_1(q+s)$

If $0 \leq \rho \leq 1$, $\mathrm{Re}(s) < 0$, define $h(s)$ by the equation

$$\frac{1}{h(s)} = \exp\{-\log(1-\rho\phi(s)) + \underset{\sim}{B}\log(1-\rho\phi(s))\}$$

$$= \exp\{\sum_{n=1}^{\infty} \frac{\rho^n}{n}(\phi^n(s) - \underset{\sim}{B}\,\phi^n(s))\}$$

Therefore $\underset{\sim}{B}\,h(s) = 1$

Define $T(s,\rho)$ to be $\exp\{-\underset{\sim}{B}\log(1-\rho\phi(s))\}$

and then $\underset{\sim}{B}\,(1-\rho\phi(s))\ T(s,\rho) = 1$

Expanding $T(s,\rho)$ in a power series

$$\sum_{n=0}^{\infty} T_n(s)\ \rho^n$$

gives $T_o(s) = 1$

and $T_{n+1}(s) = \underset{\sim}{B}\,T_n(s)\ \phi(s)$

i.e. the T_n satisfy the same defining relations as the U_n. Therefore,

$$U(s,\rho) = \sum_{n=0}^{\infty} U_n(s,q)\ \rho^n = \exp\{-\underset{\sim}{B}\log(1-\rho\phi(s))\}$$

and

$$\sum_{n=0}^{\infty} U_n(s,q) = \exp\{-\underset{\sim}{B} \log(1-\phi(s))\}$$

$$= \exp\{-\underset{\sim}{B} \log(1-\bar{A}_1(-s)\ \bar{S}_1(q+s))\} \qquad (34)$$

This equation, together with (32) gives a complete formal solution i.e.

$$q \int_0^{\infty} e^{-qt}\ E\left[e^{-s\ I(t)}\right]dt = \{1 - E\left[e^{-qW}\right]\}$$

$$+ \{1 - \bar{S}_1(q)\}\ E\left[e^{-(q+s)W}\right]\ \bar{A}_1(-s)\ \exp\{\sum_{n=1}^{\infty} \frac{1}{n} \underset{\sim}{B}\ \phi^n(s)\}$$

where

$$\underset{\sim}{B}\ \phi^n(s) = \underset{\sim}{B}\ \{\int_0^{\infty} e^{sx}\ d\ H_n(x)\ \int_0^{\infty} e^{-(q+s)u}\ d\ F_n(u)\}$$

$$= \int_0^{\infty} e^{-qu}\ \{\int_0^{u} e^{sx-su}\ d\ H_n(x) + \int_u^{\infty} d\ H_n(x)\}\ d\ F_n(u)$$

and F_n, H_n are the distribution functions of the n-fold convolutions of S_1, A_1 respectively.

These formulae are valid for the GI/G/1 priority queue, but they are too complicated to be of much use. In certain special cases however, it is possible to obtain some simplifications.

<u>Theorem</u>. Let $|\rho| < 1$ and suppose that for $\mathrm{Re}(s) = 0$

$$1 - \rho\phi(s) = \Phi^{+}(s,\rho)\ \Phi^{-}(s,\rho) \qquad (35)$$

where $\Phi^{+}(s,\rho)$ is a regular function of s in the domain $\mathrm{Re}(s) > 0$, continuous and free from zeros in $\mathrm{Re}(s) \geq 0$ and

$$\lim_{|s|\to\infty} \frac{\Phi^{+}(s,\rho)}{s} = 0 \qquad \mathrm{Re}(s) > 0$$

Similarly for $\Phi^-(s,\rho)$ but with $\mathrm{Re}(s) > 0$ changed to $\mathrm{Re}(s) < 0$ and $\mathrm{Re}(s) \geq 0$ changed to $\mathrm{Re}(s) \leq 0$. Then

$$U(s,\rho) = \left[\Phi^+(0,\rho)\ \Phi^-(s,\rho)\right]^{-1} \qquad \mathrm{Re}(s) \leq 0$$

or

$$\left[1 - \rho\ \phi(s)\right]\ U(s,\rho) = \frac{\Phi^+(s,\rho)}{\Phi^+(0,\rho)} \qquad \mathrm{Re}(s) \leq 0$$

Proof: If $\mathrm{Re}(s) < -\varepsilon < 0 \qquad (\varepsilon>0)$

$$\frac{s}{2\pi i} \int_{C_\varepsilon^-} \frac{\log \Phi^-(z,\rho)}{z(z-s)}\, dz = \log \Phi^-(s,\rho) \tag{36}$$

If $\mathrm{Re}(s) < 0$

$$\frac{s}{2\pi i} \int_{C_\varepsilon^+} \frac{\log \Phi^+(z,\rho)}{z(z-s)}\, dz = 0 \tag{37}$$

Let $\varepsilon \to 0$ in (36) and (37), giving

$$\lim_{\varepsilon\to 0} \frac{s}{2\pi i} \int_{L_\varepsilon} \frac{\log \Phi^-(z,\rho)}{z(z-s)}\, dz + \tfrac{1}{2} \log \Phi^-(0,\rho) = \log \Phi^-(s,\rho)$$

and

$$\lim_{\varepsilon\to 0} \frac{s}{2\pi i} \int_{L_\varepsilon} \frac{\log \Phi^+(z,\rho)}{z(z-s)}\, dz - \tfrac{1}{2} \log \Phi^+(0,\rho) = 0$$

Hence, for $\mathrm{Re}(s) < 0$,

$$\log \Phi^-(s,\rho) + \log \Phi^+(0,\rho) = \tfrac{1}{2} \{\log \Phi^-(0,\rho) + \log \Phi^+(0,\rho)\}$$

$$+ \lim_{\varepsilon\to 0} \frac{s}{2\pi i} \int_{L_\varepsilon} \frac{\log\{1-\rho\phi(s)\}}{z(z-s)}\, dz$$

$$= \underset{\sim}{B} \log \{1 - \rho\ \phi(s)\}$$

and for $\mathrm{Re}(s) = 0$ by continuity.

Therefore

$$U(s,\rho) = \exp\left[-\underset{\sim}{B}\log\{1-\rho\,\phi(s)\}\right]$$

$$= \left[\Phi^{-}(s,\rho)\,\Phi^{+}(0,\rho)\right]^{-1}$$

proving the theorem.

For the special case of this chapter

$$\bar{A}_1(s) = \frac{q_1}{(1+s/\lambda)^b - q_2}$$

and

$$1-\rho\,\phi(s) = \frac{(1-s/\lambda)^b - q_2 - \rho q_1\,\bar{S}_1(q+s)}{(1-s/\lambda)^b - q_2}$$

By Rouché's Theorem,

$$(1-s/\lambda)^b - q_2 - \rho q_1\,\bar{S}_1(q+s)$$

has b roots $s = \gamma_r(q,\rho)$ if $\mathrm{Re}(s) \geq 0$, $|\rho| \leq 1$ (which all lie within the circle $|s/\lambda - 1| = 1$).

Therefore, let

$$\Phi^{+}(s,\rho) = \frac{(1-s/\lambda)^b - q_2 - \rho q_1\,\bar{S}_1(q+s)}{\prod_{r=1}^{b}\left[s-\gamma_r(q,\rho)\right]}$$

and

$$\Phi^{-}(s,\rho) = \frac{\prod_{r=1}^{b}\left[s-\gamma_r(q,\rho)\right]}{(1-s/\lambda)^b - q_2}$$

then these functions satisfy the conditions of the theorem and therefore, for $|\rho| < 1$, $\mathrm{Re}(s) \leq 0$

$$U(s,\rho) = \frac{\prod_{r=1}^{b}\gamma_r(q,\rho)\left[(1-s/\lambda)^b - q_2\right]}{\left[q_1 - \rho q_1\,\bar{S}_1(q)\right]\prod_{r=1}^{b}\left[\gamma_r(q,\rho)-s\right]}$$

Substituting in equation (32) gives for $\mathrm{Re}(s) \leq 0$

$$q \int_0^\infty e^{-qt} E\left[e^{-sI(t)}\right] dt = \{1 - E\left[e^{-qW}\right]\} + \frac{\prod_{r=1}^{b} \gamma_r(q,1) E\left[e^{-(q+s)W}\right]}{\prod_{r=1}^{b} \left[\gamma_r(q,1)-s\right]}$$

CHAPTER 4

A GI/M/1 Priority Queue

4.1 The Ordinary GI/M/1 Queue

Suppose arrivals occur at epochs τ_i, $i \geq 1$ where the interarrival times $\tau_{n+1} - \tau_n$ $(n \geq 0,\ \tau_0 = 0)$ are i.i.d. positive random variables with distribution function $A(x)$, mean $1/\lambda$. Service times of customers are assumed to be i.i.d. random variables which are independent of the τ_n and have distribution function

$$S(x) = \begin{cases} 1 - e^{-\mu x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Let $\rho = \lambda/\mu$, the traffic intensity

$$\text{and } a_k = \int_0^\infty \frac{e^{-\mu x}(\mu x)^k}{k!}\, dA(x) \qquad (k \geq 0)$$

$$= P[\text{k services completed during an interarrival time}].$$

Then

$$\sum_{k=0}^{\infty} a_k z^k = \int_0^\infty \exp[-\mu x\,(1-z)]\, dA(x)$$

$$= \bar{A}(\mu - \mu z) \qquad (1)$$

where $\bar{A}(s)$ is the Laplace-Stieltjes transform of $A(x)$. The arrival epochs form a set of regeneration points and therefore an imbedded Markov chain can be defined. Let X_n = number of customers in the system at epoch $\tau_n - 0$ $(n \geq 0)$

$$\pi_k^{\,n} = P[X_n = k \mid X_0 = 0]$$

$$\pi^n(z) = \sum_{k=0}^{\infty} \pi_k^{\,n} z^n, \text{ convergent for } |z| \leq 1$$

The basic equations

$$\pi_o^{n+1} = \sum_{i=0}^{\infty} \sum_{j=1}^{\infty} \pi_i^n \, a_{i+j} \tag{2}$$

and $$\pi_k^{n+1} = \sum_{i=0}^{\infty} a_i \, \pi_{i+k-1} \qquad k \geq 1 \tag{3}$$

lead, for $|z| = 1$, to the recurrence relation

$$\pi^{n+1}(z) = z \, \pi^n(z) \, \bar{A}(\mu-\mu/z) + \sum_{j=0}^{\infty} (1-1/z^j) \, C_j^n \tag{4}$$

where

$$C_j^n = \sum_{i=0}^{\infty} \pi_i^n \, a_{i+j+1} \quad \text{and} \quad \sum_{j=0}^{\infty} C_j^n < 1$$

Define $\pi(z,w) = \sum_{n=0}^{\infty} \pi^n(z) \, w^n$: a regular function of z for $|z| \leq 1$, $|w| < 1$.
Equation (4) gives, for $|z| = 1$

$$\pi(z,w) - \pi^o(z) = zw \, \bar{A}(\mu-\mu/z) \, \pi(z,w) + \sum_{n=0}^{\infty} \sum_{j=0}^{\infty} (1-1/z^j) \, w^{n+1} \, C_j^n$$

As $X_o = 0$, $\pi^o(z) = 1$ and therefore, for $|z| = 1$, $|w| < 1$

$$\pi(z,w) = \frac{1+w \sum_{n=0}^{\infty} \sum_{j=0}^{\infty} (1-1/z^j) \, w^n \, C_j^n}{1-wz \, \bar{A}(\mu-\mu/z)} \tag{5}$$

Define $\pi(z,w)$ for $|z| > 1$ by this expression also. As it is known that $\pi(z,w)$ is a regular function of z for $|z| \leq 1$, its only singularities in the whole complex plane are the zeros of the denominator outside the unit circle $|z| = 1$.

Lemma (Takacs [17], Page 47) If $|w| < 1$, or $|w| \leq 1$ and $\rho > 1$, then the equation

$$z = w\,\bar{A}\,(\mu - \mu z)$$

has a unique root $z = \delta(w)$ in the unit circle $|z| < 1$. In particular $\delta = \delta(1)$ is the smallest positive real root of the equation

$$z = \bar{A}\,(\mu - \mu z)$$

If $\rho = \lambda/\mu < 1$ then $\delta < 1$; if $\rho \geqslant 1$ then $\delta = 1$.

Using this lemma shows that the only root of the denominator of (5) outside $|z| = 1$ is

$$z = 1/\delta(w)$$

Define $\Psi(z,w) = \{z - 1/\delta(w)\}\,\pi(z,w)$, a regular function of z in the whole complex plane. As

$$\lim_{|z|\to\infty} \frac{\Psi(z,w)}{|z|} = 0$$

it follows that $\Psi(z,w)$ is independent of z and

$$\Psi(z,w) = \Psi(1,w) = \{1 - 1/\delta(w)\}\,\{1 - w\}^{-1}$$

i.e.

$$\pi(z,w) = \frac{\delta(w)\,\Psi(z,w)}{z\,\delta(w) - 1}$$

$$= \frac{1-\delta(w)}{\{1-z\,\delta(w)\}\{1-w\}} \qquad (6)$$

The Markov chain is irreducible and aperiodic, and therefore the limiting probabilities

$$\pi_j = \lim_{n\to\infty} \pi_j^{\,n}$$

always exist; either every $\pi_j = 0$, or every $\pi_j > 0$ and $\{\pi_j\}$ is a probability distribution. Using Abel's Theorem,

$$\pi(z) = \sum_{j=0}^{\infty} \pi_j z^j = \lim_{w\to 1} (1-w)\, \pi(z,w)$$

$$= \begin{cases} 0 & \rho \geq 1 \\ (1-\delta)/(1-z\delta) & \rho < 1 \end{cases} \qquad (7)$$

Therefore, if $\rho < 1$ a stationary solution exists and is given by $\pi_j = (1-\delta)\,\delta^j$ where δ is the unique root in z within the unit circle of the equation $z = \tilde{A}(\mu-\mu z)$.

The above method is an adaptation of the method used by Takacs to derive the queue length probabilities in a $GI/E_k/1$ queue. {see [17], pages 127-133}.

4.2 The GI/M/1 Priority Queue: the Distribution of Queue Length

As before, the arrival epochs $\{\tau_n\}_{n\geq 1}$ form an ordinary renewal process, distribution function $A(x)$, $x \geq 0$, mean $1/\lambda$. Let $\tau_0 = 0$. At any arrival epoch, independently of events at all previous epochs, the arrival is with probability q_1 a 1-unit, and with probability q_2 a 2-unit, where 1-units have non-preemptive priority over 2-units. Service times of all units are assumed to be i.i.d. random variables with distribution function $F(x) = 1 - \exp(-\mu x)$, $x \geq 0$. Let $\rho = \lambda/\mu$

$$a_k = P[\text{k services completed in an interarrival time}]$$

$$= \int_0^{\infty} \frac{e^{-\mu x}(\mu x)^k}{k!}\, dA(x) \qquad k \geq 0$$

The arrival epochs form a set of regeneration points:

let X_n = number of 1-units in the system at $\tau_n - 0$

Y_n = number of 2-units in the system at $\tau_n - 0$

$$Z_n = \begin{cases} +1 \text{ if a 1-unit is being served at } \tau_n - 0 \text{ or the system is empty at } \tau_n - 0 \\ +2 \text{ if a 2-unit is being served at } \tau_n - 0 \end{cases}$$

$$\pi_{ijk}^{n} = P\left[X_n=i,\ Y_n=j,\ Z_n=k \mid X_o=0,\ Y_o=0,\ Z_o=1\right]$$

For $|y| \leqslant 1$, $|z| \leqslant 1$, $n \geqslant 0$ define

$$\Gamma_i^{n}(z) = \sum_{j=1}^{\infty} \pi_{ij2}^{n} z^j \qquad i \geqslant 0$$

$$\pi_i^{n}(z) = \sum_{j=0}^{\infty} \pi_{ij1}^{n} z^j \qquad i \geqslant 0 \qquad \{\pi_o^{n}(z) = \pi_{001}^{n}\}$$

$$\Gamma^{n}(y,z) = \sum_{i=0}^{\infty} \Gamma_i^{n}(z)\, y^i$$

$$\pi^{n}(y,z) = \sum_{i=0}^{\infty} \pi_i^{n}(z)\, y^i$$

For $|w| < 1$ define

$$\Gamma(y,z,w) = \sum_{n=0}^{\infty} \Gamma^{n}(y,z)\, w^n$$

$$\pi(y,z,w) = \sum_{n=0}^{\infty} \pi^{n}(y,z)\, w^n$$

- regular functions of z for $|y| \leqslant 1$, $|z| \leqslant 1$, $|w| < 1$. The basic equations are as follows:-

$$\pi_{001}^{n+1} = \pi_{001}^{n} \sum_{i=1}^{\infty} a_i + \sum_{s=0}^{\infty} \sum_{i=1}^{\infty} \sum_{j=0}^{\infty} \pi_{ij1}^{n} a_{i+j+s+1}$$

$$+ \sum_{s=0}^{\infty} \sum_{i=0}^{\infty} \sum_{j=1}^{\infty} \pi_{ij2}^{n} a_{i+j+s+1} \qquad (8)$$

For $i \geq 1$, $j \geq 2$

$$\pi^{n+1}_{ij2} = q_1 a_0 \pi^n_{i-1,j2} + q_2 a_0 \pi^n_{i,j-1,2} \tag{9}$$

For $i \geq 1$, $j = 1$

$$\pi^{n+1}_{i12} = q_1 a_0 \pi^n_{i-1,12} \tag{10}$$

For $j \geq 1$

$$\pi^{n+1}_{oj2} = q_2 a_0 \pi^n_{oj-12} + q_2 \sum_{i=0}^{\infty} \pi^n_{ij2} a_{i+1} + q_2 \sum_{i=1}^{\infty} a_i \pi^n_{ij-11}$$

$$+ \sum_{i=0}^{\infty} \sum_{k=j+1}^{\infty} \pi^n_{ik2} a_{i+k+1-j} + \sum_{i=1}^{\infty} \sum_{k=j}^{\infty} \pi^n_{ik1} a_{i+k+1-j} \tag{11}$$

(where the first term in (11) is $q_2 a_0 \pi^n_{001}$ if $j=1$).

For $i \geq 1$, $j \geq 0$

$$\pi^{n+1}_{ij1} = q_1 \sum_{k=i-1}^{\infty} \pi^n_{kj1} a_{k-i+1} + q_1 \sum_{k=i-1}^{\infty} \pi^n_{k\,j+1\,2} a_{k-i+2}$$

$$+ q_2 \sum_{k=i}^{\infty} \pi^n_{k\,j-1\,1} a_{k-i} + q_2 \sum_{k=i}^{\infty} \pi^n_{kj2} a_{k-i+1} \tag{12}$$

(where the last two terms in (12) are zero if $j=0$).

From equations (9) and (10), for $i \geq 1$

$$\Gamma^{n+1}_i(z) = \sum_{j=1}^{\infty} \pi^{n+1}_{ij2} z^j$$

$$= q_1 a_0 \Gamma^n_{i-1}(z) + q_2 a_0 z \Gamma^n_i(z)$$

Therefore

$$\Gamma^{n+1}(y,z) - \Gamma_o^{n+1}(z) = q_1 a_o y \Gamma^n(y,z) + q_2 a_o z \{\Gamma^n(y,z) - \Gamma_o^n(z)\}$$

$$= a_o(q_1y+q_2z) \Gamma^n(y,z) - q_2 a_o z \Gamma_o^n(z)$$

i.e. $\Gamma(y,z,w) - \Gamma(0,z,w) - \Gamma(y,z,0) + \Gamma(0,z,0)$

$$= a_o w(q_1y+q_2z) \Gamma(y,z,w) - q_2 a_o zw \Gamma(0,z,w)$$

i.e. $\Gamma(y,z,w) \{1 - a_o w(q_1y+q_2z)\} = \Gamma(0,z,w) \{1 - q_2 a_o zw\}$ (13)

using the fact that the initial state is (0,0,1).

From equation (11),

$$\Gamma_o^{n+1}(z) = \sum_{j=1}^{\infty} \pi_{oj2}^{n+1} z^j$$

$$= q_2 a_o z \Gamma_o^n(z) + q_2 z \sum_{i=0}^{\infty} a_i \pi_i^n(z) + q_2 \sum_{i=0}^{\infty} \Gamma_i^n(z) a_{i+1}$$

$$+ \sum_{j=1}^{\infty} z^j \sum_{i=0}^{\infty} \sum_{k=j+1}^{\infty} \pi_{ik2}^n a_{i+k+1-j}$$

$$+ \sum_{j=1}^{\infty} z^j \sum_{i=1}^{\infty} \sum_{k=j}^{\infty} \pi_{ik1}^n a_{i+k+1-j}$$

Now

$$\sum_{j=1}^{\infty} z^j \sum_{k=j+1}^{\infty} \pi_{ik2}^n a_{i+k+1-j} = \sum_{\nu=0}^{\infty} \frac{a_{i+\nu+2}}{z^{\nu+1}} \sum_{j=1}^{\infty} \pi_{i,\nu+j+1,2}^n z^{\nu+j+1}$$

$$= \sum_{\nu=0}^{\infty} \frac{a_{i+\nu+2}}{z^{\nu+1}} \{\Gamma_i^n(z) - \sum_{k=1}^{\nu+1} \pi_{ik2}^n z^k\}$$

$$= \sum_{\nu=0}^{\infty} \frac{a_{i+\nu+2} \Gamma_i^n(z)}{z^{\nu+1}} - \sum_{\nu=0}^{\infty} \sum_{k=1}^{\nu+1} \frac{a_{i+\nu+2} \pi_{ik2}^n}{z^{\nu-k+1}}$$

$$= \sum_{\nu=0}^{\infty} \frac{a_{i+\nu+2} \Gamma_i^n(z)}{z^{\nu+1}} - \sum_{k=1}^{\infty} \sum_{\nu=k-1}^{\infty} \frac{a_{i+\nu+2} \pi_{ik2}^n}{z^{\nu-k+1}}$$

$$= \sum_{\nu=0}^{\infty} \frac{a_{i+\nu+2}\, \Gamma_i^{n}(z)}{z^{\nu+1}} - \sum_{s=0}^{\infty} \frac{1}{z^{s}} \sum_{k=1}^{\infty} \pi_{ik2}^{n}\, a_{i+s+k+1}$$

Similarly

$$\sum_{j=1}^{\infty} z^{j} \sum_{k=j}^{\infty} \pi_{ikl}^{n}\, a_{i+k+1-j} = \sum_{\nu=-1}^{\infty} \frac{a_{i+\nu+2}\, \pi_i^{n}(z)}{z^{\nu+1}} - \sum_{s=0}^{\infty} \frac{1}{z^{s}} \sum_{k=0}^{\infty} \pi_{ikl}^{n}\, a_{i+s+k+1}$$

Therefore

$$\Gamma_o^{n+1}(z) = q_2\, a_o\, z\, \Gamma_o^{n}(z) + q_2\, z \sum_{i=0}^{\infty} a_i\, \pi_i^{n}(z) + q_2 \sum_{i=0}^{\infty} \Gamma_i^{n}(z)\, a_{i+1}$$

$$+ \sum_{i=0}^{\infty} \sum_{k=1}^{\infty} \frac{a_{i+k+1}\, \Gamma_i^{n}(z)}{z^{k}} - \sum_{s=0}^{\infty} \frac{1}{z^{s}} \sum_{i=0}^{\infty} \sum_{k=1}^{\infty} \pi_{ik2}^{n}\, a_{i+s+k+1}$$

$$+ z \sum_{i=1}^{\infty} \sum_{k=1}^{\infty} \frac{a_{i+k}\, \Gamma_i^{n}(z)}{z^{k}} - \sum_{s=0}^{\infty} \frac{1}{z^{s}} \sum_{i=1}^{\infty} \sum_{k=0}^{\infty} \pi_{ikl}^{n}\, a_{i+s+k+1}$$

From equation (8):-

$$\pi_o^{n+1}(z) = \pi_o^{n}(z) \sum_{i=1}^{\infty} a_i + \sum_{s=0}^{\infty} \sum_{i=1}^{\infty} \sum_{k=0}^{\infty} \pi_{ikl}^{n}\, a_{i+k+s+1}$$

$$+ \sum_{s=0}^{\infty} \sum_{i=0}^{\infty} \sum_{k=1}^{\infty} \pi_{ik2}^{n}\, a_{i+k+s+1}$$

Adding these two equations gives:

$$\Gamma_o^{n+1}(z) + \pi_o^{n+1}(z) = q_2\, a_o\, z\, \Gamma_o^{n}(z) + \pi_o^{n}(z) \{ \sum_{i=1}^{\infty} a_i + q_2\, a_o\, z \}$$

$$+ \sum_{s=0}^{\infty} (1-1/z^{s}) \sum_{i=0}^{\infty} \sum_{k=1}^{\infty} \pi_{ik2}^{n}\, a_{i+s+k+1}$$

$$+ \sum_{s=0}^{\infty} (1-1/z^{s}) \sum_{i=1}^{\infty} \sum_{k=0}^{\infty} \pi_{ikl}^{n}\, a_{i+s+k+1}$$

$$+ \frac{q_1}{z} \sum_{i=0}^{\infty} \sum_{\nu=0}^{\infty} \frac{a_{i+\nu+2}\, \Gamma_i^{n}(z)}{z^{\nu}} + q_2 \sum_{i=0}^{\infty} \sum_{\nu=0}^{\infty} \frac{a_{i+\nu+1}\, \Gamma_i^{n}(z)}{z^{\nu}}$$

$$+ q_1 \sum_{i=1}^{\infty} \sum_{\nu=0}^{\infty} \frac{a_{i+\nu+1}\, \pi_i^n(z)}{z^\nu} + q_2 z \sum_{i=1}^{\infty} \sum_{\nu=0}^{\infty} \frac{a_{i+\nu}\, \pi_i^n(z)}{z^\nu} \qquad (14)$$

The final equation, equation (12), gives for every $i \geq 1$:-

$$\pi_i^{n+1}(z) = q_1 \sum_{k=0}^{\infty} a_k\, \pi_{k+i-1}^n(z) + \frac{q_1}{z} \sum_{k=1}^{\infty} a_k\, \Gamma_{k+i-2}^n(z)$$

$$+ q_2 z \sum_{k=0}^{\infty} a_k\, \pi_{k+i}^n(z) + q_2 \sum_{k=1}^{\infty} a_k\, \Gamma_{k+i-1}^n(z)$$

Thus

$$\sum_{i=1}^{\infty} \pi_i^{n+1}(z)\, y^i = q_1 \sum_{k=0}^{\infty} \frac{a_k}{y^{k-1}} \{\pi^n(y,z) - \sum_{i=0}^{k-1} y^i\, \pi_i^n(z)\}$$

$$+ q_2 z \sum_{k=0}^{\infty} \frac{a_k}{y^k} \{\pi^n(y,z) - \sum_{i=0}^{k} y^i\, \pi_i^n(z)\}$$

$$+ \frac{q_1}{z} \sum_{k=1}^{\infty} \frac{a_k}{y^{k-2}} \{\Gamma^n(y,z) - \sum_{i=0}^{k-2} y^i\, \Gamma_i^n(z)\}$$

$$+ q_2 \sum_{k=1}^{\infty} \frac{a_k}{y^{k-1}} \{\Gamma^n(y,z) - \sum_{i=0}^{k-1} y^i\, \Gamma_i^n(z)\}$$

$$= \pi^n(y,z)\,(q_1 y + q_2 z) \sum_{k=0}^{\infty} \frac{a_k}{y^k} + \frac{y}{z}\, \Gamma^n(y,z)\,(q_1 y + q_2 z) \sum_{k=1}^{\infty} \frac{a_k}{y^k}$$

$$- q_1 \sum_{k=1}^{\infty} \sum_{\nu=0}^{k-1} \frac{a_k\, \pi_{k-\nu-1}^n(z)}{y^\nu} - q_2 z \sum_{k=0}^{\infty} \sum_{\nu=0}^{k} \frac{a_k\, \pi_{k-\nu}^n(z)}{y^\nu}$$

$$- \frac{q_1}{z} \sum_{k=2}^{\infty} \sum_{\nu=0}^{k-2} \frac{a_k\, \Gamma_{k-\nu-2}^n(z)}{y^\nu} - q_2 \sum_{k=1}^{\infty} \sum_{\nu=0}^{k-1} \frac{a_k\, \Gamma_{k-\nu-1}^n(z)}{y^\nu}$$

$$= \pi^n(y,z)\,(q_1 y + q_2 z) \sum_{k=0}^{\infty} \frac{a_k}{y^k} + \frac{y}{z}\, \Gamma^n(y,z)\,(q_1 y + q_2 z) \sum_{k=1}^{\infty} \frac{a_k}{y^k}$$

$$- q_1 y\, \pi_o^n(z) \sum_{k=1}^{\infty} \frac{a_k}{y^k} - q_2 z\, \pi_o^n(z) \sum_{k=0}^{\infty} \frac{a_k}{y^k}$$

$$- q_1 \sum_{\nu=0}^{\infty} \frac{1}{y^\nu} \sum_{k=\nu+2}^{\infty} a_k \pi_{k-\nu-1}^{n}(z) - q_2 z \sum_{\nu=0}^{\infty} \frac{1}{y^\nu} \sum_{k=\nu+1}^{\infty} a_k \pi_{k-\nu}^{n}(z)$$

$$- \frac{q_1}{z} \sum_{\nu=0}^{\infty} \frac{1}{y^\nu} \sum_{k=\nu+2}^{\infty} a_k \Gamma_{k-\nu-2}^{n}(z) - q_2 \sum_{\nu=0}^{\infty} \frac{1}{y^\nu} \sum_{k=\nu+1}^{\infty} a_k \Gamma_{k-\nu-1}^{n}(z) \tag{15}$$

Adding equations (14) and (15) gives

$$\pi^{n+1}(y,z) + \Gamma_o^{n+1}(z)$$

$$= \pi^n(y,z)\,(q_1 y + q_2 z) \sum_{k=0}^{\infty} \frac{a_k}{y^k} + \frac{y}{z}\,\Gamma^n(y,z)\,(q_1 y + q_2 z) \sum_{k=1}^{\infty} \frac{a_k}{y^k}$$

$$+ q_2 a_o z\,\Gamma^n(0,z) + q_1 \pi_o^n(z) \sum_{i=0}^{\infty} a_{i+1}(1-1/y^i)$$

$$+ q_2 \pi_o^n(z) \sum_{i=0}^{\infty} a_{i+1}(1-1/z^i) + \sum_{s=0}^{\infty} (1-1/z^s)\, C_s^n$$

$$+ \frac{q_1}{z} \sum_{\nu=0}^{\infty} \left(\frac{1}{z^\nu} - \frac{1}{y^\nu}\right) \sum_{k=\nu+2}^{\infty} a_k \Gamma_{k-\nu-2}^{n}(z) + q_2 \sum_{\nu=0}^{\infty} \left(\frac{1}{z^\nu} - \frac{1}{y^\nu}\right) \sum_{k=\nu+1}^{\infty} a_k \Gamma_{k-\nu-1}^{n}(z)$$

$$+ q_1 \sum_{\nu=0}^{\infty} \left(\frac{1}{z^\nu} - \frac{1}{y^\nu}\right) \sum_{k=\nu+2}^{\infty} a_k \Gamma_{k-\nu-1}^{n}(z) + q_2 z \sum_{\nu=0}^{\infty} \left(\frac{1}{z^\nu} - \frac{1}{y^\nu}\right) \sum_{k=\nu}^{\infty} a_k \pi_{k-\nu}^{n}(z)$$

where

$$C_s^n = \sum_{i=0}^{\infty} \sum_{k=1}^{\infty} \pi_{ik2}^{n} a_{i+s+k+1} + \sum_{i=1}^{\infty} \sum_{k=0}^{\infty} \pi_{ik1}^{n} a_{i+s+k+1}$$

This can be written

$$\pi^{n+1}(y,z) + \Gamma_o^{n+1}(z) = \pi^n(y,z)\,(q_1 y + q_2 z) \sum_{k=0}^{\infty} \frac{a_k}{y^k}$$

$$+ \frac{y}{z}\,\Gamma^n(y,z)\,(q_1 y + q_2 z) \sum_{k=1}^{\infty} \frac{a_k}{y^k} + q_2 a_o z\,\Gamma^n(0,z)$$

$$+ \sum_{s=0}^{\infty} (1-1/z^s)\, C_s^{\,n} + q_1\, \pi_o^{\,n}(z) \sum_{i=0}^{\infty} a_{i+1}\, (1-1/y^i)$$

$$+ q_2\, \pi_o^{\,n}(z) \sum_{i=0}^{\infty} a_{i+1}\, (1-1/z^i) + \sum_{i=0}^{\infty} \left(\frac{1}{z^i} - \frac{1}{y^i} \right) D_i^{\,n}(z)$$

where

$$D_i^{\,n}(z) = q_1 \sum_{k=i+2}^{\infty} a_k\, \Gamma_{k-i-1}^{n}(z) + q_2 z \sum_{k=i+1}^{\infty} a_k\, \pi_{k-i}^{n}(z)$$

It follows that for $|y| = 1$, $|z| = 1$, $|w| < 1$ (recall that the initial state is (0,0,1)):-

$$\pi(y,z,w) + \Gamma(0,z,w) - 1 = w(q_1 y + q_2 z)\, \pi(y,z,w) \sum_{k=0}^{\infty} \frac{a_k}{y^k}$$

$$+ \frac{wy}{z}\, \Gamma(y,z,w)(q_1 y + q_2 z) \sum_{k=1}^{\infty} \frac{a_k}{y^k} + q_2\, a_o z\, w\, \Gamma(0,z,w)$$

$$+ \sum_{s=0}^{\infty} (1-1/z^s)\, C_s(w) + q_1 w\, \pi(0,z,w) \sum_{i=0}^{\infty} a_{i+1}\, (1-1/y^i)$$

$$+ q_2 w\, \pi(0,z,w) \sum_{i=0}^{\infty} a_{i+1}\, (1-1/z^i) + \sum_{i=0}^{\infty} \left(\frac{1}{z^i} - \frac{1}{y^i} \right) D_i(z,w)$$

where

$$C_s(w) = w \sum_{n=0}^{\infty} C_s^{\,n}\, w^n \quad \text{and} \quad D_i(z,w) = w \sum_{n=0}^{\infty} D_i^{\,n}(z)\, w^n$$

i.e.

$$\pi(y,z,w)\, \{1 - w(q_1 y + q_2 z) \sum_{k=0}^{\infty} \frac{a_k}{y^k}\} = 1 + \Gamma(0,z,w)\, \{q_2\, a_o z w - 1\}$$

$$+ \frac{wy}{z}\, \Gamma(y,z,w)\, (q_1 y + q_2 z) \sum_{k=1}^{\infty} \frac{a_k}{y^k} + \sum_{s=0}^{\infty} (1-1/z^s)\, C_s(w)$$

$$+ q_1 w\, \pi(0,z,w) \sum_{i=0}^{\infty} a_{i+1}\, (1-1/y^i) + q_2 w\, \pi(0,z,w) \sum_{i=0}^{\infty} a_{i+1}\, (1-1/z^i)$$

$$+ \sum_{i=0}^{\infty} \left(\frac{1}{z^i} - \frac{1}{y^i} \right) D_i(z,w)$$

Using equation (13) gives for $|y| = 1$, $|z| = 1$, $|w| < 1$

$$\pi(y,z,w)\ \{1 - w(q_1y+q_2z) \sum_{k=0}^{\infty} \frac{a_k}{y^k}\}$$

$$= 1 + \Gamma(y,z,w)\ \{a_o w(q_1y+q_2z) + \frac{wy}{z}(q_1y+q_2z) \sum_{k=1}^{\infty} \frac{a_k}{y^k} - 1\}$$

$$+ q_1 w\ \pi(0,z,w) \sum_{i=0}^{\infty} a_{i+1}(1-1/y^i) + q_2 w\ \pi(0,z,w) \sum_{i=0}^{\infty} a_{i+1}(1-1/z^i)$$

$$+ \sum_{s=0}^{\infty} (1-1/z^s)\ C_s(w) + \sum_{i=0}^{\infty} (1/z^i - 1/y^i)\ D_i(z,w) \qquad (16)$$

Note that $\pi(0,z,w) = \sum_{n=0}^{\infty} \pi^n(0,z)w^n = \sum_{n=0}^{\infty} \pi^n_{001} w^n = \pi(0,0,w)$ which is independent of z.

Define $\pi(y,z,w)$ by this expression for $|y| > 1$ also. As it is known that $\pi(y,z,w)$ is a regular function of y for $|y| \leqslant 1$ ($|z|=1$, $|w|<1$), its only singularities in the whole complex plane are the zeros of the denominator and the singularities of the numerator outside the unit circle $|y| = 1$.

From equation (13), the numerator has its only singularity at

$$q_1\ y = (1/a_o w) - q_2\ z$$

which is outside the unit circle.

<u>Lemma</u> The equation

$$y = w(q_1+q_2zy)\ \bar{A}(\mu-\mu y) \qquad (17)$$

has a unique root in y within the unit circle $|y| < 1$ if $|w| < 1$ and $|z| \leqslant 1$, or $|w| \leqslant 1$ and $|z| < 1$, or $|w| \leqslant 1$ $|z| \leqslant 1$ and $\mu/\lambda q_1 < 1$. If this root is denoted by $\gamma(z,w)$ then $\gamma = \gamma(1,1)$ is the smallest positive real root of the equation

$$y = (q_1+q_2y)\,\bar{A}(\mu-\mu y)$$

If $\mu/\lambda q_1 > 1$ then $\gamma < 1$; if $\mu/\lambda q_1 \leq 1$ then $\gamma = 1$.

Proof If $|w| < 1$ and $|z| \leq 1$, or $|w| \leq 1$ and $|z| < 1$ then

$$|w(q_1+q_2zy)\,\bar{A}(\mu-\mu y)| < 1 - \varepsilon \quad \text{if} \quad |y| = 1 - \varepsilon$$

and ε is a sufficiently small positive number.

If $|w| \leq 1$, $|z| \leq 1$ and $\mu/\lambda q_1 > 1$

$$|w(q_1+q_2zw)\,\bar{A}(\mu-\mu y)| \leq |\bar{A}(\mu-\mu y)|\,\{q_1 + q_2|y|\}$$

$$\leq \bar{A}(\mu-\mu|y|)\,\{q_1 + q_2|y|\}$$

$$= \bar{A}(\varepsilon\mu)\,\{1 - q_2\varepsilon\} \quad \text{if} \quad |y| = 1 - \varepsilon$$

$$= \{1 - \mu\varepsilon/\lambda + o(\varepsilon)\}\,\{1 - q_2\varepsilon\}$$

$$= 1 - \varepsilon\,(\mu/\lambda+q_2) + o(\varepsilon)$$

and $\mu/\lambda + q_2 > 1$ as $\mu/\lambda q_1 > 1$. Therefore, if ε is a sufficiently small positive number

$$|w(q_1+q_2zy)\,\bar{A}(\mu-\mu y)| < 1 - \varepsilon$$

Hence, by Rouché's theorem, (17) has exactly one root in the circle $|z| < 1 - \varepsilon$ where ε is a sufficiently small positive number.

If $z = 1$ and $w = 1$ consider the equation

$$f_1(y) = f_2(y)$$

where $f_1(y) = y$ and $f_2(y) = (q_1+q_2y)\,\bar{A}(\mu-\mu y)$.

Now

$$f_1(0) = 0 \qquad \text{and} \qquad f_2(0) = q_1 \bar{A}(\mu) > 0$$

$$f_1(1) = 1 \qquad \text{and} \qquad f_2(1) = \bar{A}(0) = 1$$

Also $\quad f_1'(y) = 1 \qquad f_2'(y) = q_2 \bar{A}(\mu-\mu y) - (q_1+q_2 y)\mu \bar{A}'(\mu-\mu y)$

Thus $\quad f_1'(1) = 1$ and $f_2'(1) = q_2 + \mu/\lambda$

Also for $|y| \leq 1 \qquad f_2'(y) \leq q_2 + \mu/\lambda = 1 + (\mu/\lambda - q_1)$

i.e. if $\mu/\lambda q_1 \leq 1$ then $f_2'(y) \leq f_1'(y)$ for every $|y| \leq 1$ with equality at $y = 1$.

It follows that

$$\text{if} \quad \mu/q\lambda_1 \leq 1 \qquad \text{then} \quad \gamma = 1$$

$$\text{if} \quad \mu/q\lambda_1 > 1 \qquad \text{then} \quad \gamma < 1$$

This completes the proof of the lemma.

Using this lemma shows that the only root of the denominator of (16) outside $|y| = 1$ is

$$y = 1/\gamma(z,w)$$

Define

$$\Psi(y,z,w) = \{y - 1/\gamma(z,w)\}\{q_1 y + q_2 z - 1/a_0 w\}\pi(y,z,w) \qquad (18)$$

- a regular function of y in the whole complex plane. As

$$\lim_{|y|\to\infty} \frac{\Psi(y,z,w)}{|y|^2} = 0$$

it follows that $\Psi(y,z,w)$ is a linear function of y; i.e.

$$\Psi(y,z,w) = \Psi(0,z,w) + y\,\Psi_1(z,w) \qquad (19)$$

From equation (18),

$$\Psi(0,z,w) = \{-1/\gamma(z,w)\}\{q_2 z - 1/a_0 w\}\,\pi(0,z,w)$$

$$= \{-1/\gamma(z,w)\}\{q_2 z - 1/a_0 w\}\,\pi(0,0,w) \qquad (20)$$

From equation (16), setting $z = y$ gives

$$\pi(y,y,w) + \Gamma(y,y,w) = \frac{1+\sum_{i=0}^{\infty}(1-1/y^i)\,E_i(w)}{1-wy\sum_{k=0}^{\infty}(a_k/y^k)}$$

where $E_i(w) = wa_{i+1}\,\pi(0,0,w) + C_i(w)$

- the same equation as that for the ordinary GI/M/1 queue. Hence

$$\pi(y,y,w) + \Gamma(y,y,w) = \frac{1-\delta(w)}{\{1-y\,\delta(w)\}\{1-w\}} \qquad (21)$$

where $\delta(w)$ is the unique root in z within the unit circle of the equation

$$z = w\,\tilde{A}(\mu-\mu z)$$

Setting $y = 0$ and noting that $\Gamma(0,0,w) = 0$ gives

$$\pi(0,0,w) = \frac{1-\delta(w)}{1-w}$$

Substituting in (20) gives

$$\Psi(0,z,w) = \left\{-\frac{1}{\gamma(z,w)}\right\}\left\{q_2 z - \frac{1}{a_0 w}\right\}\frac{1-\delta(w)}{1-w} \qquad (22)$$

Consider $\Psi(y_0,z,w)$ where $q_2 y_0 = 1/a_0 w - q_2 z$, y_0 being a function of z and w. Using (18), (16) and (13):-

$$\Psi(y_0,z,w) = \frac{\{y_0 - 1/\gamma(z,w)\}\{1-q_2 a_0 z w\}\,\Gamma(0,z,w)\left\{\frac{y_0}{z\,a_0}\sum_{k=1}^{\infty}\frac{a_k}{y_0^k}\right\}}{-a_0 w\left\{1-\frac{1}{a_0}\sum_{k=0}^{\infty}\frac{a_k}{y_0^k}\right\}}$$

$$= \frac{y_o\{y_o\,\gamma(z,w)-1\}\{1-q_2a_o zw\}\,\Gamma(0,z,w)}{wz\,a_o\,\gamma(z,w)}$$

$$= \frac{\{y_o\,\gamma(z,w)-1\}\,\Gamma(0,z,w)\,y_o^{\,2}\,q_1}{z\,\gamma(z,w)}$$

Now

$$\mathcal{H}(y,z,w) = \{\,\mathcal{H}(0,z,w)\,(y_o-y) + y\,\mathcal{H}(y_o,z,w)\}\,/\,y_o$$

$$= \frac{\{1-\delta(w)\}\,q_1(y_o-y)}{\{1-w\}\,\gamma(z,w)} + \frac{y\{y_o\,\gamma(z,w)-1\}\,\Gamma(0,z,w)\,q_1\,y_o}{z\,\gamma(z,w)}$$

Thus

$$\pi(y,z,w) = \frac{\mathcal{H}(y,z,w)\,\gamma(z,w)}{\{y\,\gamma(z,w)-1\}\,q_1(y-y_o)}$$

$$= \frac{1-\delta(w)}{\{1-w\}\{1-y\,\gamma(z,w)\}} + \frac{y\,y_o\,\Gamma(0,z,w)\,\{y_o\,\gamma(z,w)-1\}}{z\{y\,\gamma(z,w)-1\}\{y-y_o\}}$$

Using (13),

$$\pi(y,z,w) = \frac{1-\delta(w)}{\{1-w\}\{1-y\,\gamma(z,w)\}} + \frac{y\,\Gamma(y,z,w)\,\{1-y_o\,\gamma(z,w)\}}{z\{y\,\gamma(z,w)-1\}} \qquad (23)$$

Setting z = y gives

$$\pi(y,y,w) = \frac{1-\delta(w)}{\{1-w\}\{1-y\,\gamma(z,w)\}} - \frac{\Gamma(y,y,w)\,\{q_1a_o w-(1-q_2a_o wy)\,\gamma(y,w)\}}{q_1a_o w\{1-y\,\gamma(y,w)\}}$$

But by (21) we also have

$$\pi(y,y,w) = \frac{1-\delta(w)}{\{1-w\}\{1-y\,\delta(w)\}} - \Gamma(y,y,w)$$

It follows that

$$\Gamma(y,y,w)\left\{1 - \frac{q_1a_o w-(1-q_2a_o wy)\,\gamma(y,w)}{q_1a_o w\left[1-y\,\gamma(y,w)\right]}\right\}$$

$$= \frac{y\{1-\delta(w)\}\{\delta(w)-\gamma(y,w)\}}{\{1-w\}\{1-y\ \delta(w)\}\{1-y\ \gamma(y,w)\}}$$

i.e.

$$\Gamma(y,y,w) = \frac{q_1 a_o wy\ \{1-\delta(w)\}\{\delta(w)-\gamma(y,w)\}}{\{1-w\}\{1-y\ \delta(w)\}\{1-a_o wy\}\ \gamma(y,w)}$$

Using (13):-

$$\Gamma(y,z,w) = \frac{1-q_2 a_o zw}{1-a_o w(q_1 y+q_2 z)}\ \Gamma(0,z,w)$$

$$= \frac{\{1-q_2 a_o zw\}\{1-a_o wz\}\ \Gamma(z,z,w)}{\{1-a_o w(q_1 y+q_2 z)\}\{1-q_2 a_o zw\}}$$

i.e.

$$\Gamma(y,z,w) = \frac{q_1 a_o wz\ \{1-\delta(w)\}\{\delta(w)-\gamma(z,w)\}}{\{1-w\}\{1-z\ \delta(w)\}\{1-a_o w(q_1 y+q_2 z)\}\ \gamma(z,w)} \qquad (24)$$

(23) and (24) give a complete solution for the generating functions of queue length probabilities.

Stationary Distribution

The Markov chain is irreducible and aperiodic, and therefore the limiting probabilities

$$\pi_{ijk} = \lim_{n\to\infty} \pi_{ijk}^{n}$$

always exist; either every $\pi_{ijk} = 0$ or every $\pi_{ijk} > 0$ and $\{\pi_{ijk}\}$ is a probability distribution. Using Abel's theorem, if $\rho = \lambda/\mu \geq 1$

$$\Gamma(y,z) = \lim_{w\to 1} (1-w)\ \Gamma(y,z,w) = 0$$

and

$$\pi(y,z) = \lim_{w\to 1} (1-w)\ \pi(y,z,w) = 0$$

If $\rho = \lambda/\mu < 1$

$$\Gamma(y,z) = \lim_{w\to 1} (1-w)\,\Gamma(y,z,w)$$

$$= \frac{q_1 a_o z\,\{1-\delta\}\{\delta-\gamma(z,1)\}}{\{1-\delta z\}\{1-a_o(q_1 y+q_2 z)\}\,\gamma(z,1)}$$

and

$$\pi(y,z) = \frac{1-\delta}{1-\gamma(z,1)y} + \frac{y\{q_1 a_o-(1-a_o q_2 z)\,\gamma(z,1)\}\,\Gamma(y,z)}{q_1 a_o z\,\{y\,\gamma(z,1)-1\}}$$

Setting $z = 1$ gives the generating functions of the number of 1-units in the system

$$\Gamma(y,1) = \frac{q_1 a_o\,\{\delta-\gamma\}}{\{1-a_o(q_1 y+q_2)\}\gamma} \tag{25}$$

and

$$\pi(y,1) = \frac{1-\delta}{1-\gamma y} - \frac{y\{q_1 a_o-(1-a_o q_2)\gamma\}\,\Gamma(y,1)}{q_1 a_o\,\{1-\gamma y\}} \tag{26}$$

Note that as $\rho < 1$, so $\lambda q_1/\mu = q_1\rho < 1$ and hence $\gamma < 1$.

Special Case M/M/1

Suppose $\bar{A}(s) = \lambda/(\lambda+s)$

Then δ is a solution of the equation $z = \bar{A}(\mu-\mu z)$

i.e. $z\lambda + z\mu - z^2\mu = \lambda$ leading to $z = \rho$ or 1

Therefore, for $\rho < 1$, $\delta = \rho$ (27)

γ is a solution of the equation $y = (q_1+q_2 y)\,\bar{A}(\mu-\mu y)$

i.e. $y\lambda + y\mu - y^2\mu = \lambda q_1 + \lambda q_2 y$

i.e. $\mu y(1-y) = \lambda q_1(1-y)$

leading to $y = 1$ or $\lambda q_1/\mu$

Therefore $\gamma = \lambda q_1/\mu = q_1\rho$ (28)

4.3 The GI/M/1 Priority Queue: the Waiting Time Distributions

The Stationary Waiting Time Distribution of a 1-unit

Let W_1 denote the waiting time of a 1-unit in the stationary state. Then the state of the system at the arrival epoch, τ, of the 1-unit can be partitioned into a number of mutually exclusive events:

A_o - the system empty at τ

A_i - i 1-units in the system, one of which is being served $(i \geq 1)$

B_i - i 1-units in the system, but a 2-unit is being served $(i \geq 0)$

Therefore

$$E[e^{-sW_1}] = \sum_{i=0}^{\infty} E[e^{-sW_1}|A_i]\, P[A_i] + \sum_{i=0}^{\infty} E[e^{-sW_1}|B_i]\, P[B_i]$$

$$= \sum_{i=0}^{\infty} \left(\frac{\mu}{\mu+s}\right)^i P[A_i] + \sum_{i=0}^{\infty} \left(\frac{\mu}{\mu+s}\right)^{i+1} P[B_i]$$

$$= \pi\left(\frac{\mu}{\mu+s}, 1\right) + \frac{\mu}{\mu+s}\, \Gamma\left(\frac{\mu}{\mu+s}, 1\right)$$

From (25) and (26),

$$\pi(y,1) + y\,\Gamma(y,1) = \frac{1-\delta}{1-\gamma y} + y\,\Gamma(y,1)\left\{1 - \frac{q_1 a_o - (1-a_o q_2)\gamma}{q_1 a_o (1-\gamma y)}\right\}$$

$$= \frac{1-\delta}{1-\gamma y} + \gamma y\,\Gamma(y,1)\left\{\frac{(1-a_o q_2) - q_1 a_o y}{q_1 a_o (1-\gamma y)}\right\}$$

$$= \frac{1-\delta}{1-\gamma y} + \frac{y\{\delta-\gamma\}}{1-\gamma y}$$

Therefore

$$E[e^{-sW_1}] = \frac{(\mu+s)(1-\delta)}{(\mu+s-\gamma\mu)} + \frac{\mu(\delta-\gamma)}{(\mu+s-\gamma\mu)} \qquad (29)$$

Special Case: M/M/1

Using (27) and (28), (29) simplifies to

$$E[e^{-sW_1}] = \frac{(\mu+s)(1-\rho)}{(\mu+s-q_1\rho\mu)} + \frac{\mu q_2 \rho}{(\mu+s-q_1\rho\mu)}$$

agreeing with Miller [14].

Differentiating (29) gives

$$E[W_1] = \frac{\delta}{\mu(1-\gamma)} \tag{30}$$

Mean Waiting Time of 2-units

Using equation (30) of Chapter 3:

$$\rho\, E[W] = q_1\rho\, E[W_1] + q_2\rho\, E[W_2]$$

Now $$E[W] = \delta/(\mu-\mu\delta)$$

Thus

$$E[W_2] = \frac{\delta}{q_2\mu}\left\{\frac{1}{1-\delta} - \frac{q_1}{1-\gamma}\right\}$$

i.e. $$E[W_2] = \frac{\delta\{q_2+q_1\delta-\gamma\}}{q_2\mu\{1-\delta\}\{1-\gamma\}} \tag{31}$$

The Stationary Waiting Time Distribution of a 2-unit

The results of the last section of Chapter 3 can be used with

$$1 - \rho\,\phi(s) = \frac{\mu+q+s-\rho\mu\,\bar{A}_1(-s)}{\mu+q+s}$$

CHAPTER 5

Inequalities

5.1 Inequalities for the GI/G/1 Queue

The problem of obtaining bounds for the mean waiting time in a GI/G/1 queue has been considered by Kingman [10,11 where references are also given to earlier work] and Marshall [12,13]. A summary of their work will now be given and then in the next section this will be extended to derive bounds for the GI/G/1 priority queue.

Suppose customers $C_0, C_1, C_2, \ldots$ arrive at a single server queue where they are served in the order of their arrival. For $n \geq 0$ let

A_n = interarrival time between the n^{th} and $(n+1)^{th}$ arrival epochs; $E[A_n] < \infty$

S_n = service time of C_n; $E[S_n] < \infty$

W_n = waiting time of C_n; initial condition W_0 with $E[W_0] < \infty$

$U_n = S_n - A_n$

where the A_n and S_n $(n \geq 0)$ are all mutually independent. Then

$$W_{n+1} = [W_n + U_n]^+ \qquad (1)$$

where

$$X^+ = \text{Max}(X,0) = \begin{cases} X & \text{if } X > 0 \\ 0 & \text{if } X \leq 0 \end{cases}$$

If the traffic intensity $\rho = E[S]/E[A] < 1$ then it is well known (Lindley's Theorem) that the distribution of W_n converges to that of a finite random variable W regardless of W_0.

Now, for any $n \geq 0$

$$E[(W_n+U_n)^+] - E[(W_n+U_n)^-] = E[W_n + U_n] = E[W_n] + E[S_n] - E[A_n]$$

where

$$X^- = - \text{Min}(X,0) = \begin{cases} -X & \text{if } X \leqslant 0 \\ 0 & \text{if } X > 0 \end{cases}$$

$$= X^+ - X$$

Taking the limit as $n \to \infty$ gives, on using (1)

$$E[(W+U)^-] = E[A] - E[S]$$

$$= - E[U] \qquad (2)$$

If Z is any random variable with finite variance, then as

$$Z = Z^+ - Z^-$$

and

$$Z^2 = (Z^+)^2 + (Z^-)^2$$

we have

$$E[Z] = E[Z^+] - E[Z^-]$$

and $$E[Z^2] = E[(Z^+)^2] + E[(Z^-)^2]$$

i.e. $$\text{var}[Z^+] + \text{var}[Z^-] = \text{var}[Z] - 2\, E[Z^+]\, E[Z^-]$$

Assuming $\text{var}[W]$ is finite

$$\text{var}[(W+U)^+] + \text{var}[(W+U)^-]$$

$$= \text{var}[W+U] - 2\, E[(W+U)^+]\, E[(W+U)^-]$$

$$= \text{var}[W] + \text{var}[U] - 2\, E[W]\, E[(W+U)^-]$$

i.e. $$E[W] = \frac{var[U]-var[(W+U)^-]}{2\{E[A]-E[S]\}}$$

$$\leqslant \frac{var[U]}{2\{E[A]-E[S]\}} \qquad (3)$$

and hence

$$E[W] \leqslant \frac{var[S]+var[A]}{2\{E[A]-E[S]\}}$$

$$= J \quad \text{say} \qquad (4)$$

the error being caused by neglection of the variance of $(W+U)^-$.

As $W \geqslant 0$,

$$[(W+U)^-]^2 \leqslant (U^-)^2$$

and therefore

$$var[(W+U)^-] = E[(W+U)^-]^2 - \{E[A] - E[S]\}^2$$

$$\leqslant E[(U^-)^2] - \{E[U]\}^2$$

$$= E[U^2 - (U^+)^2] - \{E[U]\}^2$$

$$= var[U] - E[(U^+)^2]$$

From equation (3)

$$E[W] \geqslant \frac{E[(U^+)^2]}{2\{E[A]-E[S]\}} \qquad (5)$$

Summarizing

$$\frac{E[(U^+)^2]}{2\{E[A]-E[S]\}} \leqslant E[W] \leqslant \frac{var[S]+var[A]}{2\{E[A]-E[S]\}}$$

Note that the lower bound is necessarily more complicated than the

upper one, because if it depended only on the parameters $E[S]$, $E[A]$, $\mathrm{var}[S]$, $\mathrm{var}[A]$ subject only to $E[S] < E[A]$ it would have to be zero [consider the queue D/D/1 with $S_n = E[S]$, $A_n = E[A]$, $E[S] < E[A]$ and $W_o = 0$; then $U_n = S_n - A_n < 0$ for all n and hence $W_n = 0$ for all n. Note that equality holds for the upper bound for this queue].

The above inequalities are due to J. F. C. Kingman. Another more complicated lower bound has been given by Marshall [12]. For all $w \geqslant 0$,

$$E[W_{n+1} \mid W_n = w] = E[(w+U_n)^+]$$

$$= \int_{-w}^{\infty} P[U_n > x]\,dx$$

$$= g(w) \qquad \text{say}$$

and therefore $$E[W_{n+1}] = E[g(W_n)]$$

or, in the limit

$$E[W] = E[g(W)]$$

As $g'(x) = P[U > -x]$ which increases monotonically as x increases ($x>0$), $g(x)$ is a convex function. Using Jensen's inequality, it follows that

$$E[W] = E[g(W)]$$

$$\geqslant g(E[W])$$

i.e. $$E[W] \geqslant \int_{-E[W]}^{\infty} P[U > x]\,dx \qquad (6)$$

Consider the equation

$$x = \int_{-x}^{\infty} P[U > u]\,du$$

$$= E[U^+] + \int_{-x}^{0} P[U > u]\,du \qquad (7)$$

as

$$E[U^+] = \int_0^\infty P[U > u]du$$

If $E[U^+] = 0$ then $x = 0$ is a solution of (7).

If $E[U^+] > 0$ and $E[S] < E[A]$ then

$$E[U] = E[U^+] - E[U^-] < 0$$

i.e. $$E[U^+] < E[U^-]$$

$$E[U^+] < \int_{-\infty}^0 P[U \leqslant u]du$$

Therefore, for x sufficiently large,

$$E[U^+] < \int_{-x}^0 P[U \leqslant u]du$$

$$< x - \int_{-x}^0 P[U > u]du$$

i.e. $$x > E[U^+] + \int_{-x}^0 P[U > u]du$$

and hence (7) has a solution. Similarly if $E[S] > E[A]$ then (7) has no solution.

If x is a solution of (7) over some range $[a,b]$ say, then

$$g'(x) = P[U > -x] = 1 \quad \text{for } x\varepsilon[a,b]$$

and hence $g'(x) = 1$ for $x\varepsilon[a,\infty)$, and therefore the curves never cross, only meet. As $g(x)$ is convex, it follows that if $E[S] < E[A]$, equation (7) has a unique root ℓ, say.

If $\ell = 0$, then trivially $E[W] \geqslant \ell$

If $\ell > 0$, then $E[U^+] > 0$ and for all $x\varepsilon[0,\ell)$

$$x < E[U^+] + \int_{-x}^0 P[U > u]du$$

Hence if $E[W] < \ell$, then

$$E[W] < E[U^+] + \int_{-E[W]}^{0} P[U > u]du$$

$$= \int_{-E[W]}^{\infty} P[U > u]du$$

contradicting (6). Therefore $E[W] \geq \ell$.

Summarizing, if $E[S] < E[A]$ then $E[W] \geq \ell$ where ℓ is the unique root of the equation

$$x = \int_{-x}^{\infty} P[U > u]du \qquad (x \geq 0) \qquad (8)$$

Marshall has also given improvements of this lower bound when restrictions are placed on the distribution of interarrival times. There are three possible assumptions:-

(i) Suppose A has its mean residual life bounded above by γ ($\gamma < \infty$)

i.e. $$\frac{\int_t^{\infty} P[A>u]du}{P[A>t]} \leq \gamma, \qquad \text{for every } t \geq 0$$

Let I denote the length of an idle period; then I can be expressed in the form $I = A - X$ where $X > 0$. Using the notation $f_Z(t)$ to denote the density function of any random variable Z and $Z(t)$ to denote its distribution function:-

$$\frac{f_I(t)}{P[I>t]} = \frac{1}{P[I>t]} \int_0^{\infty} f_A(t+x)\, dX(x)$$

$$\int_t^{\infty} P[I > u]du = \int_{u=t}^{\infty} P[A > X + u]du$$

$$= \int_{u=t}^{\infty} \int_0^{\infty} P[A > u + x]\, dX(x)\, du$$

$$= \int_{x=0}^{\infty} P[A > t + x] \int_{v=t+x}^{\infty} \frac{P[A>v]dv}{P[A>t+x]} \, dX(x)$$

$$\leqslant \gamma \int_{x=0}^{\infty} P[A > t + x] \, dX(x)$$

i.e. $$\frac{\int_t^{\infty} P[I>u]du}{E[I]} \leqslant \gamma \frac{P[I>t]}{E[I]} \qquad \text{for every } t \geqslant 0$$

Integrating over t gives

$$\frac{E[I^2]}{2\,E[I]} \leqslant \gamma \tag{9}$$

(ii) Suppose A has decreasing mean residual life

i.e. $$\frac{\int_t^{\infty} P[A>u]du}{P[A>t]}$$ decreases monotonically as t increases

($t \geqslant 0$, $P[A > t] > 0$). Note that decreasing mean residual life implies mean residual life bounded above by $E[A]$.

As for (i)

$$\int_t^{\infty} P[I > u]du = \int_{x=0}^{\infty} P[A > t + x] \int_{v=t+x}^{\infty} \frac{P[A>u]}{P[A>t+x]} du \, dX(x)$$

$$\leqslant \int_{x=0}^{\infty} P[A > t + x] \int_{v=t}^{\infty} \frac{P[A>u]}{P[A>t]} du \, dX(x)$$

i.e. $$\frac{\int_t^{\infty} P[I>u]du}{P[I>t]} \leqslant \int_{v=t}^{\infty} \frac{P[A>u]}{P[A>t]} du, \text{ for every } t \geqslant 0 \tag{10}$$

(iii) Suppose A has increasing failure rate

i.e. $\dfrac{f_A(t)}{P[A>t]}$ increases monotonically as t increases ($t \geq 0$, $P[A > t] > 0$) and hence

$$\frac{\int_t^{t+u} f_A(x)dx}{P[A>t]}$$

increases monotonically with t ($u \geq 0$).

[Note that this implies

$$f_A(z)\, P[A > u] \leq f_A(u)\, P[A > z] \qquad \text{for every } z \leq u$$

and hence for every $t \leq v$

$$\int_{z=t}^{v} f_A(z)dz \int_v^{\infty} P[A > u]du \leq P[A > v] \int_{z=t}^{v} P[A > z]dz \tag{11}$$

and

$$\int_{z=v}^{\infty} f_A(z)dz \int_v^{\infty} P[A > u]du = P[A > v] \int_{z=v}^{\infty} P[A > z]dz \tag{12}$$

Adding (11) and (12) gives

$$P[A > t] \int_v^{\infty} P[A > u]du \leq P[A > v] \int_{z=t}^{\infty} P[A > z]dz$$

i.e. A has decreasing mean residual life].

With this assumption of increasing failure rate, for every $v \geq t \geq 0$

$$\frac{\int_t^v f_I(u)du}{P[I>t]} = \frac{1}{P[I>t]} \int_{u=t}^{v} \int_o^{\infty} f_A(u+x)\, dX(x)du$$

$$= \frac{1}{P[I>t]} \int_{u=t}^{v} \int_o^{\infty} \frac{f_A(u+x)}{P[A>t+x]} P[A > t + x]\, dX(x)du$$

$$\geq \frac{1}{P[I>t]} \int_{u=t}^{v} \int_{o}^{\infty} \frac{f_A(u)}{P[A>t]} P[A > t + x]\, dX(x)du$$

$$= \frac{\int_t^v f_A(u)du}{P[A>t]}$$

i.e. $$\frac{P[I>t]-P[I>v]}{P[I>t]} \geq \frac{P[A>t]-P[A>v]}{P[A>t]}$$

$$\frac{P[I>v]}{P[I>t]} \leq \frac{P[A>v]}{P[A>t]} \qquad \text{for every } v \geq t \geq 0$$

As A has also decreasing mean residual life, from equation (10) for every $0 \leq v \leq t$

$$\frac{\int_t^\infty P[I>u]du}{\int_t^\infty P[A>u]du} \leq \frac{P[I>t]}{P[A>t]} \leq \frac{P[I>v]}{P[A>v]}$$

i.e. $$P[A > v] \int_t^\infty P[I > u]du \leq P[I > v] \int_t^\infty P[A > u]du$$

Integrating over v

$$\int_o^t P[A > v]dv \int_t^\infty P[I > u]du \leq \int_o^t P[I > v]dv \int_t^\infty P[A > u]du$$

i.e. $$E[A] \int_t^\infty P[I > u]du \leq E[I] \int_t^\infty P[A > u]du$$

and therefore

$$\frac{E[I^2]}{2E[I]} \leq \frac{E[A^2]}{2E[A]} \qquad (13)$$

To apply these inequalities, note that equation (3) gives

$$E[W] = \frac{var[U]-var[(W+U)^-]}{-2E[U]}$$

On using (2) this becomes

$$E[W] = \frac{E[U^2]-E[\{(W+U)^-\}^2]}{-2E[U]}$$

If $(W+U)^- > 0$ then $(W+U)^- = I$. Therefore, if

$$a_o = P[\text{arrival finds system empty}]$$

$$= P[W + U < 0]$$

then

$$E[W] = \frac{E[U^2]-a_o E[I^2]}{-2E[U]}$$

$$= \frac{E[U^2]}{-2E[U]} - \frac{E[I^2]}{2E[I]}$$

$$= \frac{var[S]+var[A]}{2\{E[A]-E[S]\}} + \frac{E[A]-E[S]}{2} - \frac{E[I^2]}{2E[I]}$$

Thus if A has mean residual life bounded above by γ

$$E[W] \geq J + \frac{\{E[A]-E[S]\}}{2} - \gamma$$

If A has decreasing mean residual life, $\gamma = E[A]$ and

$$E[W] \geq J - \frac{E[A]+E[S]}{2}$$

If A has increasing failure rate

$$E[W] \geq J + \frac{E[A]-E[S]}{2} - \frac{E[A^2]}{2E[A]}$$

$$= J - \frac{var[A]+E[S]E[A]}{2\,E[A]}$$

Recall that J denotes the upper bound - see equation (4).

5.2 Inequalities for the GI/G/1 Priority Queue

In this section the superscripts NP and PR will be used to denote the non-preemptive and preemptive resume disciplines. Therefore

$$E[W_2^{PR}] = E[W_2^{NP}]$$

and from Chapter 3, equation (30)

$$\rho\, E[W] = \rho_1\, E[W_1^{NP}] + \rho_2\, E[W_2^{NP}] \qquad (14)$$

where W is the equilibrium waiting time in the pooled queue and W_i^{NP} is the equilibrium waiting time of an i-unit (i=1,2) in a non-preemptive priority queue.

The problem of obtaining approximations for $E[W]$ and $E[W_1^{PR}]$ reduces to that of the last section, the results of which can be used to give simple bounds A, B, A_1, B_1 such that

$$A \leqslant E[W] \leqslant B$$

$$A_1 \leqslant E[W_1^{PR}] \leqslant B_1$$

Clearly $E[W_1^{NP}] \geqslant E[W_1^{PR}] \geqslant A_1$. Equation (14) then gives

$$E[W_2^{NP}] \leqslant \frac{\rho\, E[W] - \rho_1\, A_1}{\rho_2} \leqslant \frac{\rho\, B - \rho_1\, A_1}{\rho_2}$$

To obtain a lower bound for $E[W_2^{NP}]$ (and hence, by equation (14) an upper bound for $E[W_1^{NP}]$ note that

$$W_2 = T_0 + T_1 + T_2 + \ldots,$$

where

T_0 = time to clear system of all units which arrived before the 2-unit.

T_i = time to clear system of all 1-units which arrived during T_{i-1} $(i \geq 1)$.

Therefore

$$E[W_2] = E[W] + \sum_{i=1}^{\infty} E[T_i]$$

$$= E[W] + \frac{1}{\mu_1} \sum_{i=1}^{\infty} E[N_i]$$

where $1/\mu_1 = E[S_1]$ = mean service time of a 1-unit, and

N_i = number of 1-units arriving during T_{i-1}.

Therefore, using the result on page 53 of [3],

$$E[N_1] \geq \lambda_1 E[W] - 1$$

$$E[N_2] \geq \lambda_1 E[T_1] - 1 \qquad \text{etc.}$$

Thus

$$E[W_2^{NP}] \geq \text{Max}\{E[W], E[W](1+\rho_1) - \frac{1}{\mu_1},$$

$$E[W](1+\rho_1+\rho_1^2) - \frac{1}{\mu_1}\rho_1 - \frac{2}{\mu_1}, \ldots\}$$

giving a lower bound for $E[W_2^{NP}]$.

Note that the second term is better (i.e. greater) than the first if

$$\rho_1 E[W] \geq 1/\mu$$

i.e. $$\lambda_1 E[W] \geq 1$$

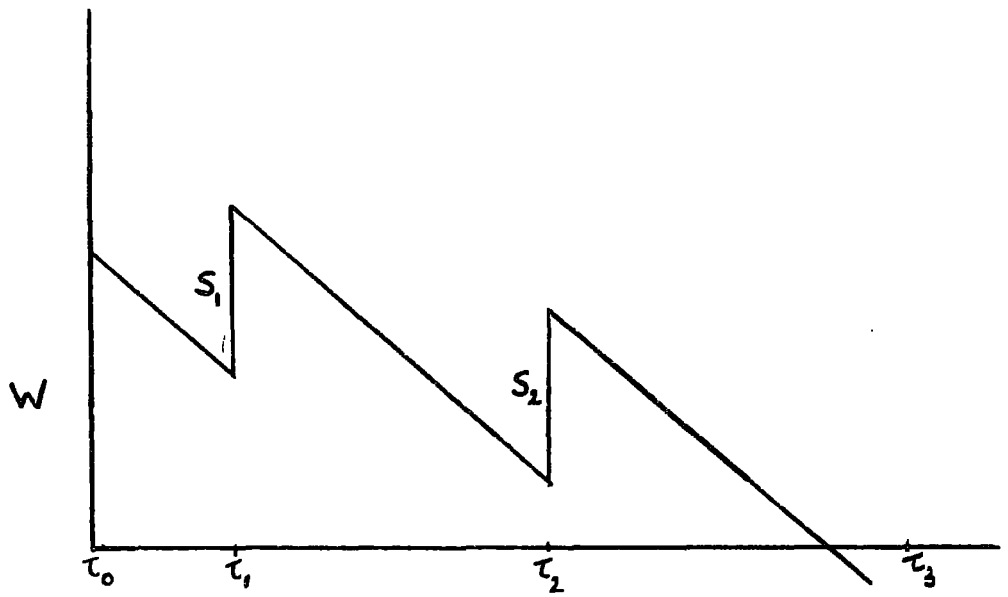
and the third term is better than the second if

$$\lambda_1^2 E[W] \geq \mu_1 + \lambda_1 \qquad \text{etc.}$$

Clearly this method is very crude and better methods are needed: if, on

inspection, the distribution of the interarrival time between two 1-units can be bounded by a distribution for which results are known [i.e. E_k or M] or alternatively if the service time distribution function can be bounded by exponential distribution functions, then the following results can be used:

W_2, the waiting time of a 2-unit in the stationary state, has the same distribution as the first passage time to 0 in the following stochastic process, V(t)



$$V(t) = W - t + \Sigma S_i$$

where the summation is taken over all i such that $\tau_i \leq t$ and W is the waiting time in the pooled queue

$S_1, S_2, \ldots$, are the service times of 1-units.

$A_n = \tau_n - \tau_{n-1}$, $n \geq 1$ are interarrival times of 1-units and all these random variables are independent.

Compare two such processes: $V_1(t)$ in which the A_n have distribution function $F_1(t) = P[A \leq t]$, density $f_1(t)$ and $V_2(t)$ in which the A_n have

distribution function $F_2(t) = P[A \leq t]$, density $f_2(t)$.

Suppose $F_1(t) \geq F_2(t)$, for all $t \geq 0$.

Then if for $i = 1,2$,

$\bar{F}_i(t) = 1 - F_i(t)$ and $F_i^{(n)}(t)$ denotes the distribution function of the n-fold convolution of $F_i(t)$

$$\bar{F}_1^{(2)}(t) = 1 - F_1^{(2)}(t)$$

$$= P[A_1 + A_2 > t \text{ in the first process}]$$

$$= \int_0^\infty P[A_1 > t - u]\, f_1(u)\, du$$

$$= \int_0^\infty \bar{F}_1(t-u)\, f_1(u)\, du$$

$$\leq \int_0^\infty \bar{F}_2(t-u)\, f_1(u)\, du$$

$$= \int_0^\infty \bar{F}_1(t-u)\, f_2(u)\, du$$

$$\leq \int_0^\infty \bar{F}_2(t-u)\, f_2(u)\, du$$

$$= \bar{F}_2^{(2)}(t)$$

and by induction

$$\bar{F}_1^{(n)}(t) \leq \bar{F}_2^{(n)}(t) \qquad \text{for every } n \geq 1,\ t \geq 0$$

If $N_i(t)$ denotes the number of arrivals in $[0,t]$ in the $V_i(t)$ process (i=1,2) then

$$P[N_1(t) < n] = P[A_1 + A_2 + \ldots + A_n > t \text{ in the first process}]$$

$$= \bar{F}_1^{(n)}(t)$$

$$\leq \bar{F}_2^{(n)}(t)$$

$$= P[N_2(t) < n]$$

i.e. $P[N_1(t) \geq n] \geq P[N_2(t) \geq n]$ for every $n \geq 0$, $t \geq 0$

Now

$$V_i(t) = W - t + X_i(t) \qquad i = 1,2;\ t \geq 0$$

where $X_i(t) = \Sigma S_j$, the summation being taken over all j such that $\tau_j \leq t$ in the i^{th} process. Therefore

$$P[V_1(t) > 0] = P[X_1(t) > t - W]$$

$$= \sum_{k=0}^{\infty} P[S_1 + \ldots + S_k > t - W]\ P[N_1(t) = k]$$

$$= \sum_{k=0}^{\infty} \mathbb{P}_k\ a_k^{\,1}$$

where $a_k^{\,1} = P[N_1(t) = k] \qquad k \geq 0$

and $\mathbb{P}_k = P[S_1 + \ldots + S_k > t - W]$

$$0 \leq p_0 \leq p_1 \leq \ldots \leq 1$$

Thus

$$P[V_1(t) > 0] - P[V_2(t) > 0] = \sum_{k=0}^{\infty} p_k(a_k^{\,1} - a_k^{\,2})$$

$$= \sum_{k=0}^{\infty} p_k\ a_k$$

which is absolutely convergent.

Now

$$\sum_{k=n}^{\infty} a_k = P[N_1(t) \geq n] - P[N_2(t) \geq n] \geq 0 \qquad \text{for all } n \geq 0$$

Also

$$\sum_{k=0}^{\infty} |a_k| \quad \text{converges} \quad (\text{and is} \leqslant 2)$$

Hence, there exists an m such that

$$\sum_{k=m}^{\infty} |a_k| < \varepsilon$$

$$\sum_{k=0}^{\infty} p_k a_k = \sum_{k=1}^{\infty} (p_k - p_o) a_k + \sum_{k=0}^{\infty} p_o a_k$$

$$\geqslant \sum_{k=1}^{\infty} (p_k - p_o) a_k$$

$$\text{as} \quad \sum_{k=0}^{\infty} a_k \geqslant 0 \ , \quad p_o \geqslant 0$$

$$= \sum_{k=1}^{\infty} p_k^{\,1} a_k \ , \qquad p_k^{\,1} = p_k - p_o$$

$$0 \leqslant p_1^{\,1} \leqslant p_2^{\,1} \leqslant \ldots \leqslant 1$$

Similarly,

$$\sum_{k=1}^{\infty} p_k^{\,1} a_k = \sum_{k=2}^{\infty} (p_k^{\,1} - p_1^{\,1}) a_k + \sum_{k=1}^{\infty} p_1^{\,1} a_k$$

$$\geqslant \sum_{k=2}^{\infty} p_k^{\,2} a_k \qquad 0 \leqslant p_2^{\,2} \leqslant p_3^{\,2} \leqslant \ldots \leqslant 1$$

and so on, giving

$$\sum_{k=0}^{\infty} p_k a_k \geqslant \sum_{k=m}^{\infty} p_k^{\,m} a_k \qquad 0 \leqslant p_m^{\,m} \leqslant p_{m+1}^{\,m} \leqslant \ldots \leqslant 1$$

But

$$\left| \sum_{k=m}^{\infty} p_k^{\,m} a_k \right| \leqslant \sum_{k=m}^{\infty} |a_k| < \varepsilon$$

i.e. $$\sum_{k=m}^{\infty} p_k^{\,m} a_k > -\varepsilon \qquad \text{for every } \varepsilon > 0$$

i.e. $\sum_{k=0}^{\infty} p_k a_k > -\varepsilon$ for every $\varepsilon > 0$

i.e. $\sum_{k=0}^{\infty} p_k a_k \geq 0$

and therefore

$$P[V_1(t) > 0] \geq P[V_2(t) > 0] \quad \text{for every } t \geq 0$$

It follows that, for every $t \geq 0$,

$$P[V_1(u) > 0,\ 0 \leq u \leq t] \geq P[V_2(u) > 0,\ 0 \leq u \leq t]$$

i.e. $P[T_1 > t] \geq P[T_2 > t]$ for every $t \geq 0$

where T_i denotes the first passage time to zero in the i^{th} process.

Integrating gives $E[T_1] \geq E[T_2]$

This result shows that for two arrival processes with distribution functions satisfying

$$F_1(t) \geq F_2(t) \quad \text{for every } t \geq 0$$

then the mean waiting time of a 2-unit in the first process is not less than the mean waiting time of a 2-unit in the second process.

Example. Suppose the age specific failure rate $\phi(t)$ of the distribution function of the interarrival time between two 1-units satisfies

$$1/\alpha \leq \phi(t) \leq 1/\beta \quad \text{for all } t$$

then $\exp(-t/\beta) \leq \bar{F}(t) \leq \exp(-t/\alpha)$

and using the above results gives

$$\frac{E[W]}{1-(1/\mu_1\alpha)} \leq E[W_2^{NP}] \leq \frac{E[W]}{1-(1/\mu_1\beta)}$$

$E[W]$ referring to the pooled queue.

Similarly for two processes with service time distribution functions satisfying

$$S_1(t) \geq S_2(t) \qquad \text{for every } t \geq 0$$

$$p_k^1 = P[S_1 + \ldots + S_k > t - W \text{ for the first process}]$$

$$\leq p_k^2$$

and hence

$$P[V_1(t) > 0] - P[V_2(t) > 0] \leq 0$$

and the mean waiting time of a 2-unit in the first process is not greater than the mean waiting time of a 2-unit in the second process.

Unfortunately the above method requires knowledge of the whole distribution of either the interarrival times or the service times of 1-units and so particular values obtained may be very sensitive to small changes in the distributions. To obtain more robust inequalities it seems necessary to place restrictions on the distribution of the service time of a 2-unit: suppose S_2 has mean residual life bounded above by α. Then, for a 2-class non-preemptive queue:

$$W_{n+1} = \text{Waiting Time of } (n+1)^n \text{ 1-unit}$$

$$= W_n + S_n - A_n + X_n \qquad (15)$$

where

S_n = service time of n^{th} 1-unit, mean $1/\mu_1$

A_n = interarrival time between n^{th} and $(n+1)^{th}$ 1-units, mean $1/\lambda_1$

$$X_n = \begin{cases} 0 & \text{if } W_n + S_n - A_n \geqslant 0 \\ I_n + R_n & \text{if } W_n + S_n - A_n = -I_n < 0 \end{cases}$$

I_n = time from departure of n^{th} 1-unit to arrival of $(n+1)^{th}$ 1-unit

R_n = time $(n+1)^{th}$ 1-unit must wait before commencing service due to 2-unit in service.

Therefore

$$W_{n+1}^2 = W_n^2 + (S_n - A_n)^2 + X_n^2 + 2W_n(S_n - A_n + X_n) + 2X_n(S_n - A_n)$$

$$= W_n^2 + (S_n - A_n)^2 + X_n[X_n + 2S_n - 2A_n + 2W_n] + 2W_n[S_n - A_n]$$

Taking expected values and assuming stationarity,

$$2E[W_1^{NP}]E[A - S] = E[(S-A)^2] + \rho E[(I+R)(R-I)]$$

where $\rho = P[W_1^{NP} + S - A < 0]$

$= P[1\text{-unit finds server idle}]$

i.e.

$$E[W_1^{NP}] = \frac{E[(S-A)^2] + \rho E[R^2] - \rho E[I^2]}{2(1/\lambda_1 - 1/\mu_1)} \qquad (16)$$

From equation (15)

$$E[X] = \rho E[I + R] = E[A] - E[S] = 1/\lambda_1 - 1/\mu_1$$

i.e. $\rho E[I] = 1/\lambda_1 - 1/\mu_1 - \rho E[R]$

Now S_2 has mean residual life bounded above by α, and therefore $E[R] \leq \alpha$

i.e. $$p\,E[I] \geq 1/\lambda_1 - 1/\mu_1 - \alpha$$

Also

$$p\,E[I^2] \geq p\{E[I]\}^2$$

$$\geq \{p\,E[I]\}^2$$

$$\geq b$$

where $$b = \begin{cases} 0 & \text{if } 1/\lambda_1 - \frac{1}{\mu_1} < \alpha \\ 1/\lambda_1 - 1/\mu_1 - \alpha & \text{if } 1/\lambda_1 - \frac{1}{\mu_1} \geq \alpha \end{cases}$$

Substituting in (16) gives

$$E[W_1^{NP}] \leq \frac{E[(S-A)^2]+E[R^2]-b}{2(1/\lambda_1-1/\mu_1)}$$

Finally $R = S_2 - B$ where B is a strictly positive random variable and therefore, as for equation (9) above

$$E[R^2] \leq 2\alpha E(R)$$

giving the final inequality

$$E[W_1^{NP}] \leq \frac{E[(S-A)^2]+2\alpha/\mu_2-b}{2(1/\lambda_1-1/\mu_1)}$$

Van Mildert College, Durham September, 1971

REFERENCES

[1] AVI-ITZHAK, B., I. BROSH and P. NAOR On Discretionary Priority Queueing. Zeit. angew. Math. Mech. (1964) 44, 235-242.

[2] BALACHANDRAN, K.R. Parametric Priority Queues: an Approach to Optimization in Priority Queues. Operat. Res. (1970) 18, 526-540.

[3] BARLOW, R.E. and F. PROSCHAN Mathematical Theory of Reliability. J. Wiley (1965).

[4] COX, D.R. Renewal Theory. Methuen (1962).

[5] COX, D.R. and W.L. SMITH Queues. Methuen (1961).

[6] DOWNTON, F. Bivariate Exponential Distributions in Reliability Theory. J.R. Statist. Soc. B (1970) 32, 408-417.

[7] GAVER, D.P. A Waiting Line with Interrupted Service including Priorities. J.R. Statist. Soc. B (1962) 24, 73-90.

[8] JAISWAL, N.K. Priority Queues. Academic Press (1968).

[9] JAISWAL, N.K. and K. THIRUVENGADAM Preemptive Resume Priority Queue with Erlangian Inputs. Indian J. Math. (1962) 4, 53-70.

[10] KINGMAN, J.F.C. The Heavy Traffic Approximation in the Theory of Queues. Proceedings of the Symposium on Congestion Theory, W.L. Smith and W.E. Wilkinson (eds.) University of North Carolina Press (1965).

[11] ———— Inequalities in the Theory of Queues. J.R. Statist. Soc. B (1970) 32, 102-110.

[12] MARSHALL, K.T. Some Inequalities in Queueing. Operat. Res. (1968) 16, 651-668.

[13] ———— Bounds for Some Generalisations of the GI/G/1 Queue. Operat. Res. (1968) 16, 841-848.

[14] MILLER, R.G. Priority Queues. Ann. Math. Statist. (1960) 31, 86-103.

[15] MIRASOL, N.M. The Output of an M/G/∞ Queueing System is Poisson. Operat. Res. (1963) 11, 282-284.

[16] SCHRAGE, L.E. and L.W. MILLER The Queue M/G/1 with the Shortest Remaining Processing Time Discipline. Operat. Res. (1966) 14, 670-684.

[17] TAKACS, L. Introduction to the Theory of Queues. Oxford University Press (1962).

[18] ———— Priority Queues. Operat. Res. (1964) 12, 63-74.

[19] ———— Combinatorial Methods in the Theory of Stochastic Processes. J. Wiley (1966).

[20] ———— On the Distribution of the Supremum for Stochastic Processes. Ann. Inst. H. Poincaré Sect. B (1970) 6, 237-247.

[21] ———— On the Distribution of the Maximum of Sums of Mutually Independent and Identically Distributed Random Variables. Advances in Appl. Probability (1970) 2, 344-354.

[22] VERE-JONES, D. Some Applications of Probability Generating Functionals to the Study of Input-Output Streams. J.R. Statist. Soc. B (1968) 30, 321-333.