CONTROLLING SINGLE SERVER QUEUES

by David Edward Matthews

Thesis submitted for the degree of Doctor of Philosophy in the University of London by a candidate registered at Imperial College of Science and Technology

ABSTRACT

Controlling or reducing congestion by changing basic features of a queueing process is an important aspect of applied queueing theory. Each change should alter statistical properties of the process to benefit either customers or servers.

If customers with shorter service times are served first, or if service is faster when the queue is long, the queueing times of most customers will be reduced. If longer idle periods are created by closing the service counter, servers are free to do ancillary work. These three changes in a queueing process are considered, individually, as ways of controlling congestion in a single server queueing system with Poisson arrivals.

A simple queue discipline with only two non-preemptive priority classes is shown to be an effective method of reducing queueing times if prior information about service times is available.

Faster service when the queue is long is the aim of hysteresis control. An equilibrium solution is obtained for a generalized model of hysteresis control with k pairs of control levels and arbitrary service time distributions. A special case, unilevel control, is shown to act automatically to prevent long queues from forming.

How long a service counter should remain closed for ancillary work can be decided by referring to the number of waiting customers or to the virtual queueing time. In each case the inconvenience to customers of shutting down the server is determined by deriving the equilibrium queueing time distribution.

A series of numerical studies explores the practical effects of these suggested methods of controlling congestion.

THANKS BE TO GOD

ACKNOWLEDGMENTS

I would like to thank Professor D.R. Cox for his patient supervision and timely encouragement during the course of this work. I am also indebted to both Dr. Valerie Isham and Dr. Mark Westcott for carefully reading an earlier draft; their pertinent comments were appreciated. Dr. Agnes Herzberg also deserves my gratitude for her assistance on many occasions.

The financial support of the Commonwealth Scholarship Commission in the United Kingdom is gratefully acknowledged. I would also like to thank the members of the Commonwealth Education and Awards Department of the British Council for all their efforts on my behalf.

TABLE OF CONTENTS

ABSTRACT 2

ACKNOWLEDGMENTS 3

LIST OF TABLES AND FIGURES 6

CHAPTER 1. Introduction

1.1 A general model for a queueing process 8

1.2 Measures of congestion 9

1.3 Some general aspects of the problem of control 11

CHAPTER 2. Choosing a queue discipline to control congestion

2.1 Simple control techniques — two-class non-preemptive priority disciplines 13

2.2 Evaluating some effects of a change in queue discipline 18

2.3 More complicated control techniques — k-class non-preemptive priority disciplines 29

2.4 Idealized control techniques — a different priority class for each customer 32

CHAPTER 3. Controlling congestion behind the counter

3.1 Linking shut-down control of the service process to queue size 41

3.2 Linking shut-down control of the service process to virtual queueing time 48

3.3 The effect of shut-down control on a queueing process 52

3.3.1 The effect of (0,N) control on queueing time 54

3.3.2 The effect of (0,V) control on queueing time 59

3.3.3 Choosing between (0,N) and (0,V) control 64

CHAPTER 4. Adaptive control of the service process

4.1 Linking the service process to the line size 67

4.2 Unilevel control of the service process 70

4.3 The effect of unilevel control on the distribution of L 76

4.4 Bilevel hysteresis control of the service process 98

CHAPTER 5. Generalized hysteresis control of the service process

5.1 2k-level hysteresis control 107

CHAPTER 6. Concluding remarks

6.1 An alternative to optimal control 114

6.2 Some outstanding problems 116

REFERENCES 119

LIST OF TABLES AND FIGURES

Figure 2.1.1	Service times which optimally divide customers into two priority classes	16
Table 2.1.1	The ratio of mean queueing times for service in order of arrival and a sub-optimal division of customers into two priority classes	18
Table 2.2.1	Skewness coefficients for theoretical and fitted conditional queueing time distributions	23
Figure 2.2.1	The effect of a change in queue discipline on the distribution of positive queueing times	25
Figure 2.2.2	The effect of a change in queue discipline on the distribution of positive queueing times	26
Figure 2.2.3	The effect of a change in queue discipline on the distribution of positive queueing times	27
Table 2.2.2	Relative accuracies of gamma approximations to theoretical queueing time distributions	29
Table 2.4.1	Mean queueing times for five different queue disciplines	37
Table 2.4.2	Queueing time variances for five different queue disciplines	38
Table 3.3.1	Probabilities of standardized queueing time events for (0,N) control	55
Figure 3.3.1	The effect of (0,N) control on queueing time	56-7
Table 3.3.2	Probabilities of standardized queueing time events for (0,V) control	60

Figure 3.3.2	The effect of (0,V) control on queueing time	61-2
Table 3.3.3	Comparing alternative versions of (0,N) and (0,V) control	65
Figure 4.1.1	A bilevel hysteresis control pattern	69
Table 4.3.1	Probabilities of standardized events in different unilevel control line size processes	80-3
Figure 4.3.1	Probabilities of standardized events in different unilevel control line size processes as a function of the control threshold	84-5
Figure 4.3.2	Probabilities of standardized events in different unilevel control line size processes as a function of the control threshold	87-8
Table 4.3.2	Probabilities of standardized events in corresponding line size processes without unilevel control	92-4
Figure 4.3.3	The effects of unilevel control on the distribution of line size	95-6
Table 4.4.1	Equilibrium marginal probability distributions for line length in six different hysteresis control queues	105
Figure 5.1.1	A generalized hysteresis control pattern	107

CHAPTER 1. Introduction

1.1 A general model for a queueing process

Models for queueing processes, whether simple or elaborate, generally describe the interaction of an arrival process and a pattern of service. Customers join the system, are selected for service from the pool of waiting customers, and leave after being served. Usually, an adequate model of a particular queueing situation can be specified by identifying the essential details of the arrival process, the selection pattern or queue discipline, and the service process.

Two features generally suffice to describe the arrival process; these are the size of the customer population and the joint probability distribution of the intervals between arrivals. Obviously, many different arrival patterns are possible. In subsequent chapters attention concentrates on those situations for which the arrival pattern can be adequately represented by a stationary Poisson process with rate $\lambda > 0$.

Arriving customers form one or more queues. Sometimes the possible size of a queue is limited; this particular case is usually called a limited waiting room model. In Chapters 2-5 the implicit assumption is made that arrivals form a single queue of unrestricted length.

The queue discipline is a rule by which customers are selected for service. Many different rules are possible. For example, customers could be served in order of arrival, or could be assigned service priorities. Military communications traffic is an excellent example of a queueing system with a queue discipline involving priority classes.

Two features generally determine the service process. These are the number of servers and the joint probability distribution of customer service times. In the single server situations which we consider, customer service times are independent realizations of a non-negative random variable with distribution function of arbitrary form. According to the well-established classification system which Kendall (1953) introduced, the results of succeeding chapters apply to queueing situations which can be represented by the familiar model M/G/1.

The following definitions should eliminate any confusion which might arise because terminology in queueing theory has not been standardized.

Definitions

A customer's queueing time, W_q, is the time between his arrival in the queue and the start of his service.

A customer's waiting time, W, is the time between his arrival in the queue and his departure from the system.

The queue length, L_q, is the number of customers queueing for service.

The line length, L, is the number of customers in the queueing system.

It follows that waiting time equals queueing time plus service time, and line length equals the number of customers being served plus the queue length.

1.2 Measures of congestion

Models of queueing processes not only describe but also quantify the amount of congestion in a queueing situation in terms of several different properties. Perhaps the simplest measure of congestion is the traffic intensity, ρ; this is generally defined as the ratio of mean service time to mean inter-arrival time. Usually, if ρ exceeds unity, the system will be very congested. Conversely, if $\rho<1$, most systems will reach a state of statistical equilibrium. If customers arrive in a stationary Poisson process and $\rho<1$, it is well-known that, with probability $1-\rho$, a given customer will not have to queue for service. To know the probability of this event in a practical situation is often quite important.

Often, the mean queue length is used to measure congestion. However, specific knowledge of the probability distribution of L_q can be useful, particularly when the size of the waiting room must be restricted. For example, with randomly arriving customers, the long-run proportion of customers who are turned away because the system is full can be evaluated. Occasionally, either the probability distribution of L or its mean value may be easier to determine. Since L equals L_q plus the number of customers being served, it is usually a simple matter to derive the statistical properties of one quantity from those of the other.

Mean queueing time and the probability distribution of W_q are important properties of the system in relation to the amount of congestion. This is particularly true whenever customer delays represent economic losses. Provided the loss per unit delay per customer is constant, Cox & Smith(1961, p.26) state that only mean queueing times need be considered. In other situations it would probably be helpful to know the queueing time distribution as well as its mean. For example, if standards of service are defined in terms of the long-run proportion of customers who queue for more than a fixed time, tail probabilities for the distribution of W_q will need to be evaluated.

Sometimes other properties of a queueing system may best characterize the important aspects of congestion. For example, if serving costs are particularly high and idle time represents costly economic losses, it would be useful to know the distribution of the length of the busy period. In this case, congestion behind the service counter may be more important than delays to customers.

In general, then, and whenever possible, congestion should be measured in terms of quantities which have an obvious physical or economic significance.

1.3 Some general aspects of the problem of control

Situations in which the level of congestion is likely to exceed tolerable limits give rise to the problem of controlling a queueing process. Theoretically, by modifying one or more basic features of the system, i.e. the service process, queue discipline, etc., reductions in congestion can be obtained. Evaluating the effect of proposed modifications on the level of congestion in the system is an important aspect of the problem.

Clearly, the level of congestion can be decreased by restricting or interrupting the arrival process. Sometimes this filtering effect can be achieved by taxing customers who decide to join the queue. The success of taxation as a method for controlling congestion depends crucially on the assumption that customers can be selectively discouraged from joining the queue by increasing the tax. Naor(1969) considers the use of a fixed tax in order to filter the arrival process of an M/M/1 queue. Adler & Naor(1969) examine a similar problem for the case of an M/D/1 queue. By assuming a linear structure of customer rewards and operating costs the same authors show that optimal joining decisions by individual customers do not necessarily determine a social optimum for the customer population. More recently, Yechiali(1971) has analyzed the problem of determining individual balking rules and social toll charges for a GI/M/1 queueing process. When customer rewards and queueing costs per unit time are linear, Yechiali is able to use Markov decision process methods to determine the form of control rules which maximize either the individual or population, infinite-horizon, average reward.

Instead of taxing customers who join the queue in order to reduce congestion it may be simpler to limit the size of the queue. Customers who arrive when the system is full are turned away. This method has obvious applications in telephone engineering and related fields. When the arrival process is Poisson, the long-run proportion of blocked customers can be evaluated using the Erlang loss formula [cf. Saaty(1961, p.303)].

Arguments which lead to the Pollaczek-Khintchine formula [cf. Cox & Smith(1961, p.55)] show that mean queueing time in the M/G/1 queue does not depend on the queue discipline if customers are indistinguishable from the point of view of service time. However, if customer delays are measured in relation to the queueing time distribution, then the choice of a queue discipline will be an important one. The number of possible choices in any situation may be considerable. In Chapter 2 we examine in greater detail queue discipline choices which can help to reduce congestion. In particular, the use of available information to minimize individual and overall customer delays will be emphasized.

Sometimes it may be more important to reorganize server idle time than to reduce customer delays. This situation has been mentioned already in §1.2. In Chapter 3 we consider detailed results regarding two methods of modifying the service process in order to restructure the server's busy and idle periods.

Changes in the service process are frequently suggested whenever the primary aim of any control method is to reduce customer delays. Such changes might include an increase in the number of servers or a change in the service time distribution. In Chapter 4 we obtain equilibrium solutions for two related control methods which monitor the level of congestion in a system and regulate the service process accordingly. In this respect, each method is analogous to modern industrial feedback control.

The theoretical results of Chapter 4 point to a generalized model for adaptive control of the service process. In Chapter 5 we derive an equilibrium solution in this wider frame of reference and identify the results of §§4.2 and 4.4 as two important special cases of the general problem.

CHAPTER 2. Choosing a queue discipline to control congestion

2.1 Simple control techniques — two-class non-preemptive priority disciplines

For a fixed pattern of arrivals and customer service times the queue discipline determines how long customers are delayed. Kingman(1962) proves that, for the class of queue disciplines which do not affect the distribution of the number in the queue at any time, the mean of the queueing time distribution is independent of the discipline, but the variance is minimized by serving customers in order of arrival. If minimum variance for the queueing time distribution determines the preferred queue discipline, service in order of arrival would be the natural choice provided the alternative disciplines satisfy the above conditions.

However, not all queue disciplines satisfy the conditions which Kingman specifies. Among those which do not are service patterns with mean queueing times which are less than the value specified by the Pollaczek-Khintchine formula; for the M/G/1 queue it is this value to which Kingman's result refers. Schrage & Miller(1966) state that when the customer with the shortest remaining processing time is given a preemptive resume priority for service the line length at any time is minimized. However, it would often be impossible to follow this rule. Instead, what is needed is a practical queue discipline which can use available information to reduce congestion.

Perhaps the simplest of all such rules is a two-class non-preemptive priority discipline requiring some prior knowledge of customers' service times. The Pollaczek-Khintchine formula specifies that if g_j is the jth moment about the origin of the service time distribution ($j=1,2$), then $E(W_q)$ is equal to $\frac{1}{2}\lambda g_2/(1-\rho)$, where $\lambda g_1=\rho<1$. However, if customers are classified as "long" and "short" according to their future service times, and if the "short" class is given non-preemptive priority, then the mean queueing time can be much less than $\frac{1}{2}\lambda g_2/(1-\rho)$. Within each priority class, customers are served in order of arrival.

The credit for pioneering work on non-preemptive priority queue disciplines belongs to Cobham(1954,1955). Using Cobham's results, Phipps(1956) shows that a considerable reduction in $E(W_q)$ can be obtained by giving non-preemptive priority to the waiting customer with the shortest future service time. Some implications of Phipps's shortest service time rule will be considered in §2.4.

Let $G(\cdot)$ be an arbitrary service time distribution with corresponding density function $g(\cdot)$ and jth moment $g_j=\int_0^\infty t^j g(t)dt$, $(j=1,2,\ldots)$. Schrage & Miller(1966) show that when customers with service times not exceeding ϕ are assigned to class 1 and all others are relegated to class 2, the mean queueing time, $E(W_q|\phi)$, is given by

$$E(W_q|\phi)=\frac{1}{2}\lambda g_2\frac{1-\rho G(\phi)}{(1-\rho)\{1-\rho(\phi)\}}\quad, \tag{2.1.1}$$

where $\rho(\phi)=\lambda\int_0^\phi tg(t)dt\leq\rho<1$ for all $\phi>0$. This queue discipline is obviously simple to administer and only requires a moderate amount of prior information concerning customers' service times. However, Schrage & Miller(1966) only briefly discuss the possibility that this discipline could be an effective, practical way of reducing congestion; many authors do not consider this same rule at all.

We can show that $E(W_q|\phi)$ is smaller than $\frac{1}{2}\lambda g_2/(1-\rho)$ for any finite, positive ϕ. If service times can be accurately estimated or are known in advance, it seems sensible to select ϕ to minimize $E(W_q|\phi)$. Provided $g(x)>0$, (2.1.1) is minimized when $\phi=\phi^*$, where ϕ^* satisfies the equation

$$\phi^*=g_1+\rho\int_0^{\phi^*}G(t)\,dt\quad. \tag{2.1.2}$$

Then, if all customers are correctly classified, the ratio $E(W_q)/E(W_q|\phi=\phi^*)$ attains a maximum value of ϕ^*/g_1. The following examples show the solution of (2.1.2) for several common service time distributions.

Example 2.1.1

Let $g(x)=\frac{1}{2}b^{-1}$ $(0 \leq x \leq 2b)$, where $\lambda b<1$. Then ϕ^* is a solution of the quadratic equation $\rho\phi^{*2}-4b\phi^*+4b^2=0$, and the optimum separation point is $(2-2\sqrt{1-\rho})/\lambda$.

Example 2.1.2

When $g(x)=\frac{k\mu(k\mu x)^{k-1}}{\Gamma(k)}e^{-k\mu x}$ $(k=1,2,\ldots)$ ϕ^* must satisfy the equation $\phi^*=\frac{1}{\mu(1-\rho)}\left\{1-\rho+\frac{\rho}{k}e^{-k\mu\phi^*}\sum_{r=0}^{k-1}\sum_{j=0}^{r}\frac{(k\mu\phi^*)^j}{j!}\right\}$. If $k=2$, we obtain the simple relation

$$\phi^*=\frac{1-\rho+\rho e^{-2\mu\phi^*}}{\mu\left(1-\rho-\rho e^{-2\mu\phi^*}\right)}\ . \tag{2.1.3}$$

When $k=2$ and $\mu=\frac{1}{2}$, i.e. $g(x)=xe^{-x}$, we will call this service time distribution D_1.

Example 2.1.3

Let $g(x)=\mu e^{-\mu x}$. Then

$$\phi^*=\left(1+\frac{\rho}{1-\rho}e^{-\mu\phi^*}\right)\frac{1}{\mu}\ . \tag{2.1.4}$$

When $\mu=\frac{1}{2}$ we will call this particular service time distribution D_2.

Example 2.1.4

Let $g(x)=\theta\mu e^{-\mu x}+(1-\theta)k\mu e^{-k\mu x}$ $(0<\theta<1;\ k>0)$. Then (2.1.2) becomes

$$\phi^*=\frac{1}{k\mu}\left[k\theta+(1-\theta)+\frac{\rho}{1-\rho}\left\{k\theta e^{-\mu\phi^*}+(1-\theta)e^{-k\mu\phi^*}\right\}\right]\ . \tag{2.1.5}$$

By specifying values for θ, k and μ, we can obtain a distribution with any desired mean and a range of coefficients of variation greater than unity. Thus, service times from this mixed exponential distribution are more variable than service times from distributions in any of the preceding examples. When $k=1/3$, $\theta=\frac{1}{2}$ and $\mu=1$, i.e. $g(x)=\frac{1}{2}(e^{-x}+1/3e^{-x/3})$, we will call this service time distribution D_3.

By specifying the parameters in the preceding examples we can solve (2.1.3), (2.1.4) and (2.1.5) numerically for any values of ρ between 0 and 1. Particular solutions to (2.1.3), (2.1.4) and (2.1.5) for the

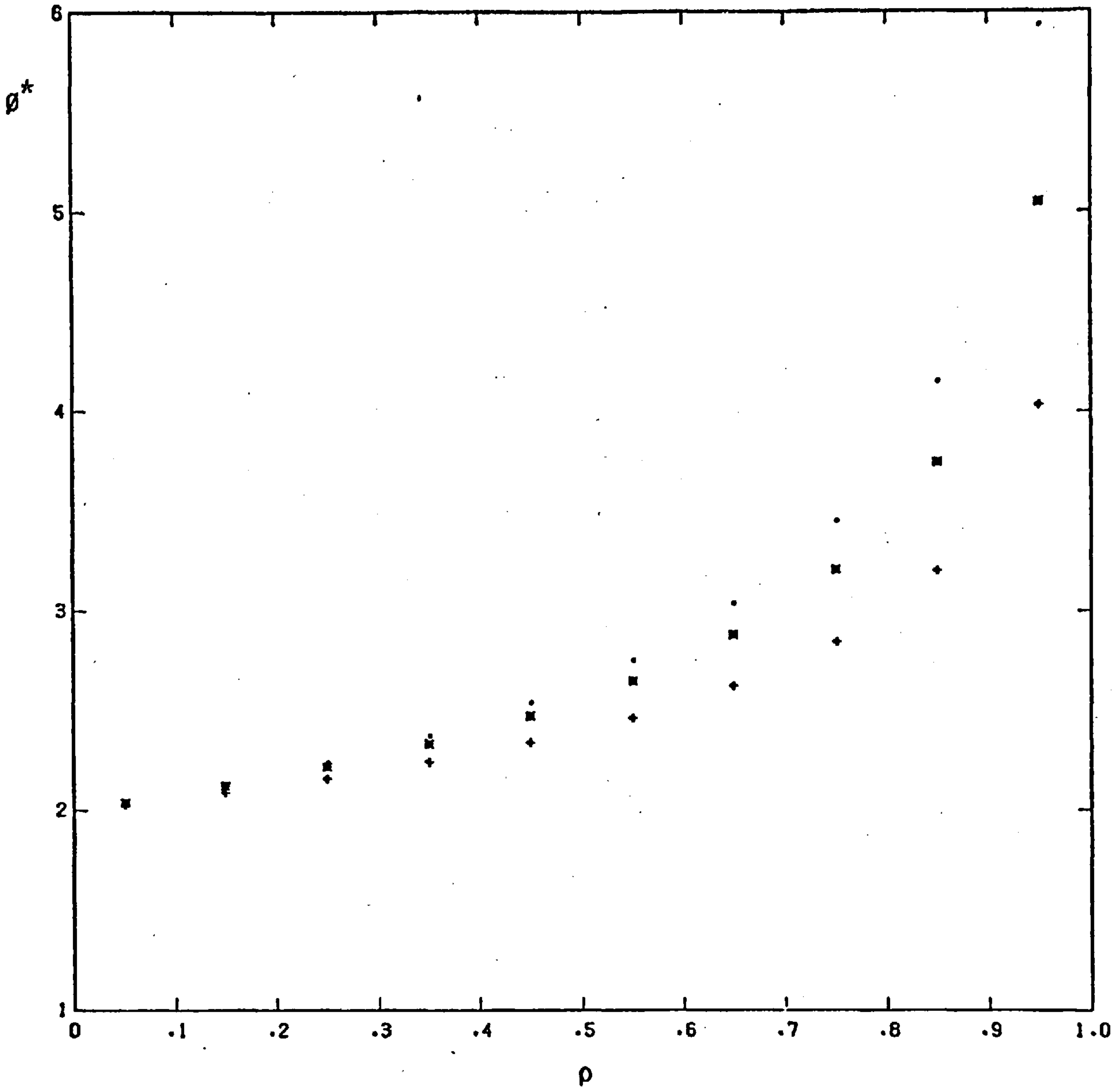


Figure 2.1.1 Service times, ϕ^*, which optimally divide customers into two priority classes when service times are D_1, D_2 or D_3 with traffic intensity ρ.

+ D_1 (Erlang)

■ D_2 (exponential)

• D_3 (mixed exponential)

distributions D_1, D_2 and D_3 have been obtained; the values of $\emptyset^*$ are plotted in Fig. 2.1.1.

Let $G_1(\cdot)$ and $G_2(\cdot)$ be two service time distributions with the same mean, and suppose that $G_1(t) \leqq G_2(t)$ for all $t \geqq t_o$. Then, for any $\rho < 1$ we can show that $\emptyset_1^* \geqq \emptyset_2^*$, where $\emptyset_j^*$ is the solution to (2.1.2) when $G(\cdot) \equiv G_j(\cdot)$ $(j=1,2)$. For large enough service times, the distribution function for D_2 is bounded above by the distribution function for D_1 and bounded below by the distribution function for D_3. Therefore, as Fig 2.1.1 indicates, the values of $\emptyset^*$ for D_1 and D_3 are always the smallest and the largest, respectively. This ordering of the solutions to (2.1.2) by increasing magnitude corresponds to an ordering of the D_i's by increasing variance. It is not obvious, however, that uniformly greater solutions to (2.1.2) will always be obtained for any other service time distribution which has the same mean as the D_i's but which is overdispersed with respect to D_3.

It was stated previously that $\emptyset^*/g_1$ is equal to the ratio $E(W_q)/E(W_q|\emptyset=\emptyset^*)$, where the mean value in the numerator is given by the Pollaczek-Khintchine formula. Figure 2.1.1 indicates that when ρ exceeds 0.8, changing from service in order of arrival to this simple priority discipline could reduce the mean queueing time by at least 1/3. If $\emptyset^*$ cannot be determined accurately, (2.1.2) shows that $\emptyset$ should be at least as large as the mean service time. Under these conditions, i.e. $\emptyset=g_1$, the ratio $E(W_q)/E(W_q|\emptyset=g_1)$ is equal to $\{1-\rho(g_1)\}/\{1-\rho G(g_1)\}$. Table 2.1.1 gives values of this ratio for the three distributions D_1, D_2 and D_3. Provided traffic is quite heavy and it is possible to predict whether individual service times are shorter or longer than the average service time, practical reductions in mean queueing time can be obtained by serving customers with shorter service times first.

ρ	D_1	D_2	D_3
0.05	1.014	1.019	1.022
0.15	1.045	1.061	1.070
0.25	1.079	1.109	1.126
0.35	1.120	1.165	1.192
0.45	1.166	1.231	1.271
0.55	1.221	1.310	1.367
0.65	1.287	1.406	1.485
0.75	1.366	1.525	1.637
0.85	1.465	1.676	1.836
0.95	1.590	1.875	2.111

Table 2.1.1 Mean queueing time ratio showing the advantage of a priority discipline favouring customers with service times shorter than mean service time compared to service in order of arrival for traffic intensity ρ and D_1(Erlang), D_2(exponential) or D_3(mixed exponential) service times.

2.2 Evaluating some effects of a change in queue discipline

The results of §2.1 show that by using prior knowledge of customers' service times effectively, considerably reduced mean queueing times can be obtained in many situations. To explore this further we consider the changes in the queueing time distribution which this reduction in the mean value reflects.

Kesten & Runnenburg(1957) derive a general expression for the Laplace transform of the equilibrium queueing time distribution for customers in the kth class of an r-class non-preemptive priority discipline queue. Using their expressions for the case r=2, we can show that if $W_{q,j}$ is the queueing time for customers belonging to class j (j=1,2) and if $\emptyset>0$ is the service time which separates classes 1 and 2, then

$$E\left(e^{-sW_{q,1}} \,\middle|\, \phi\right) = -\frac{s(1-\rho)+\lambda\,\mathcal{G}(\phi)-\lambda\int_{\phi}^{\infty} e^{-sx} g(x)\,dx}{\lambda\, G(\phi) - s - \lambda\int_{0}^{\phi} e^{-sx} g(x)\,dx}\,, \qquad (2.2.1)$$

$$E\left(e^{-sW_{q,2}}\,\middle|\,\phi\right)=\frac{(1-\rho)\left\{z^*(s)-\lambda G(\phi)-s\right\}}{\lambda\mathcal{G}(\phi)-s-\lambda\int_\phi^\infty g(x)\exp\left[-x\left\{\lambda G(\phi)-z^*(s)+s\right\}\right]dx}\,, \quad (2.2.2)$$

where $\mathcal{G}(\phi)=1-G(\phi)$ and $z^*(s)$ satisfies the equation

$$z^*(s)=\lambda\int_0^\phi g(x)\exp\left[-x\left\{\lambda G(\phi)-z^*(s)+s\right\}\right]dx\ .$$

The Laplace transform of W_q is therefore a linear combination of (2.2.1) and (2.2.2) which cannot be simply expressed. Even when service times are independent, exponentially distributed with mean $1/\mu$, (2.2.1) becomes

$$E\left(e^{-sW_{q,1}}\,\middle|\,\phi\right)=1-\rho+\frac{s\lambda(1-\rho)+\lambda(\mu+s)\rho e^{-\mu\phi}+\lambda\mu(2-\rho)e^{-\phi(\mu+s)}}{s(\mu+s)+\lambda\mu\left\{1-e^{-\phi(\mu+s)}\right\}-\lambda(\mu+s)\left(1-e^{-\mu\phi}\right)}\ .$$

The queueing time distribution for each class of customers is a mixture of a discrete probability, $1-\rho$, at $t=0$ and a density for positive values of t. This is a form which direct considerations of queueing time distributions for M/G/1 systems would lead us to expect [cf. Cox & Smith (1961, pp.50-58)]. A non-preemptive queue discipline does not change the distribution of the length of the server's busy and idle periods; therefore, the long-run proportion of time that the server is idle is $1-\rho$. It follows that, with probability $1-\rho$, an arriving customer of either priority class will find the system empty and will thus avoid queueing. Since the proportion of queueing times which are zero is the same for service in order of arrival and for the priority discipline of §2.1, a reduction in mean queueing time indicates that the distribution of positive queueing times must be affected.

To determine more precisely how changes in the queue discipline affect the queueing time distribution we compare the queueing time distributions induced by service in order of arrival and by the priority discipline of §2.1 in identical circumstances. A second aspect of the analysis may suggest qualitative conclusions regarding queueing situations in which this simple priority discipline is most effective in reducing congestion.

The complicated forms of (2.2.1) and (2.2.2) indicate that a combination of numerical and analytical methods will be required. To examine both the quantitative and qualitative aspects of the comparison we need to consider specific service time distributions. The distributions D_1, D_2 and D_3 mentioned in §2.1 are simple examples of service time distributions with coefficients of variation, τ, less than, equal to and greater than unity, respectively. By using
D_1: $g(x)=xe^{-x}$ $(\tau=1/\sqrt{2})$, D_2: $g(x)=\frac{1}{2}e^{-\frac{1}{2}x}$ $(\tau=1)$, D_3: $g(x)=\frac{1}{2}(e^{-x}+\frac{1}{3}e^{-\frac{x}{3}})$ $(\tau=\sqrt{1.5})$ we should be able to draw practical, qualitative conclusions. The mean of each D_i is 2; therefore, changes in the traffic intensity can be obtained by adjusting the rate, λ, of the Poisson arrival process.

When customers are served in order of arrival, the Laplace transform of the equilibrium queueing time distribution is given by

$$E\left(e^{-sW_q}\right) = 1-\rho+(1-\rho)\frac{\lambda-\lambda g^*(s)}{s-\lambda+\lambda g^*(s)} \quad , \tag{2.2.3}$$

[cf. Cox & Smith(1961, p.57)], where $g^*(s)$ is the Laplace transform of $g(x)$, the service time probability density function. If the service time distribution is D_j $(j=1,2,3)$, the queueing time distribution may be obtained by inverting the particular form of (2.2.3). Thus

$$D_1: \quad \Pr\left(W_q>x \mid W_q>0\right) = \tfrac{1}{2}(1-\rho)\left\{\frac{a-2}{a(a-b)}e^{-ax} - \frac{b-2}{b(a-b)}e^{-bx}\right\}, \quad (x>0)$$

where $\quad a=\frac{1}{2}\left\{2-\lambda-(\lambda^2+4\lambda)^{\frac{1}{2}}\right\}, \qquad b=\frac{1}{2}\left\{2-\lambda+(\lambda^2+4\lambda)^{\frac{1}{2}}\right\}$;

$$D_2: \quad \Pr\left(W_q>x \mid W_q>0\right) = e^{-x(\frac{1}{2}-\lambda)}, \quad (x>0) \; ; \tag{2.2.4}$$

$$D_3: \quad \Pr\left(W_q>x \mid W_q>0\right) = \frac{1-\rho}{6}\left\{\frac{3c-2}{c(c-d)}e^{-cx} - \frac{3d-2}{d(c-d)}e^{-dx}\right\}, \quad (x>0)$$

where $\quad c=\frac{1}{6}\left\{4-3\lambda-(4+9\lambda^2)^{\frac{1}{2}}\right\}, \quad d=\frac{1}{6}\left\{4-3\lambda+(4+9\lambda^2)^{\frac{1}{2}}\right\}$.

Since (2.2.1) and (2.2.2) cannot be inverted, exact probabilities of various queueing times for the particular cases D_1, D_2 and D_3 cannot be obtained. However, the moments $E(W^k_{q,j}|\phi)$ $(j=1,2;\ k=1,2,\ldots)$ can be evaluated and some of these are given by

$$E\left(W_{q,1}|\phi\right)=\frac{1}{2}\,\frac{\lambda g_2}{1-\rho(\phi)}\,,\qquad E\left(W_{q,2}|\phi\right)=\frac{1}{2}\,\frac{\lambda g_2}{(1-\rho)\{1-\rho(\phi)\}}\,, \quad (2.2.5)$$

$$E\left(W^2_{q,1}|\phi\right)=\frac{\lambda g_3}{3\{1-\rho(\phi)\}}+\frac{1}{2}\,\frac{\lambda^2 g_2 g_2(\phi)}{\{1-\rho(\phi)\}^2}\,,$$

$$(2.2.6)$$

$$E\left(W^2_{q,2}|\phi\right)=\frac{\lambda g_3}{3\{1-\rho(\phi)\}^2(1-\rho)}+\frac{1}{2}\,\frac{\lambda^2 g_2 g_2(\phi)}{(1-\rho)\{1-\rho(\phi)\}^3}+\frac{1}{2}\,\frac{(\lambda g_2)^2}{\{1-\rho(\phi)\}^2(1-\rho)^2}\,,$$

$$E\left(W^3_{q,1}|\phi\right)=\frac{1}{4}\,\frac{\lambda g_4}{1-\rho(\phi)}+\frac{1}{2}\,\frac{\lambda^2 g_2 g_3(\phi)}{\{1-\rho(\phi)\}^2}+\frac{1}{2}\,\frac{\lambda^2 g_2(\phi) g_3}{\{1-\rho(\phi)\}^2}+\frac{1}{4}\,\frac{3\lambda^3 g_2\{g_2(\phi)\}^2}{\{1-\rho(\phi)\}^3}\,,$$

$$(2.2.7)$$

$$E\left(W^3_{q,2}|\phi\right)=\frac{1}{4}\,\frac{\lambda g_4}{(1-\rho)\{1-\rho(\phi)\}^3}+\frac{1}{4}\,\frac{3(\lambda g_2)^3}{(1-\rho)^3\{1-\rho(\phi)\}^3}+\frac{1}{2}\,\frac{3\lambda^3 g_2^2 g_2(\phi)}{(1-\rho)^2\{1-\rho(\phi)\}^4}+$$

$$\frac{1}{2}\,\frac{3\lambda^3 g_2\{g_2(\phi)\}^2}{(1-\rho)\{1-\rho(\phi)\}^5}+\frac{\lambda^2 g_2 g_3}{(1-\rho)^2\{1-\rho(\phi)\}^3}+\frac{\lambda^2 g_2(\phi) g_3}{(1-\rho)\{1-\rho(\phi)\}^4}+\frac{1}{2}\,\frac{\lambda^2 g_2 g_3(\phi)}{(1-\rho)\{1-\rho(\phi)\}^4}\,,$$

where $g_j(\phi)=\int_0^{\phi} t^j g(t)\,dt$ $(j=1,2,\ldots)$.

By using a probability distribution of known form to fit a queueing time distribution with unknown form but with known moments, the required probabilities can be estimated from the fitted distribution. To fit queueing time distributions we can use either the two-parameter lognormal or the gamma distribution [cf. Kendall & Stuart(1963, pp.152,168)].

The precise form of a fitted probability density function is determined by solving two equations which respectively equate the parametric mean and variance to specific values for the theoretical mean and variance of the queueing time distribution. The accuracy of this approximation can be estimated by comparing the skewness coefficients of the fitted and theoretical distributions.

A better approximation will be obtained if we only use a gamma or lognormal distribution to fit the conditional distribution of positive queueing times. Theoretical expressions for the moments of this conditional distribution can be obtained by multiplying the expressions in (2.2.5), (2.2.6) and (2.2.7) by $1/\rho$; then, the conditional queueing time distributions for priority classes 1 and 2 can be fitted as previously indicated. Conditional probabilities can be estimated using the fitted distributions and then linearly combined using the factors $G(\phi)$ and $\mathcal{G}(\phi)$ as class 1 and class 2 weighting factors, respectively.

Since the results will depend on ϕ, we use the values of ϕ^* plotted in Fig. 2.1.1 to fit the conditional queueing time distributions.

Table 2.2.1 shows the skewness coefficients $\gamma_{j,L}$ and $\gamma_{j,G}$ for the fitted lognormal and fitted gamma distributions, respectively, for class j (j=1,2) conditional queueing time distributions. The corresponding theoretical skewness coefficients, γ_j, are also given. In every case $\gamma_{j,G}$ more closely approximates γ_j than does $\gamma_{j,L}$ (j=1,2). Therefore, we use only gamma distributions to fit the 20 conditional queueing time distributions for each D_j (j=1,2,3). Secondly, apart from the seven instances $\rho=0.05,\ldots,0.65$ for the D_3 class 1 queueing time distributions, γ_j is usually less than $\gamma_{j,G}$ (j=1,2). Since the means and variances of the fitted and theoretical distributions are equal, this suggests that tail probabilities of the fitted densities for D_1, D_2 or $D_3(\rho=0.75,0.85,0.95)$ service times will be upper bounds for the exact probabilities specified by the theoretical queueing time distributions. Conveniently, the conditional probability that a customer who arrives during a busy period queues longer than a given length of time is an important indicator of the level of congestion in the system.

We can now calculate estimates, $\hat{Q}_k(\rho,\phi^*)$, which are probably upper bounds for the exact values of $Q_k(\rho,\phi^*)$, the conditional probability that a busy period arrival queues longer than k times the mean of the same conditional queueing time distribution. This conditional mean queueing time, $E(W_q|\phi=\phi^*,W_q>0)$, is equal to $E(W_q|\phi=\phi^*)/\rho$; a formula for $E(W_q|\phi)$ is given in (2.1.1). Thus, the mean of a given queueing time distribution is taken to be the unit of scale for that distribution; comparisons of probabili-

Service Times	Traffic Intensity ρ	Class 1 Distributions γ_1	$\gamma_{1,G}$	$\gamma_{1,L}$	Class 2 Distributions γ_2	$\gamma_{2,G}$	$\gamma_{2,L}$
	0.05	1.59	1.75	3.29	1.69	1.79	3.41
	0.15	1.54	1.71	3.20	1.80	1.85	3.56
	0.25	1.48	1.67	3.10	1.89	1.90	3.71
	0.35	1.41	1.63	3.00	1.96	1.95	3.87
	0.45	1.33	1.59	2.89	2.01	2.01	4.02
D_1	0.55	1.25	1.55	2.79	2.06	2.05	4.16
	0.65	1.18	1.50	2.68	2.08	2.10	4.30
	0.75	1.14	1.46	2.58	2.10	2.14	4.44
	0.85	1.22	1.44	2.53	2.11	2.17	4.54
	0.95	1.62	1.51	2.70	2.08	2.18	4.57
	0.05	1.98	1.98	3.94	2.01	2.01	4.02
	0.15	1.94	1.94	3.82	2.02	2.02	4.07
	0.25	1.89	1.90	3.70	2.03	2.04	4.12
	0.35	1.83	1.85	3.55	2.04	2.06	4.17
	0.45	1.75	1.79	3.40	2.04	2.07	4.23
D_2	0.55	1.65	1.72	3.23	2.05	2.09	4.28
	0.65	1.53	1.65	3.04	2.06	2.11	4.33
	0.75	1.38	1.56	2.83	2.06	2.12	4.38
	0.85	1.23	1.47	2.61	2.06	2.14	4.42
	0.95	1.37	1.44	2.52	2.04	2.13	4.41
	0.05	2.27	2.21	4.65	2.26	2.22	4.70
	0.15	2.23	2.16	4.50	2.22	2.21	4.67
	0.25	2.18	2.11	4.34	2.18	2.20	4.63
	0.35	2.12	2.05	4.16	2.15	2.19	4.60
	0.45	2.05	1.99	3.96	2.12	2.18	4.57
D_3	0.55	1.96	1.91	3.73	2.10	2.17	4.53
	0.65	1.84	1.82	3.47	2.08	2.16	4.50
	0.75	1.67	1.70	3.17	2.06	2.15	4.47
	0.85	1.43	1.56	2.82	2.05	2.14	4.42
	0.95	1.24	1.41	2.47	2.03	2.11	4.35

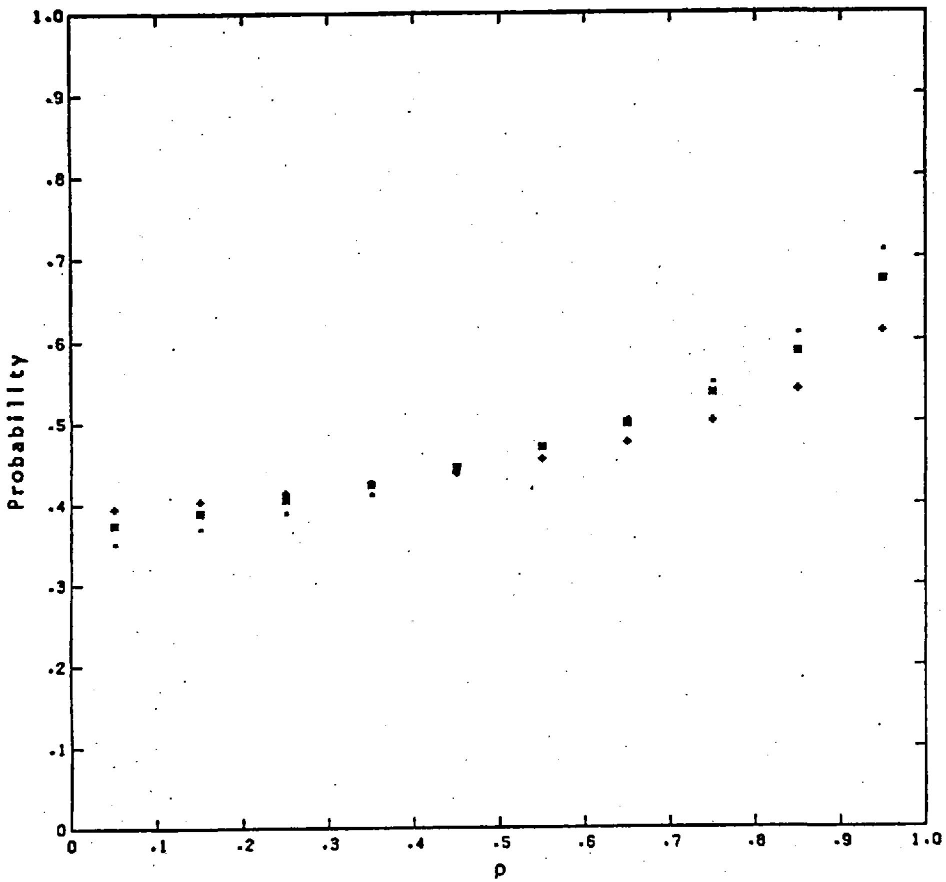
Table 2.2.1 Skewness coefficients, γ_j, $\gamma_{j,G}$ and $\gamma_{j,L}$ for class j conditional queueing time, fitted gamma and fitted lognormal distributions, respectively, (j=1,2) when service times are D (Erlang), D (ex onential` or D ´m´

ties among different queueing time distributions for the priority discipline will be at the same multiples of mean queueing time, the means being different. For fixed ρ and each service time distribution D_j we can also use the conditional mean queueing time, $E(W_q|\emptyset=\emptyset^*,W_q>0)$, to compare the distributions of positive queueing times determined by service in order of arrival and by the two-class priority discipline in identical circumstances, i.e. identical service time distributions and traffic intensities. To make this comparison we require $Q_k(\rho)$, the conditional probability that a busy period arrival in a service in order of arrival queue waits longer than k times the conditional mean queueing time, $E(W_q|\emptyset=\emptyset^*,W_q>0)$, in the priority queue. The $Q_k(\rho)$ can be calculated using the formulae given in (2.2.4).

The required probabilities are plotted in Figures 2.2.1, 2.2.2 and 2.2.3. Each figure is paired; the left half shows the exact probabilities, $Q_k(\rho)$, for service in order of arrival while the right half shows the estimated probabilities, $\hat{Q}_k(\rho,\emptyset^*)$, for the priority discipline of §2.1. The estimated probabilities were calculated by means of an algorithmic routine for evaluating the incomplete gamma function.

To determine the effect of the two-class priority discipline on the level of congestion consider the two halves of each figure individually. When ρ is very small, the two queue disciplines are negligibly different. Since $\hat{Q}_k(\rho,\emptyset^*)$ is an upper bound, the ratio $Q_k(\rho,\emptyset^*)/Q_k(\rho)$ is less than 0.9 when ρ=0.25. For moderately large ρ the difference between $Q_k(\rho)$ and $\hat{Q}_k(\rho,\emptyset^*)$ is more pronounced, and in conditions of very heavy traffic, e.g. ρ=0.95, the ratio $Q_k(\rho,\emptyset^*)/Q_k(\rho)$ is less than 0.25. By comparing the ratios $E(W_q)/E(W_q|\emptyset=\emptyset^*)$ (cf. §2.1) and $Q_k(\rho)/\hat{Q}_k(\rho,\emptyset^*)$ for $\rho \geq 0.55$ it may be seen that the reduction in long queueing times is more pronounced than the reduction in mean.

To determine the type of queueing situation in which this priority discipline is most effective in reducing congestion is more difficult. The difference, $Q_k(\rho)-\hat{Q}_k(\rho,\emptyset^*)$, is generally greatest for D_3, that is, when



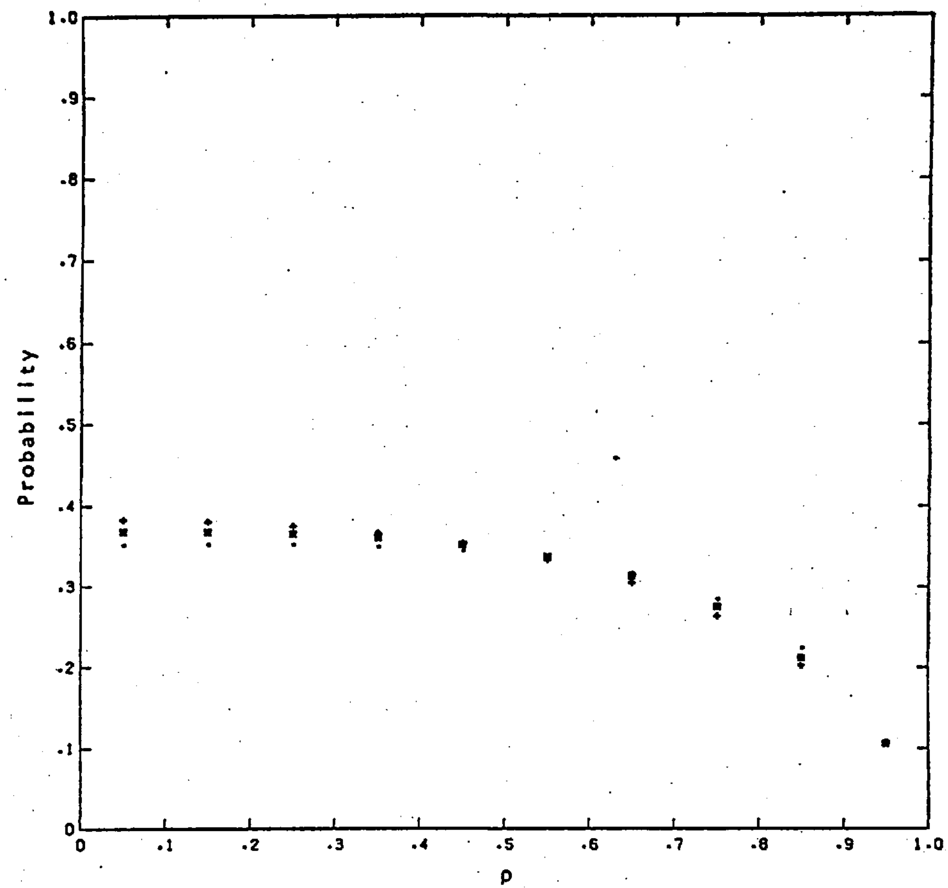
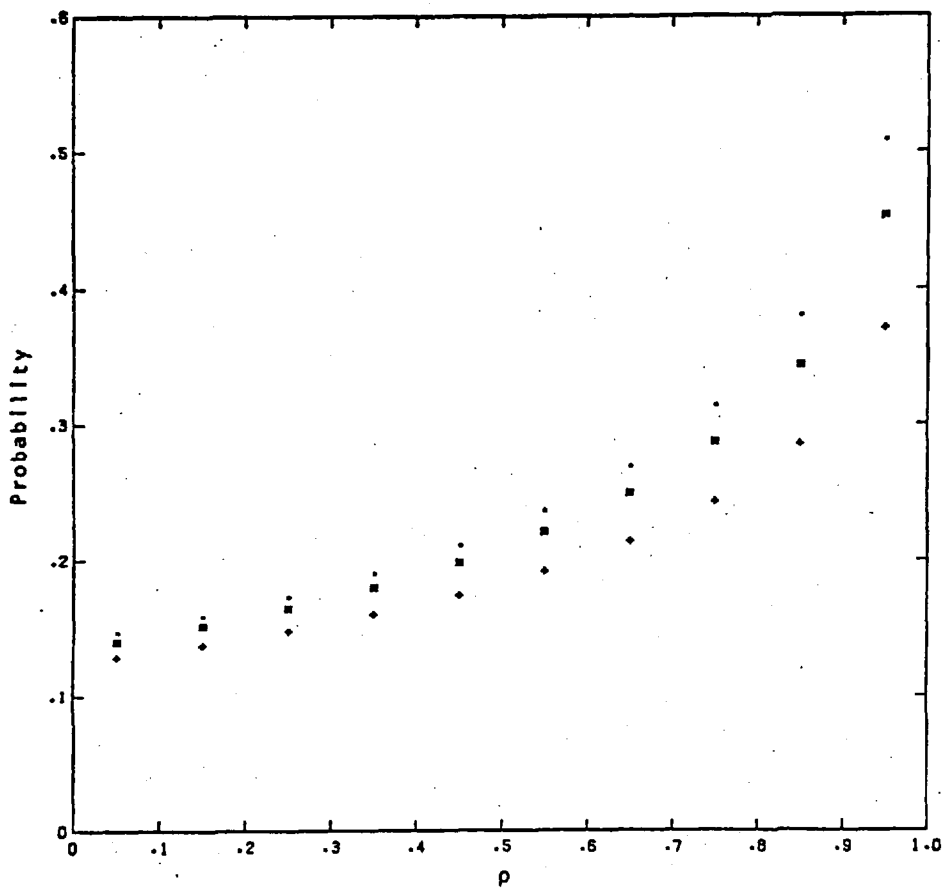


Fig. 2.2.1 a Conditional probabilities that busy period arrivals, served in order of arrival, queue longer than the conditional mean queueing time in an otherwise identical queue with two, optimal, service time dependent priorities, traffic intensity ρ, and D_1, D_2 or D_3 service times.

Fig. 2.2.1 b Estimated conditional probabilities that busy period arrivals in a queue with two, optimal, service time dependent priorities queue longer than the conditional mean queueing time. Service times are D_1, D_2 or D_3 with traffic intensity ρ.

+ D_1 (Erlang) ■ D_2 (exponential) • D_3 (mixed exponential)



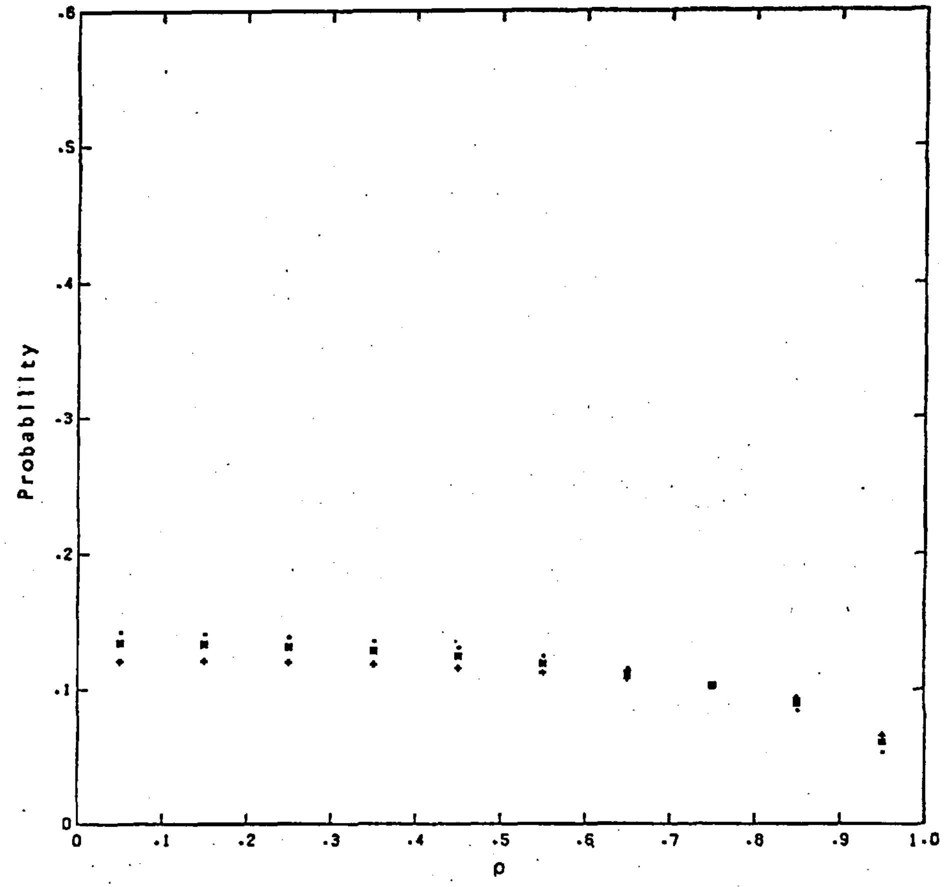
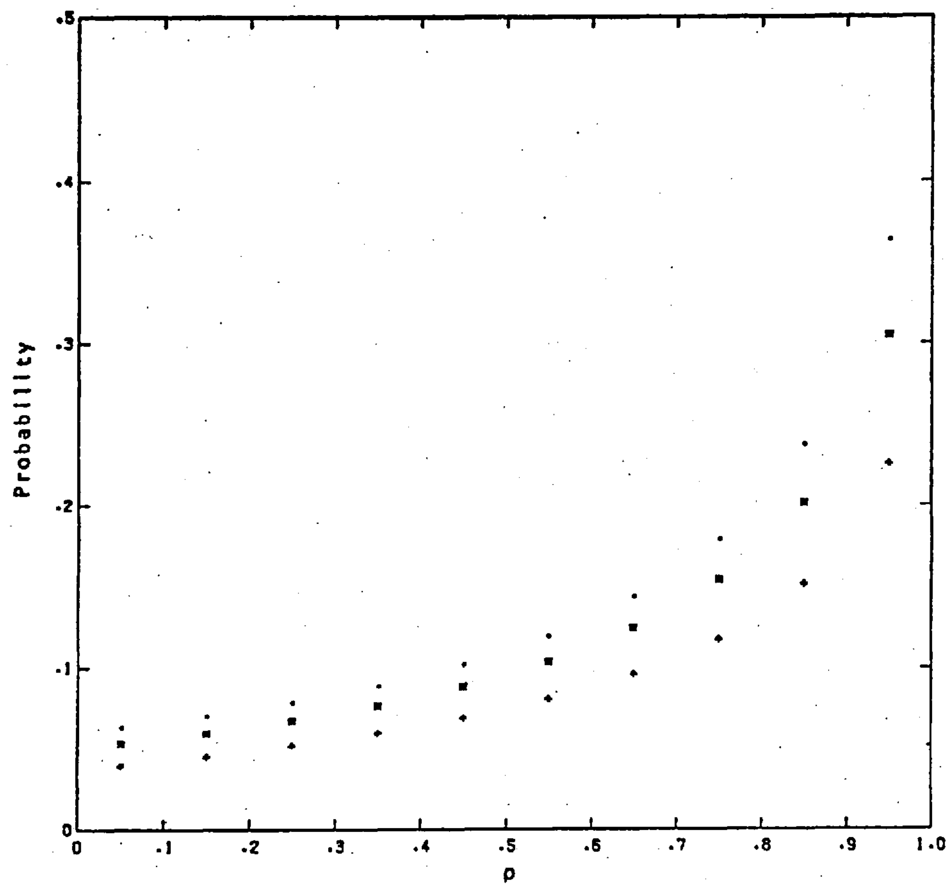


Fig. 2.2.2 a Conditional probabilities that busy period arrivals, served in order of arrival, queue longer than twice the conditional mean queueing time in an otherwise identical queue with two, optimal, service time dependent priorities, traffic intensity ρ, and D_1, D_2 or D_3 service times.

Fig. 2.2.2 b Estimated conditional probabilities that busy period arrivals in a queue with two, optimal, service time dependent priorities queue longer than twice the conditional mean queueing time. Service times are D_1, D_2 or D_3 with traffic intensity ρ.

+ D_1 (Erlang) ■ D_2 (exponential) • D_3 (mixed exponential)



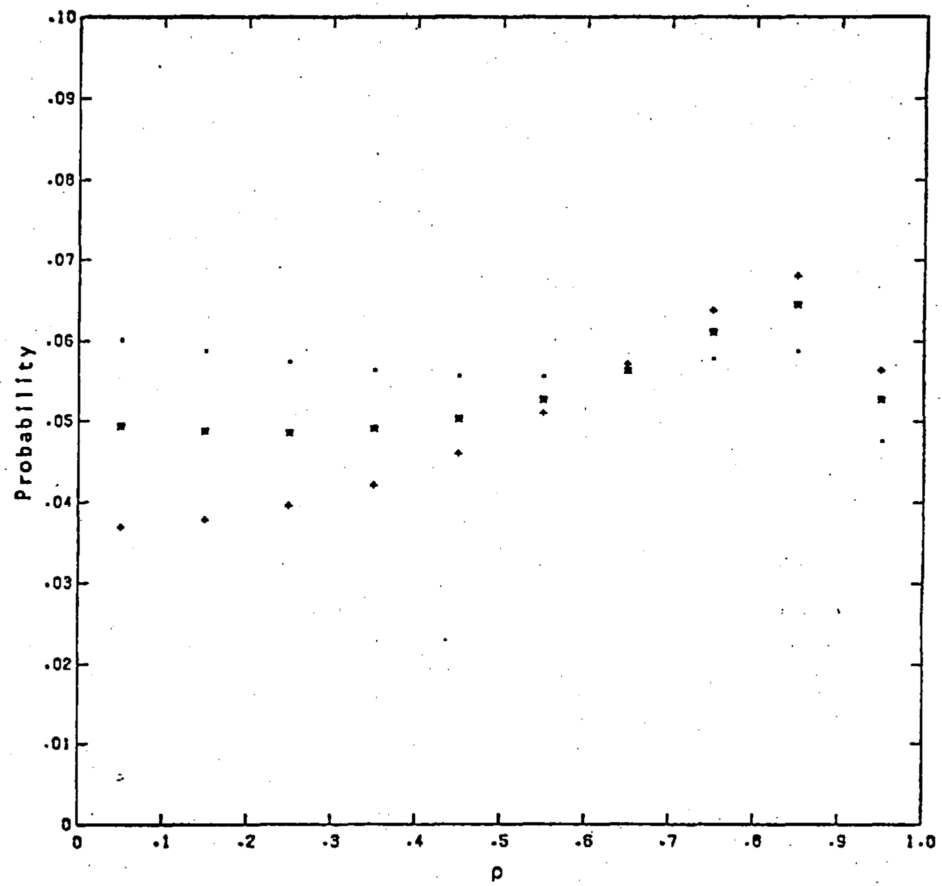


Fig. 2.2.3 a Conditional probabilities that busy period arrivals, served in order of arrival, queue longer than three times the conditional mean queueing time in an otherwise identical queue with two, optimal, service time dependent priorities, traffic intensity ρ, and D_1, D_2 or D_3 service times.

Fig. 2.2.3 b Estimated conditional probabilities that busy period arrivals in a queue with two, optimal, service time dependent priorities queue longer than three times the conditional mean queueing time. Service times are D_1, D_2 or D_3 with traffic intensity ρ.

\+ D_1 (Erlang) ■ D_2 (exponential) • D_3 (mixed exponential) service times vary considerably. While most customers are assigned to the priority 1 class, customers with very long service times, who cause much of the queueing, are relegated to class 2. Therefore, when long service times occur more frequently, i.e. $\tau>1$, the effect of the priority discipline should be more pronounced. Joint comparison of Figs. 2.2.1b, 2.2.2b and 2.2.3b appears to indicate that by using the same multiples of $E(W_q|\emptyset=\emptyset^*, W_q>0)$ differences among the priority discipline queueing time distributions for the D_j's have been eliminated. Although this effect was intended, real distinctions among the distributions of positive queueing times may be hidden by the different accuracies of the estimates, $\hat{Q}_k(\rho,\emptyset^*)$, for each D_j. This differing degree of accuracy is reflected in the columns for γ_j and $\gamma_{j,G}$ in Table 2.2.1. If we interpret the dimensionless quantity, δ, where $\delta=G(\emptyset^*)\frac{\gamma_{1,G}}{\gamma_1}+\bar{G}(\emptyset^*)\frac{\gamma_{2,G}}{\gamma_2}-1$, as indicating the accuracy with which the estimates, $\hat{Q}_k(\rho,\emptyset^*)$, bound the true probabilities $Q_k(\rho,\emptyset^*)$, then probable underestimates or over-estimates of $Q_k(\rho,\emptyset^*)$ correspond to negative and positive values of δ, respectively. A value of δ for each estimation situation is given in Table 2.2.2. The $\hat{Q}_k(\rho,\emptyset^*)$ appear to be most accurate when service times are exponentially distributed. Table 2.2.2 also indicates that unless $\rho\geq 0.75$, we need to regard $\hat{Q}_k(\rho,\emptyset^*)$ with some suspicion for D_3 service times.

Provided service times are not constant, the simple priority discipline of §2.1 substantially reduces most queueing times. Only customers with rather long service times are inconvenienced; these individuals generally experience very long queueing times as well.

	δ		
ρ	D_1	D_2	D_3
0.05	0.083	0.001	-0.023
0.15	0.080	0.002	-0.022
0.25	0.089	0.006	-0.020
0.35	0.106	0.011	-0.019
0.45	0.132	0.020	-0.018
0.55	0.167	0.036	-0.014
0.65	0.203	0.064	-0.004
0.75	0.220	0.112	0.021
0.85	0.155	0.174	0.086
0.95	-0.056	0.044	0.131

Table 2.2.2 Relative accuracy, δ, of a gamma approximation to a conditional queueing time distribution with traffic intensity, ρ, and D_1(Erlang), D_2(exponential) or D_3(mixed exponential) service times. If $\delta>0(<0)$, the approximation over-estimates (underestimates) tail probabilities.

2.3 More complicated control techniques — k-class non-preemptive priority disciplines

The results of §§2.1 and 2.2 suggest that if customers can be accurately divided into k (k=3,4,...) homogeneous classes according to their service times, and if non-preemptive priorities are assigned to classes sensibly, average queueing times should decrease as k increases.

Suppose that $0=\emptyset_0<\emptyset_1<\cdots<\emptyset_k=\infty$ and an arriving customer is assigned to priority class j if his future service time, X, is such that $\emptyset_{j-1}\leq X<\emptyset_j$ (j=1,...,k). Provided the arrival process is independent of the queue discipline and the traffic intensity is less than unity, we can show, by a relatively simple argument, that average queueing time always decreases whenever an additional priority class is added to the priority discipline defined above. The following background, which is due to Kingman(1962), serves as a basis for the argument.

Suppose that, during a single busy period of any stable ($\rho<1$) queueing process, n customers, labelled 1,2,...,n in order of arrival, join the queue and are served. Let $\{A_i\}$ and $\{S_i\}$ ($i=1,\ldots,n$) be the sequences of times at which the ith customer arrives in the queue and enters service, respectively. Then $Q_i=S_i-A_i>0$ ($i=1,\ldots,n$) is the queueing time of the ith customer and the average queueing time during the busy period is $\frac{1}{n}\sum_{i=1}^{n} Q_i$. A queue discipline is therefore an ordering of the n customers who are served during the busy period. More precisely, a queue discipline is a permutation $\pi \in \Pi$ on the n-tuple $(1,\ldots,n)$ determined by the chronological ordering of the S_i's, where Π is the set of all permitted queue disciplines. Thus, the identity permutation on $(1,\ldots,n)$ represents service in order of arrival. Since S_i and Q_i vary according to the queue discipline, these times are better represented by $S_i(\pi)$ and $Q_i(\pi)$ ($\pi \in \Pi$; $i=1,\ldots,n$), respectively. We can now prove the result; the argument depends on an interchange technique similar to that used by Schrage (1968).

Let π be the priority discipline which serves customers according to the classes $[\emptyset_{j-1},\emptyset_j)$ ($j=1,\ldots,k$). From any class $[\emptyset_{r-1},\emptyset_r)$ create two new classes, $[\emptyset_{r-1},\emptyset_s)$ and $[\emptyset_s,\emptyset_r)$, where $\emptyset_{r-1}<\emptyset_s<\emptyset_r$, and let π^+ be the queue discipline which serves customers according to this expanded version of the k-class priority scheme. Obviously, on each busy period of the queueing process, either π and π^+ are identical or π and π^+ differ. Whenever π and π^+ are identical, $\frac{1}{n}\sum_{i=1}^{n} Q_i(\pi)=\frac{1}{n}\sum_{i=1}^{n} Q_i(\pi^+)$; therefore, consider any busy period on which π and π^+ differ.

Let X_i be the service time of the ith arrival ($i=1,\ldots,n$). Since π and π^+ differ, there is a first instant w when, for two customers p and q, $\emptyset_r>X_p>\emptyset_s>X_q\geq\emptyset_{r-1}$, $S_p(\pi)=w$, $A_q<w$ and $S_q(\pi)>w$. Let π_1 be a discipline which coincides with π in every respect save that π_1 interchanges q and p in the serving order. Clearly,

$$\frac{1}{n}\sum_{i=1}^{n} Q_i(\pi_1) < \frac{1}{n}\left\{\sum_{i=1}^{n} Q_i(\pi_1)+X_p-X_q\right\} \leq \frac{1}{n}\sum_{i=1}^{n} Q_i(\pi) \quad .$$

Either π_1 is identical to π^+ or there exists a first instant $z>w$ when, for two customers u and v, $\emptyset_r>X_u>\emptyset_s>X_v\geq\emptyset_{r-1}$, $S_u(\pi_1)=z$, $A_v<z$ and $S_v(\pi_1)>z$. Let π_2 be a discipline which coincides with π_1 in every respect save that π_2 interchanges u and v in the serving order. Then

$$\frac{1}{n}\sum_{i=1}^{n} Q_i(\pi_2) < \frac{1}{n}\sum_{i=1}^{n} Q_i(\pi_1) < \frac{1}{n}\sum_{i=1}^{n} Q_i(\pi) \quad .$$

If π_2 is not π^+ then, by a finite sequence of pairwise interchanges we can obtain π^+ and

$$\frac{1}{n}\sum_{i=1}^{n} Q_i(\pi^+) < \cdots < \frac{1}{n}\sum_{i=1}^{n} Q_i(\pi) \quad .$$

By averaging over a large number of busy periods we can see that the average queueing time for π^+ is less than the average queueing time for π.

The preceding argument does not depend on assumptions regarding specific arrival or service time distributions. When arrivals are Poisson, Cobham's(1954) results give the expression

$$E\left(W_q \mid \underline{\phi}\right) = \frac{1}{2}\lambda g_2 \sum_{i=1}^{k} \frac{G(\phi_i) - G(\phi_{i-1})}{\{1-\rho(\phi_{i-1})\}\{1-\rho(\phi_i)\}} \quad . \tag{2.3.1}$$

for the mean queueing time, where $\underline{\emptyset}=(\emptyset_0,\ldots,\emptyset_k)$, $G(\cdot)$ is the service time distribution function with derivative $g(\cdot)$ and $\rho(x)=\lambda\int_0^x t\,g(t)dt<\rho<1 \quad (x\geq 0)$.

For fixed $k>2$, Oliver & Pestalozzi(1965) use dynamic programming to determine $\underline{\emptyset}^*$, the value of $\underline{\emptyset}$ which minimizes (2.3.1). Whether the reduction in mean queueing time will compensate sufficiently for the increased administrative load will depend upon many factors including the number of extra classes added and the accuracy with which customers are assigned to their respective priority classes; it is difficult to quantify this.

2.4 Idealized control techniques — a different priority class for each customer

The central argument of §2.3 suggests that if customers' exact service times are known in advance, then, to minimize average queueing time over all customers, the waiting customer with the smallest service time should always be served next. Phipps(1956) was the first to consider this shortest service time discipline. Schrage & Miller(1966) point out that Phipps's shortest service time rule is a non-preemptive special case of their shortest remaining processing time discipline. Subsequently, Schrage(1968) proves that the shortest remaining processing time rule minimizes the number of customers in the system; that is, if the queue discipline is to serve, preemptively, the customer with the shortest remaining processing time, then the number of customers in the queueing system never exceeds the line length for any other rule simultaneously acting on the same sequence of arrivals and processing times.

By retaining the assumptions of §2.3 and by slightly modifying the argument given there we can show that, if the queue discipline is to serve the customer with the shortest service time at each service epoch, the mean queueing time, averaged over all customers, is minimized with respect to all non-preemptive rules applied to the same sequence of arrival and service times. Although the proof is similar to that which Schrage(1968) gives, the addition of Kingman's(1962) framework (cf. §2.3) substantially improves the argument. The proof begins with any permissible discipline π_1, say, which is not shortest service time and uses the pairwise interchange technique to establish the result. No other details will be furnished since the proof is very similar to that outlined in §2.3.

As in §2.3, the argument does not depend on assumptions regarding specific arrival or service time distributions. For Poisson arrivals, Phipps(1956) has derived expressions for $E(L_q)$ and $E(W_q|s)$, the mean queueing time for a customer whose service time is s. These expressions depend on the assumption that customers' service times are known exactly before they are served. Since estimates of customers' service times may be somewhat uncertain, we now consider how to incorporate this uncertainty into an expression for the mean queueing time when the queue discipline is the shortest service time rule.

Suppose that customers who join an M/G/1 queue which is in equilibrium are assigned non-preemptive priorities S_1, S_2,... which are independent, identically distributed observations from a priority assignment probability distribution; the distribution is arbitrary up to monotonic transformation. For convenience we assume that $P(s)=\mathrm{pr}(S_i \leq s)$ $(i=1,2,\ldots;$ $s \geq 0)$. Whenever a customer departs, the next customer to be served is always the one whose priority is greatest, i.e. the customer whose assigned priority is numerically smallest. However priorities are assigned, e.g. randomly, we shall require that $\mu_{j,s}=E(X_s^j)$ $(j=1,2;\ s \geq 0)$ can be evaluated, where $\mu_{j,s}$ is the conditional jth moment about the origin of the service time, X_s, of customers belonging to the priority class with index $s \in [0,\infty)$. Let $\mu_s \equiv \mu_{1,s}$, and suppose that s is a continuity point of $P(\cdot)$. By adapting Phipps's(1956) argument to the results of Kesten & Runnenburg (1957) we can show that the first two moments of the queueing time distribution for a customer with priority s are given by

$$E\left(W_q \mid s\right) = \frac{1}{2}\,\frac{\lambda g_2}{\left\{1-\lambda\int_0^s \mu_t\, dP(t)\right\}^2}\ , \tag{2.4.1}$$

$$E\left(W_q^2 \mid s\right) = \frac{1}{3}\,\frac{\lambda g_3}{\left\{1-\lambda\int_0^s \mu_t\, dP(t)\right\}^3} + \frac{\lambda^2 g_2\left\{\int_0^s \mu_{2,t}\, dP(t)\right\}}{\left\{1-\lambda\int_0^s \mu_t\, dP(t)\right\}^4}\ . \tag{2.4.2}$$

By using Laplace transform techniques, both Takács(1964) and Cohen(1969, p.454) derive expressions which agree with (2.4.1) and (2.4.2).

Integrate (2.4.1) and (2.4.2) with respect to s, the assigned priority. It follows that

$$E\left(W_q\right) = \frac{1}{2}\lambda g_2 \int_0^\infty \frac{dP(s)}{\left\{1-\lambda\int_0^s \mu_t\, dP(t)\right\}^2}\ , \tag{2.4.3}$$

$$E\left(W_q^2\right) = \lambda g_3 \int_0^\infty \frac{dP(s)}{3\left\{1-\lambda\int_0^s \mu_t\, dP(t)\right\}^3} + \lambda^2 g_2 \int_0^\infty \frac{\left\{\int_0^s \mu_{2,t}\, dP(t)\right\} dP(s)}{\left\{1-\lambda\int_0^s \mu_t\, dP(t)\right\}^4} \quad . \quad (2.4.4)$$

We can use (2.4.3) and (2.4.4) to obtain both familiar and new results for the M/G/1 queue. The next three examples are important special cases.

Example 2.4.1

Since (2.4.3) is a generalization of Phipps's(1956) result, we can obtain Phipps's expression for $E(W_q)$ by setting $P(s)=G(s)$ and $E(X_s^j)=s^j$ $(j=1,2;\ s\geq 0)$. Hence, if $g_j(t)=\int_0^t x^j g(x)dx$ $(j=1,2;\ t>0)$ and $\lambda g_1(t)=\rho(t)<1$ for $t>0$,

$$E\left(W_q\right) = \frac{1}{2}\lambda g_2 \int_0^\infty \frac{g(t)\,dt}{\{1-\rho(t)\}^2} \ ,$$

and

$$E\left(W_q^2\right) = \frac{\lambda g_3}{3}\int_0^\infty \frac{g(s)\,ds}{\{1-\rho(s)\}^3} + \lambda^2 g_2 \int_0^\infty \frac{g_2(s)\, g(s)\, ds}{\{1-\rho(s)\}^4}$$

may be used to evaluate $\mathrm{Var}(W_q)$.

Suppose that the priorities assigned to customers are uniformly distributed over some fixed interval, say [0,1], and this assignment does not depend on service times. Obviously, the advantages of the shortest service time discipline will be eliminated. Moreover, since assigned priorities cannot be changed, low priority customers generally have long queueing times. When customers are served in random order, however, at any service epoch each customer has an equal chance of being served next. Therefore, queueing times should be more regular if customers are served in random order than if customers are served according to priorities assigned at random.

Example 2.4.2

If fixed priorities are randomly assigned, independent of each customer's service time, then

$$P(s)=\begin{cases} s & 0\leq s\leq 1 \\ 1 & s>1 \end{cases} \quad , \quad E(X_s^j)=g_j \quad (j=1,2;\ s\geq 0)$$

After simplifying, (2.4.3) and (2.4.4) become

$$E(W_q) = \frac{1}{2}\,\frac{\lambda g_2}{1-\rho}\,, \qquad Var(W_q) = \frac{\lambda g_3(2-\rho)}{6(1-\rho)^2} + \frac{\lambda^2 g_2^2(3+\rho)}{12(1-\rho)^3}\,. \qquad (2.4.5)$$

This random priority discipline satisfies the conditions of Kingman (1962); therefore, the mean queueing time is identical to the Pollaczek-Khintchine formula and to mean queueing time for service in random order. Cohen(1969, p.431) indicates that the variance of W_q for service in random order is $\frac{2}{3}\,\frac{\lambda g_3}{(1-\rho)(2-\rho)} + \frac{(\lambda g_2)^2(2+\rho)}{4(1-\rho)^2(2-\rho)}$. This is smaller than (2.4.5), indicating that queueing times are more regular if customers are served in random order than if customers are assigned fixed priorities at random.

Conway & Maxwell(1962) mention a queue discipline which is the antithesis of the shortest service time rule — at each service epoch, always serve the customer with the longest service time. According to these authors, this longest service time discipline can occur in practical situations. Apparently, a customer's importance is associated with the length of his service time; therefore, longer jobs are given priority over shorter ones.

When long waiting times are subject to severe economic penalties it is probably sensible to give priority to the customer with the longest service time. Obviously, this particular queue discipline increases queueing times in contrast to, say, service in order of arrival. Although the longest service time rule can occur in practice, this queue discipline has been given little attention in the literature; the expressions derived in Example 2.4.3 appear to be new.

Example 2.4.3

Since a customer's priority is inversely proportional to his service time, $S \leq s$ if and only if $X \geq s^{-1}$, where S is the customer's assigned priority and X is his service time. Therefore, $P(s) = \mathcal{G}(s^{-1})$ and (2.4.1) and (2.4.2) become

$$E(W_q|s) = \frac{1}{2}\,\frac{\lambda g_2}{\{1-\rho+\rho(s^{-1})\}^2}\,, \quad E(W_q^2|s) = \frac{\lambda g_3}{3\{1-\rho+\rho(s^{-1})\}^3} + \frac{\lambda^2 g_2\{g_2 - g_2(s^{-1})\}}{\{1-\rho+\rho(s^{-1})\}^4}\,.$$

The averaged first and second moments of W_q are given by

$$E(W_q) = \frac{1}{2}\lambda g_2 \int_0^\infty \frac{g(z)\,dz}{\{1-\rho+\rho(z)\}^2}$$

$$E(W_q^2) = \lambda g_3 \int_0^\infty \frac{g(z)\,dz}{3\{1-\rho+\rho(z)\}^3} + \lambda^2 g_2 \int_0^\infty \frac{\{g_2 - g_2(z)\}\,g(z)\,dz}{\{1-\rho+\rho(z)\}^4} .$$

We have already established that no non-preemptive queue discipline can be devised which is superior to the shortest service time rule in minimizing the mean queueing time, averaged over all customers. Under the same assumptions, i.e. that the arrival process is independent of the queue discipline and $\rho < 1$, we can show that when the queue discipline is to serve the customer with the longest service time, the mean queueing time, averaged over all customers, will be maximized for all non-preemptive rules applied to the same sequence of arrival and service times. The similarities between the two results are obvious; therefore, details of an argument proving the result for the longest service time discipline can be omitted.

To illustrate some of the differences among the queue disciplines which we have considered, expressions for $E(W_q)$ and $Var(W_q)$ from §§2.2 and 2.4 have been evaluated numerically for different values of ρ. The service time distributions used in the calculations are the three simple examples D_1, D_2 and D_3 of §§2.1 and 2.2. Tables 2.4.1 and 2.4.2 give the results of the calculations.

The tabulated mean values show how much mean queueing times can be reduced if advance information about service times is used sensibly. The column corresponding to the longest service time queue discipline indicates, quantitatively, the effect which this rule has on mean queueing time, averaged over all customers. For the two queue disciplines with mean queueing times smaller than that for service in order of arrival, the reduction in mean value is greatest as $\rho \to 1$; similarly, for fixed ρ, the same reduc-

Service Times	Traffic Intensity ρ	Shortest Service Time	Optimal 2-class Priorities	Service in Order of Arrival	Random Priority Service	Longest Service Time
D_1	0.05	0.077	0.078	0.079	0.079	0.082
	0.15	0.249	0.253	0.265	0.265	0.299
	0.25	0.449	0.463	0.500	0.500	0.620
	0.35	0.689	0.720	0.808	0.808	1.12
	0.45	0.985	1.05	1.23	1.23	1.92
	0.55	1.37	1.49	1.83	1.83	3.34
	0.65	1.90	2.12	2.79	2.79	6.11
	0.75	2.74	3.16	4.50	4.50	12.7
	0.85	4.40	5.31	8.50	8.50	34.9
	0.95	10.8	14.1	28.5	28.5	261.
D_2	0.05	0.103	0.103	0.105	0.105	0.108
	0.15	0.326	0.333	0.353	0.353	0.383
	0.25	0.578	0.601	0.667	0.667	0.771
	0.35	0.871	0.922	1.08	1.08	1.34
	0.45	1.22	1.32	1.64	1.64	2.22
	0.55	1.66	1.85	2.44	2.44	3.67
	0.65	2.24	2.58	3.71	3.71	6.35
	0.75	3.11	3.74	6.00	6.00	12.2
	0.85	4.73	6.05	11.3	11.3	30.1
	0.95	10.5	15.1	38.0	38.0	179.
D_3	0.05	0.128	0.129	0.132	0.132	0.135
	0.15	0.403	0.412	0.441	0.441	0.483
	0.25	0.709	0.740	0.833	0.833	0.980
	0.35	1.06	1.13	1.35	1.35	1.72
	0.45	1.47	1.60	2.05	2.05	2.87
	0.55	1.96	2.21	3.06	3.06	4.79
	0.65	2.60	3.04	4.64	4.64	8.39
	0.75	3.52	4.33	7.50	7.50	16.4
	0.85	5.14	6.80	14.2	14.2	41.1
	0.95	10.6	16.0	47.5	47.5	250.

Table 2.4.1 Mean queueing times for five different queue disciplines when service times are D_1(Erlang), D_2(exponential) or D_3(mixed exponential) with traffic intensity ρ.

Service Times	Traffic Intensity ρ	Shortest Service Time	Optimal 2-class Priorities	Service in Order of Arrival	Random Priority Service	Longest Service Time
	0.05	0.209	0.210	0.217	0.223	0.248
	0.15	0.697	0.706	0.776	0.855	1.17
	0.25	1.34	1.36	1.58	1.92	3.25
	0.35	2.30	2.32	2.81	3.85	8.07
	0.45	3.95	3.86	4.78	7.76	20.1
D_1	0.55	7.30	6.74	8.25	16.7	54.1
	0.65	15.7	13.2	15.2	41.3	170.
	0.75	44.3	32.4	32.3	131.	719.
	0.85	211.	124.	94.9	705.	5770.
	0.95	5510.	2100.	888.	22200.	406000.
	0.05	0.411	0.415	0.432	0.444	0.479
	0.15	1.32	1.36	1.54	1.69	2.12
	0.25	2.43	2.53	3.11	3.75	5.49
	0.35	3.94	4.13	5.47	7.46	12.8
	0.45	6.32	6.55	9.22	14.8	29.7
D_2	0.55	10.8	10.8	15.8	31.5	74.2
	0.65	21.5	19.8	28.7	76.6	215.
	0.75	56.4	45.1	60.0	240.	831.
	0.85	253.	160.	174.	1270.	5890.
	0.95	6280.	2400.	1600.	39600.	330000.
	0.05	0.714	0.721	0.754	0.775	0.839
	0.15	2.26	2.32	2.67	2.93	3.72
	0.25	4.05	4.26	5.36	6.45	9.65
	0.35	6.34	6.76	9.35	12.7	22.4
	0.45	9.67	10.4	15.6	24.9	52.1
D_3	0.55	15.5	16.4	26.4	52.1	130.
	0.65	28.6	28.7	47.6	125.	378.
	0.75	69.2	61.3	98.3	386.	1450.
	0.85	292.	202.	280.	2020.	10300.
	0.95	6920.	2740.	2520.	62200.	585000.

Table 2.4.2 Queueing time variances for five different queue disciplines when service times are D_1(Erlang), D_2(exponential) or D_3(mixed exponential) with traffic intensity ρ.

tion is an increasing function of τ, the service time distribution coefficient of variation. This latter aspect underlines a previous suggestion that queue disciplines which use prior information regarding service times are probably most effective in reducing congestion when service times are frequently quite different from their mean value.

While Table 2.4.1 verifies that the shortest service time discipline minimizes mean queueing time, it is also apparent that the optimal, two-class priority rule which was discussed in §§2.1 and 2.2 has at least one important characteristic. Both tables show that this practical queue discipline generally induces a queueing time distribution which has a smaller mean and variance than the comparable distribution when the queue discipline is service in order of arrival. The entries in Table 2.4.2 also reflect the hidden disadvantages of the shortest service time rule. Conway & Maxwell (1962) point out that this ideal queue discipline reduces mean queueing times, overall, at the expense of customers whose service times are long. The shortest service time rule usually means that their queueing times will be excessively long as well. This disadvantage is shown in Table 2.4.2 in the values for queueing time variance; these exceed corresponding entries for service in order of arrival or the optimal, two-class priority rule in conditions of heavy traffic. The rule of §§2.1 and 2.2 generally appears to be a better method of reducing congestion than service in order of arrival or the queue discipline which always serves the customer with the shortest service time.

The entries in Tables 2.4.1 and 2.4.2 also show that service in order of arrival causes less congestion than a queue discipline which assigns fixed priorities at random. Similarly, the values of $E(W_q)$ and $Var(W_q)$ for the longest service time rule emphasize that the reasons for choosing this queue discipline must be economic, since this rule appears to maximize not only the mean but also the variance of the queueing time.

There are many other priority queueing disciplines which we have not considered, including several types of preemptive disciplines. No doubt some of these disciplines would be more suited to individual queueing situations, particularly if service preemptions involve little loss of time. However, the optimal, two-class priority rule of §§2.1 and 2.2 is a practical alternative to service in order of arrival whenever administrative simplicity and effectiveness in controlling congestion are major considerations.

CHAPTER 3. Controlling congestion behind the counter

3.1 Linking shut-down control of the service process to queue size

Ordinarily, congestion in the queue is the primary concern. However, the situation behind the service counter may also require attention. Yadin & Naor(1963) point out that if, for example, an M/G/1 queue is organized so that $E(L_q) \leq 1$, the server may be idle nearly 30% of the time. Unless the unit cost of serving time is quite small, this idle fraction probably means resources are being wasted. Suppose that to do ancillary tasks the server could use, profitably, idle periods exceeding ω times the mean length of an idle period. Such idle periods occur with probability $e^{-\omega}$ which may be quite small. By reorganizing the service process, however, it might be possible to create some longer idle periods which the server could suitably exploit.

Yadin & Naor(1963) proposed shut-down control as a means of reorganizing the service process in order to increase the length of individual idle periods. Simplicity is an important feature of this method; two operating phases characterize the reorganized service process. During a shut-down phase the server does ancillary work and a queue of customers forms. Once the queue equals a predetermined size, the server begins to serve the waiting customers. This latter phase is usually called the busy period. A busy period ends and a new shut-down phase begins when a departing customer leaves behind an empty system.

Shut-down control is usually linked to the queue size; the critical value, N say, gives rise to the name (0,N) control which is sometimes used in the literature [cf. Yadin & Naor(1963), Bell(1971)]. However, other properties of the queueing process can be used to determine when a shut-down phase should end. In §3.2 we consider a control method which monitors the virtual queueing time. When this quantity exceeds a threshold level V during a shut-down phase, another busy period begins.

The (0,N) version of shut-down control has been investigated from two main perspectives. In their introductory paper, Yadin & Naor(1963) derive an expression for E(L), the average line length in the steady-state. In addition to the usual assumptions of (0,N) control, these authors include random start-up and shut-down intervals at the phase change epochs. In a brief discussion of costs, a total system operating cost which is linear with respect to the average value of each of its components is proposed. This assumption enables the authors to derive a simple expression for the value of N which minimizes the marginal cost per unit time of introducing (0,N) control into a steady-state M/G/1 queue.

Heyman(1968) and Bell(1971) arrive at (0,N) control from a different starting-point; Bell corrects and improves the results which Heyman obtained. Initially, both authors consider an M/G/1 queueing system with a removable server. The system operating cost is assumed to be linear, consisting of different unit costs for customer waiting time, server idle time, server running time and fixed start-up and shut-down charges. Each author sets out to prove that shut-down control is the optimal stationary operating policy for continuously discounted, infinite-horizon, expected operating costs. Bell first establishes that an optimal stationary operating policy has a form which is no more than a simple variation on shut-down control. Then, using a Markov renewal programming formulation he proves that shut-down control is the optimal stationary operating policy for an M/G/1 queue with the given operating cost structure.

The theorems upon which the results of both Heyman(1968) and Bell(1971) depend only require bounded expected costs between transitions in the state space. The simple assumptions concerning operating costs probably simplify the necessary algebraic manipulations. Neither author indicates how different assumptions about system operating costs might affect their conclusions.

Adopting (0,N) control in a queueing system changes important stochastic processes, such as the queue size distribution. To assess how effectively shut-down control reorganizes the queue we need to analyze the changed processes. Previous authors have usually restricted their attention to average values. In the following discussion we derive new results concerning the probability distributions of various stochastic processes in the reorganized queueing system.

Shut-down control naturally divides the queueing process into alternating shut-down phases and busy periods, say S(N) and B(N), respectively. Let $T_{S(N)}$ and $T_{B(N)}$ be the corresponding lengths of S(N) and B(N). Since $T_{S(N)}$ is the sum of N independent, identically distributed exponential random variables, $T_{S(N)}$ has a gamma distribution with mean N/λ.

Suppose $b^*(s)=\int_0^\infty e^{-st}b(t)dt$, where $b(\cdot)$ is the probability density function associated with the random variable $T_{B(1)} \equiv T_B$. Cox & Smith(1961, pp.145-147) show that $b^*(s)$ satisfies the functional equation $b^*(s)=g^*\{s+\lambda-\lambda b^*(s)\}$, where $g^*(\cdot)$ is the Laplace transform of the service time probability density function $g(\cdot)$. Since $T_{B(N)}$ is the sum of N independent, identically distributed random variables with Laplace transform $b^*(s)$ it follows that

$$E\left\{e^{-sT_{B(N)}}\right\}=\left[g^*\left\{s+\lambda-\lambda b^*(s)\right\}\right]^N . \qquad (3.1.1)$$

For $\rho<1$, Cox and Smith have shown that $E(T_B)=g_1/(1-\rho)$; hence $E\{T_{B(N)}\}=\frac{Ng_1}{1-\rho}$, a result which Yadin & Naor(1963) obtain by using direct arguments with averages.

Example 3.1.1

When $g(x)=\mu e^{-\mu x}$ and $\rho \leq 1$, Cox & Smith(1961, p.148) have shown that

$$b^*(s)=\frac{1}{2\rho}\left[1+\rho+\frac{s}{\mu}-\left\{\left(1+\rho+\frac{s}{\mu}\right)^2-4\rho\right\}^{1/2}\right]$$

Therefore,

$$E\left\{e^{-sT_{B(N)}}\right\}=\left(\frac{1}{2\rho}\left[1+\rho+\frac{s}{\mu}-\left\{\left(1+\rho+\frac{s}{\mu}\right)^2-4\rho\right\}^{1/2}\right]\right)^N$$

It follows(Abramowitz & Stegun,1964,29.3.58) that the probability density of $T_{B(N)}$ is

$$\frac{N e^{-t(\mu+\lambda)}}{t \rho^{\frac{1}{2}N}} I_N\left(2\sqrt{\lambda\mu}\, t\right) \quad ,$$

where $I_N(t)$ is the Bessel function of imaginary argument and Nth order.

From the server's perspective, shut-down control links N idle periods to form one shut-down phase of average length N/λ. However, customers generally experience longer queues and increased queueing times; therefore, we consider the equilibrium queueing time distribution.

Since the queueing time process for an arbitrary customer, C, is generally non-Markov, we consider the Takács virtual queueing time process (sometimes called the waiting time process) [cf. Takács(1962, p.49)] . Initially, we assume that customers who arrive during a shut-down phase first contribute to the virtual queueing time process when the next busy period begins. This definition makes the virtual queueing time process Markov, and at any time t either the server is shut down, j customers are present ($j=0,\ldots,N-1$) and the virtual queueing time, $\eta(t)$, is zero, or the server is busy and $\eta(t)=x>0$.

Let $p_j(t)$=pr(server is shut down and j customers are queueing) ($t \geq 0$; $j=0,\ldots,N-1$) and $p(x,t)$=pr{server is busy and $\eta(t)=x$} ($x>0$; $t \geq 0$). Since arrivals during S(N) first contribute to $\eta(t)$ at the phase change epoch, the distribution function for $\eta(t)$ is discontinuous at $x=0$, and continuous for $x>0$. Time-dependent forward Kolmogorov differential equations for the probability distribution of $\eta(t)$ are given by

$$\left.\begin{aligned} &\frac{\partial}{\partial t} p_0(t) = -\lambda p_0(t) + p(0,t) \quad , \\ &\frac{\partial}{\partial t} p_j(t) = -\lambda p_j(t) + \lambda p_{j-1}(t) \quad , \quad (j=1,\ldots,N-1) \\ &\frac{\partial}{\partial t} p(x,t) = \frac{\partial}{\partial x} p(x,t) - \lambda p(x,t) + \lambda p_{N-1}(t) g_N(x) + \lambda \int_0^x p(x-u,t) g(u)\, du , \end{aligned}\right\} \quad (3.1.2)$$

where $g_N(x)$ is the N-fold convolution of $g(x)$.

When $\rho<1$, the equilibrium queueing time distribution for C coincides with the equilibrium probability distribution for the virtual queueing time. If we replace $p_j(t)$ and $p(x,t)$ by p_j and $p(x)$, then steady-state equations corresponding to (3.1.2) are

$$\lambda p_0 = p(0) \quad , \tag{3.1.3}$$

$$\lambda p_j = \lambda p_{j-1} \quad , \quad (j=1,\dots,N-1) \tag{3.1.4}$$

$$0 = p'(x) - \lambda p(x) + \lambda p_{N-1} g_N(x) + \lambda \int_0^x p(x-u) g(u) du. \tag{3.1.5}$$

In equilibrium, according to (3.1.3), the rate at which departing customers leave an empty queue behind equals the rate at which arriving customers find the system empty.

Let $p^*(s) = \int_{0+}^{\infty} e^{-sx} p(x) dx$ and take Laplace transforms in (3.1.5) with respect to x. Then

$$0 = s p^*(s) - p(0) - \lambda p^*(s) + \lambda p_{N-1} \{g^*(s)\}^N + \lambda p^*(s) g^*(s) . \tag{3.1.6}$$

Using (3.1.3) and (3.1.4) we obtain the expression

$$p^*(s) = \frac{p(0) \left[1 - \{g^*(s)\}^N \right]}{s - \lambda + \lambda g^*(s)} \quad . \tag{3.1.7}$$

Since $\lambda g_1 = \rho$, as $s \to 0+$ (3.1.7) becomes

$$p^*(0) = \frac{p(0) N g_1}{1-\rho} \quad .$$

Using (3.1.3), (3.1.4) and the normalizing condition $\sum_{j=0}^{N-1} p_j + \int_0^{\infty} p(x) dx = 1$

we obtain the equation $Np_0 + p^*(0) = 1$,

$$\text{i.e.} \qquad \frac{N p(0)}{\lambda} + \frac{N g_1 p(0)}{1-\rho} = 1 \quad .$$

Therefore

$$p^*(s) = \frac{\lambda(1-\rho) \left[1 - \{g^*(s)\}^N \right]}{N \{ s - \lambda + \lambda g^*(s) \}} \quad , \tag{3.1.8}$$

and $\lim_{s \to 0+} p^*(s) = \rho$ is the probability that C's queueing time is positive, i.e. that C arrives during B(N) [$C \varepsilon B(N)$] . Hence C arrives during S(N) [$C \varepsilon S(N)$] with probability $1-\rho$, as direct considerations indicate we should expect.

By definition, customers arriving during S(N) first contribute to the virtual queueing time process when B(N) begins. Therefore, if $W_q(N)$ is the equilibrium queueing time for (0,N) control

$$
\begin{aligned}
E\left\{e^{-sW_q(N)}\right\} &= \sum_{j=0}^{N-1} P_j + p^*(s) \\
&= 1-\rho + \rho\,\frac{(1-\rho)\,\dfrac{1-\{g^*(s)\}^N}{sNg_1}}{1-\rho\,\dfrac{1-g^*(s)}{sg_1}} \qquad (3.1.9)\\
&= (1-\rho)\,E\left\{e^{-sW_q(N)}\,\middle|\, C\in S(N)\right\} + \rho\, E\left\{e^{-sW_q(N)}\,\middle|\, C\in B(N)\right\}.
\end{aligned}
$$

It follows that the queueing time distribution for a busy period arrival is not affected by the previous definition of queueing time for shut-down phase arrivals. We now relax that definition and consider the conditional distribution of $W_q(N)$ for customers arriving during a shut-down phase.

Exactly N customers arrive during each S(N); therefore C is the jth arrival, C_j, with probability 1/N, i.e. $pr\{C=C_j\varepsilon S(N)\}=1/N$, $(j=1,\dots,N)$. If customers are served in order of arrival the jth arrival queues while N-j additional customers arrive and then while the j-1 customers preceding him in the queue are served. Therefore,

$$E\left\{e^{-sW_q(N)}\,\middle|\, C=C_j\in S(N)\right\} = \left(\frac{\lambda}{\lambda+s}\right)^{N-j}\left\{g^*(s)\right\}^{j-1}, \qquad (j=1,\dots,N).$$

Hence

$$
\begin{aligned}
E\left\{e^{-sW_q(N)}\,\middle|\, C\in S(N)\right\} &= \sum_{j=1}^{N} E\left\{e^{-sW_q(N)}\,\middle|\, C=C_j\in S(N)\right\} pr\left\{C=C_j\in S(N)\right\}\\
&= \frac{1}{N}\sum_{j=1}^{N}\left(\frac{\lambda}{\lambda+s}\right)^{N-j}\{g^*(s)\}^{j-1},
\end{aligned}
$$

i.e.

$$E\left\{e^{-sW_q(N)}\,\middle|\, C\in S(N)\right\} = \frac{1}{N}\,\frac{\lambda+s}{\lambda-(\lambda+s)g^*(s)}\left[\left(\frac{\lambda}{\lambda+s}\right)^N - \{g^*(s)\}^N\right]. \qquad (3.1.10)$$

It follows from (3.1.9) and (3.1.10) that

$$
\begin{aligned}
E\left\{e^{-sW_q(N)}\right\} &= \frac{1-\rho}{N}\,\frac{\lambda+s}{\lambda-(\lambda+s)g^*(s)}\left[\left(\frac{\lambda}{\lambda+s}\right)^N - \{g^*(s)\}^N\right]\\
&\quad + \rho\,\frac{(1-\rho)\,\dfrac{1-\{g^*(s)\}^N}{sNg_1}}{1-\rho\,\dfrac{1-g^*(s)}{sg_1}} \qquad (3.1.11)
\end{aligned}
$$

In the limit, as $s \to 0+$, the right-hand side of (3.1.11) tends to unity; hence (3.1.11) is the Laplace transform of a proper probability distribution. When N=1, (3.1.11) reduces to the Laplace transform of the equilibrium queueing time distribution in an M/G/1 queue without shut-down control [cf. (2.2.3)].

Using (3.1.11) we can show that

$$E\left\{W_q(N)\right\} = \frac{1}{2}\frac{\lambda g_2}{1-\rho} + \frac{1}{2}\frac{N-1}{\lambda} \quad . \qquad (3.1.12)$$

An identical expression for $E\{W_q(N)\}$ can be obtained by analyzing the total queueing time of all customers served during a shut-down phase and the subsequent busy period. Yadin & Naor(1963) obtain (3.1.12) by yet another argument.

In an M/G/1 queue without shut-down control the equilibrium mean queueing time is $\frac{1}{2}\lambda g_2/(1-\rho)$. It follows, from (3.1.12), that the increase in mean queueing time caused by shut-down control is $\frac{1}{2}(N-1)/\lambda$.

The form of (3.1.11) suggest an interpretation of the equilibrium queueing time process for busy period arrivals. Notice that

$$\begin{aligned} \frac{1-\{g^*(s)\}^k}{skg_1} &= \int_0^\infty \frac{\mathcal{G}_k(x)}{kg_1} e^{-sx} dx \\ &= h_k^*(s) \qquad , \qquad (k=1,2,\ldots) \end{aligned}$$

where $g_k(t)$ is the probability density function corresponding to $G_k(t)$, the k-fold convolution of $G(\cdot)$, $\mathcal{G}_k(x)=1-G_k(x)$ and $kg_1=\int_0^\infty tg_k(t)dt$. Then the equilibrium queueing time distribution for busy period arrivals has the Laplace transform

$$h_N^*(s) \frac{(1-\rho)s}{s-\lambda+\lambda g^*(s)} \quad . \qquad (3.1.13)$$

The function $h_k^*(s)$ (k=1,2,...) is the Laplace transform of the probability density function for the equilibrium forward (or backward) recurrence-time in a renewal process with interval probability density function $g_k(x)$. Notice, also, that $\frac{(1-\rho)s}{s-\lambda+\lambda g^*(s)}$ is the Laplace transform of the equilibrium queueing time distribution in an M/G/1 queue without shut-down control [cf. (2.2.3)]. Since (3.1.13) is the product of two Laplace transforms, the equilibrium queueing time for busy period arrivals has two independent, additive components. The first is a residual length of time related to the clearing of the initial N customers. The second component is the steady-state queueing time in an M/G/1 queue without shut-down control.

The probability distribution for $W_q(N)$ can, in principle, be obtained by inverting (3.1.11); this may be difficult in practice. However, the moments of $W_q(N)$ are easily recovered.

3.2 Linking shut-down control of the service process to virtual queueing time

One possible disadvantage of (0,N) control is that all shut-down phase customers are regarded alike. If service times are well dispersed about the mean value, several shut-down phase customers with long service times could cause considerable unnecessary queueing. Since queue size is sometimes only a rough measure of the workload in a queueing process, a more refined indicator of system workload might overcome this disadvantage.

This substitute indicator should account for the workload associated with each shut-down phase arrival. If the future service times of customers are known, or can be accurately estimated, Takács(1962, p.49) virtual queueing time process would be a suitable alternative. It is surprising, therefore, that the idea of linking shut-down control to the virtual queueing time has not appeared in the literature. To analyze a queueing process in which virtual queueing time determines server availability, we need a definition of virtual queueing time which applies both to the shut-down phase and to the busy period.

Definition

The <u>virtual queueing time</u>, $\eta(t)$, is the time which a customer arriving at time t would need to wait until his service began if customers were actually being served at t.

Thus, $\eta(t)$ is equal to the sum of the future service times, residual or otherwise, of all customers in the system at t. The obvious analogue of the (0,N) rule is (0,V) control, where V is a fixed, positive value. As each shut-down phase customer arrives, $\eta(t)$ increases by the equivalent of that customer's service time. When $\eta(t)$ $(t>0)$ first equals or exceeds V, a busy period begins. Subsequent shut-down phases begin whenever $\eta(t)$ is reduced to 0, i.e. whenever a departing customer leaves an empty system behind.

Much of the discussion in §3.1 applies to the analysis of (0,V) control. Generally, we will replace N by V in the notation, e.g. S(V) for S(N), etc., but the basic assumptions of §3.1 will not be changed.

Let N_I represent the number of customers who arrive during an interval I, and let S_i $(i=1,2,\ldots)$ be the future service time of the ith arrival. The S_i's are independent, identically distributed, non-negative random variables; therefore, standard renewal theory arguments [cf. Cox(1962, pp.36,45)] give

$$\text{(i)}\quad \text{pr}\{N_{S(V)}=k\}=G_{k-1}(V)-G_k(V) \qquad (k=1,2,\ldots), \quad (3.2.1)$$

$$\text{(ii)}\quad E\{N_{S(V)}\}=1+H(V), \qquad (3.2.2)$$

where $G_0(\cdot)\equiv 1$ and $H(\cdot)=\sum_{i=1}^{\infty}G_i(\cdot)$ is the renewal function defined by Cox(1962, p.45).

If $g^*(s)$ is a rational function of s, an explicit expression for (3.2.2) can be obtained by first inverting the Laplace transform $\dfrac{1}{s\{1-g^*(s)\}}$ and then evaluating the resulting real-valued function at V.

There is a simple explanation for (3.2.2). Although the renewal function, H(t), specifies the expected number of renewals in the interval (0,t) given that a renewal occurred at t=0, H(t) does not include the event at the origin. In (0,V) control the analogue of a renewal at time zero is the arrival of the first customer during a shut-down phase. In determining $E\{N_{S(V)}\}$ this customer must also be counted; therefore, when a busy period begins the queue contains, on average, $1+H(V)$ customers.

Using elementary conditional arguments we can show that

$$E\left\{e^{-sT_{S(V)}}\right\} = \sum_{n=1}^{\infty} pr\left\{N_{S(V)}=n\right\} E\left\{e^{-sT_{S(V)}} \mid N_{S(V)}=n\right\}$$

$$= \sum_{n=1}^{\infty} \left\{G_{n-1}(V) - G_n(V)\right\} \left(\frac{\lambda}{\lambda+s}\right)^n$$

$$= \frac{\lambda}{\lambda+s} - \frac{s}{\lambda+s}\left\{\sum_{n=1}^{\infty} G_n(V)\left(\frac{\lambda}{\lambda+s}\right)^n\right\} , \qquad (3.2.3)$$

$$E\left\{T_{S(V)}\right\} = \frac{1}{\lambda} E\left\{N_{S(V)}\right\} = \frac{1+H(V)}{\lambda} . \qquad (3.2.4)$$

To obtain an expression for the Laplace transform of $T_{B(V)}$ we use results derived by Cox & Miller(1965, pp.244-246) for the busy period in the Takács process. Thus

$$E\left[e^{-sT_{B(V)}} \mid \eta\left\{S(V)\right\} = \eta_0\right] = e^{-\omega(s)\eta_0} , \qquad (3.2.5)$$

where $\eta\{S(V)\}=\eta_0$ is the value of $\eta(t)$ when $B(V)$ begins, $\eta_0 \geqq V$. The function $\omega(s)$ is that particular root of the equation $s=\omega-\lambda+\lambda g^*(\omega)$ which is positive when $Re(s)>0$ and which satisfies $\omega(0)=0$.

Integrate (3.2.5) with respect to the distribution of η_0; then,

$$E\left\{e^{-sT_{B(V)}}\right\} = E\left\{e^{-\omega(s)\eta_0}\right\} . \qquad (3.2.6)$$

Other results in Cox & Miller(1965, pp.51-55) may be used to show that

$$E\left\{e^{-\theta\eta_0} z^{N_{S(V)}}\right\} = 1 - \left\{1 - z g^*(\theta)\right\} K(\theta,z) , \qquad (3.2.7)$$

where $K(\theta,z)=\sum_{n=0}^{\infty} z^n\{\int_0^V e^{-\theta x} g_n(x)dx\}$ for $Re(\theta)>0$ ($|z|\leqq 1$). By evaluating the limit of (3.2.7) as $z\to 1-$ we can express $E(e^{-\theta\eta_0})$ in terms of $g^*(\theta)$ and $K(\theta,1)$. Thus (3.2.6) becomes

$$E\left\{e^{-sT_{B(V)}}\right\} = 1 - \left[1 - g^*\left\{\omega(s)\right\}\right]\left[\sum_{n=0}^{\infty}\left\{\int_0^V e^{-\omega(s)x} g_n(x)dx\right\}\right]. \qquad (3.2.8)$$

An expression for $E\{T_{B(V)}\}$ can be obtained from (3.2.8). However, it follows from the results of §3.1 that

$$E\left\{T_{B(V)}\right\} = \frac{g_1}{1-\rho} E\left\{N_{S(V)}\right\} = \frac{g_1+g_1 H(V)}{1-\rho} . \qquad (3.2.9)$$

Example 3.2.1

If $g(x)=\mu e^{-\mu x}$, then $g_n(x)=\frac{\mu(\mu x)^{n-1}}{\Gamma(n)}e^{-\mu x}$ and $H(t)=\mu t$ $(t\geqq 0)$. After simplification, (3.2.3) becomes

$$E\left\{e^{-sT_{S(V)}}\right\}=\frac{\lambda}{\lambda+s}\,e^{-\frac{s\mu V}{\lambda+s}}.$$

Therefore (Abramowitz & Stegun, 1964, 29.3.81) the probability density function of $T_{S(V)}$ is $\lambda e^{-(\mu V+\lambda t)}I_0(2\sqrt{\lambda\mu Vt})$ $(t\geqq 0)$, where $I_0(z)$ is the Bessel function of imaginary argument and order 0.

Similarly, (3.2.8) becomes

$$E\left\{e^{-sT_{B(V)}}\right\}=\frac{\mu}{\mu+\omega(s)}\,e^{-\omega(s)V},$$

where $\omega(s)=\frac{1}{2}[s+\lambda-\mu+\{(s+\lambda-\mu)^2+4\mu s\}^{\frac{1}{2}}]$ and we take the modulus of the square root.

Since $H(x)=\mu x$, (3.2.4) and (3.2.9) become

$$E\left\{T_{S(V)}\right\}=\frac{1+\mu V}{\lambda},\qquad E\left\{T_{B(V)}\right\}=\frac{1+\mu V}{\mu(1-\rho)}.$$

To evaluate the queueing time distribution of an arbitrary customer we condition on $N_{S(V)}$ and use the results of §3.1.

If $N_{S(V)}=n$, (3.1.11) gives

$$E\left\{e^{-sW_q(V)}\,\middle|\,N_{S(V)}=n\right\}=\frac{1-\rho}{n}\,\frac{\lambda+s}{\lambda-(\lambda+s)g^*(s)}\left[\left(\frac{\lambda}{\lambda+s}\right)^n-\left\{g^*(s)\right\}^n\right]+\rho\,\frac{(1-\rho)\dfrac{1-\{g^*(s)\}^n}{sng_1}}{1-\rho\,\dfrac{1-g^*(s)}{sg_1}}.\tag{3.2.10}$$

It follows from (3.2.1) and (3.2.10) that

$$E\left\{e^{-sW_q(V)}\right\}=\sum_{n=1}^{\infty}E\left\{e^{-sW_q(V)}\,\middle|\,N_{S(V)}=n\right\}\left\{G_{n-1}(V)-G_n(V)\right\}.\tag{3.2.11}$$

This transform cannot be simplified without first specifying the service time distribution. However, by using (3.1.13) and (3.2.2) the simple formula

$$E\left\{W_q(V)\right\}=\frac{1}{2}\,\frac{\lambda g_2}{1-\rho}+\frac{1}{2}\,\frac{H(V)}{\lambda},\tag{3.2.12}$$

can be obtained; obviously, $\frac{1}{2}\frac{H(V)}{\lambda}$ is the increase in queueing time caused by (0,V) control.

Example 3.2.2

Let $g(x)=\mu e^{-\mu x}$. Then $G_{n-1}(V)-G_n(V)=\frac{(\mu V)^{n-1}}{\Gamma(n)}e^{-\mu V}$ $(n=1,2,\ldots)$ and $g^*(s)=\frac{\mu}{\mu+s}$. Substituting in (3.2.11) gives

$$E\left\{e^{-sW_q(V)}\right\}=\frac{1}{s\mu^2 V}\left\{\lambda\mu-(\lambda+s)(\mu+s)e^{-\frac{s\mu V}{\lambda+s}}\right\}+\frac{1}{\mu^2 V(s+\mu-\lambda)}\left\{(\mu+s)^2 e^{-\frac{s\mu V}{\lambda+s}}-\lambda\right\}$$

Preliminary attempts to invert this transform suggest that the probability distribution of $W_q(V)$ can be expressed as the sum of a constant factor, an exponential term and a complicated linear combination of several Bessel functions of imaginary argument. The mean queueing time, $E\{W_q(V)\}$, is equal to $\frac{\rho}{\mu(1-\rho)}+\frac{1}{2}\frac{V}{\rho}$.

The preceding discussion is based on the assumption that server availability should be linked to virtual queueing time instead of queue size. It is quite possible that in a more refined version of shut-down control server availability would depend on both queue size and virtual queueing time. The optimal combination has not been obtained.

3.3 The effect of shut-down control on a queueing process

Users of shut-down control will probably be interested in quantitatively assessing its two major effects. The service process is reorganized to create some longer idle periods. For (0,N) control, $T_{S(N)}$, the length of the shut-down phase, has a gamma distribution with mean N/λ and coefficient of variation $N^{-\frac{1}{2}}$. The distribution of $T_{S(V)}$, the length of the (0,V) shut-down phase, depends both on V and on the service time distribution. If $T_{S(N)}$ and $T_{S(V)}$ are made to have the same mean value, the distribution of $T_{S(V)}$ is more dispersed than that of $T_{S(N)}$.

An effect of equal importance in shut-down control is the general increase in queueing times; every customer must queue because the server is never idle. Without shut-down control a proportion, $1-\rho$, of all customers avoid queueing altogether. If shut-down control queueing times are too long, the overall functioning of the system may be impaired. Therefore, we need to investigate the shut-down control queueing time distribution.

Laplace transforms of the queueing time distribution for (0,N) and (0,V) control are given by (3.1.11) and (3.2.11), respectively. To evaluate these expressions, a service time distribution must be specified. In a similar situation in Chapter 2 we considered three service time distributions, D_1, D_2 and D_3 (cf. §§2.1,2.2). These were chosen as simple examples of service time distributions with coefficients of variation less than, equal to, and greater than unity, respectively. For each D_i (i=1, 2,3), the probabilities of various queueing times have been calculated for different values of the shut-down control parameters. These probabilities will be used to determine more precisely how shut-down control affects queueing times.

Two queueing systems will be said to correspond if their respective arrival processes and service time distributions are identical; we suppose that the time scale for each system is the same. In the remainder of §3.3 we compare the queueing time distribution in an M/G/1 queue without shut-down control to queueing time distributions in corresponding shut-down control queues. Comparisons for (0,N) and (0,V) control are considered separately.

Customers in a shut-down queue are likely to observe that queueing times are longer than necessary because the server is not available at all times. Therefore, from the customer's perspective, it is important to compare the probability of queueing longer than a fixed length of time in a queue without shut-down control to the probability of queueing longer than the same fixed length of time in a corresponding queue with shut-down control. Consequently, for the queue without shut-down control specified by D_i and ρ we have calculated $P_k(\rho)$, the probability of queueing longer than k times $E(W_q)$, where $E(W_q)$ is the mean queueing time in that particular queue; for the corresponding (0,j) shut-down control queue we have evaluated $P_k(j,\rho)$, the probability of queueing longer than the same fixed time, k times $E(W_q)$, $(j=N,V)$.

3.3.1 The effect of (0,N) control on queueing time

The Laplace transform of the queueing time distribution for (0,N) control is given by (3.1.11). The transform is a linear combination of two separate transforms; each separate transform corresponds to a probability distribution. Therefore, by separately inverting two individual transforms for each D_i, exact probabilities can be calculated. Details of the transform inversions are unimportant. In the worst possible case, D_3, the probability distribution for $W_q(N)$ is a linear combination of gamma distributions; hence, the required probabilities were evaluated numerically using the algorithm referred to in §2.2. For each service time distribution, the unconditional queueing time distribution for a queue without shut-down control can be derived from the conditional queueing time distributions given in (2.2.4).

Calculations were carried out for three values of N (2,5,8), three values of k (1,2,3) and 10 values of ρ (0.05,...,0.95) for each D_i. The calculated probabilities, $P_k(N,\rho)$, are arranged in Table 3.3.1. Due to rounding errors the results for $\rho=0.95$ are probably larger than the exact probabilities. Some of the probabilities are also plotted in Figs. 3.3.1 a, b and c.

A column of the same table also gives the mean queueing time, $E(W_q)$, for corresponding queues without shut-down control; for a given D_i and ρ, k times $E(W_q)$ is the fixed length of queueing time used in calculating both $P_k(\rho)$ and $P_k(N,\rho)$. If required, mean queueing times in corresponding shut-down queues can be obtained by adding $(N-1)/\rho$ to the given values of $E(W_q)$.

For fixed values of k and ρ the probabilities, $P_k(\rho)$, are negligibly different for the three queues without shut-down control. Therefore,

ρ	$E(W_q)$			N	$P_k(N,\rho)$ k=1			k=2			k=3		
	D_1	D_2	D_3		D_1	D_2	D_3	D_1	D_2	D_3	D_1	D_2	D_3
				2	0.997	0.973	0.957	0.991	0.948	0.919	0.983	0.923	0.884
0.05	0.079	0.105	0.132	5	0.999	0.999	0.999	0.999	0.999	0.998	0.999	0.998	0.998
				8	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999
				2	0.971	0.909	0.867	0.924	0.829	0.766	0.869	0.759	0.688
0.15	0.265	0.353	0.441	5	0.997	0.995	0.994	0.993	0.990	0.987	0.989	0.985	0.978
				8	0.998	0.997	0.996	0.996	0.994	0.993	0.994	0.991	0.989
				2	0.919	0.832	0.772	0.807	0.697	0.622	0.695	0.587	0.516
0.25	0.500	0.667	0.833	5	0.990	0.987	0.982	0.980	0.971	0.959	0.968	0.953	0.930
				8	0.994	0.992	0.990	0.988	0.983	0.979	0.981	0.974	0.967
				2	0.845	0.745	0.675	0.661	0.558	0.486	0.506	0.421	0.362
0.35	0.808	1.08	1.35	5	0.980	0.973	0.963	0.957	0.937	0.908	0.930	0.891	0.841
				8	0.988	0.983	0.979	0.974	0.965	0.955	0.959	0.944	0.927
				2	0.752	0.653	0.580	0.508	0.426	0.368	0.337	0.279	0.240
0.45	1.23	1.64	2.05	5	0.965	0.951	0.931	0.920	0.876	0.822	0.861	0.778	0.694
				8	0.978	0.970	0.963	0.952	0.935	0.915	0.923	0.894	0.858
				2	0.646	0.562	0.495	0.369	0.317	0.279	0.213	0.179	0.163
0.55	1.83	2.44	3.06	5	0.942	0.914	0.877	0.852	0.768	0.682	0.723	0.590	0.486
				8	0.964	0.952	0.937	0.917	0.886	0.848	0.862	0.803	0.731
				2	0.536	0.483	0.428	0.269	0.241	0.228	0.144	0.122	0.126
0.65	2.79	3.71	4.64	5	0.902	0.850	0.786	0.716	0.589	0.487	0.480	0.347	0.270
				8	0.942	0.921	0.896	0.856	0.795	0.720	0.746	0.623	0.507
				2	0.436	0.427	0.391	0.228	0.197	0.211	0.112	0.092	0.106
0.75	4.50	6.00	7.50	5	0.817	0.728	0.634	0.466	0.365	0.306	0.226	0.164	0.152
				8	0.902	0.863	0.810	0.725	0.600	0.486	0.475	0.325	0.243
				2	0.414	0.396	0.404	0.215	0.169	0.206	0.084	0.072	0.084
0.85	8.50	11.3	14.2	5	0.586	0.538	0.463	0.275	0.222	0.240	0.127	0.094	0.106
				8	0.793	0.708	0.597	0.388	0.306	0.278	0.181	0.126	0.132
				2	0.476	0.377	0.470	0.189	0.146	0.186	0.056	0.056	0.057
0.95	28.5	38.0	47.5	5	0.486	0.408	0.477	0.211	0.158	0.200	0.065	0.061	0.062
				8	0.485	0.423	0.481	0.236	0.171	0.214	0.077	0.066	0.068

Table 3.3.1 Probabilities, $P_k(N,\rho)$, that (0,N) shut-down queueing times exceed k times $E(W_q)$, mean queueing time in a non-shut-down queue with identical traffic intensity, ρ, and D_1(Erlang), D_2(exponential) or D_3(mixed exponential) service times. Shut-down phases end when N customers are waiting. The entries in this table are approximations calculated

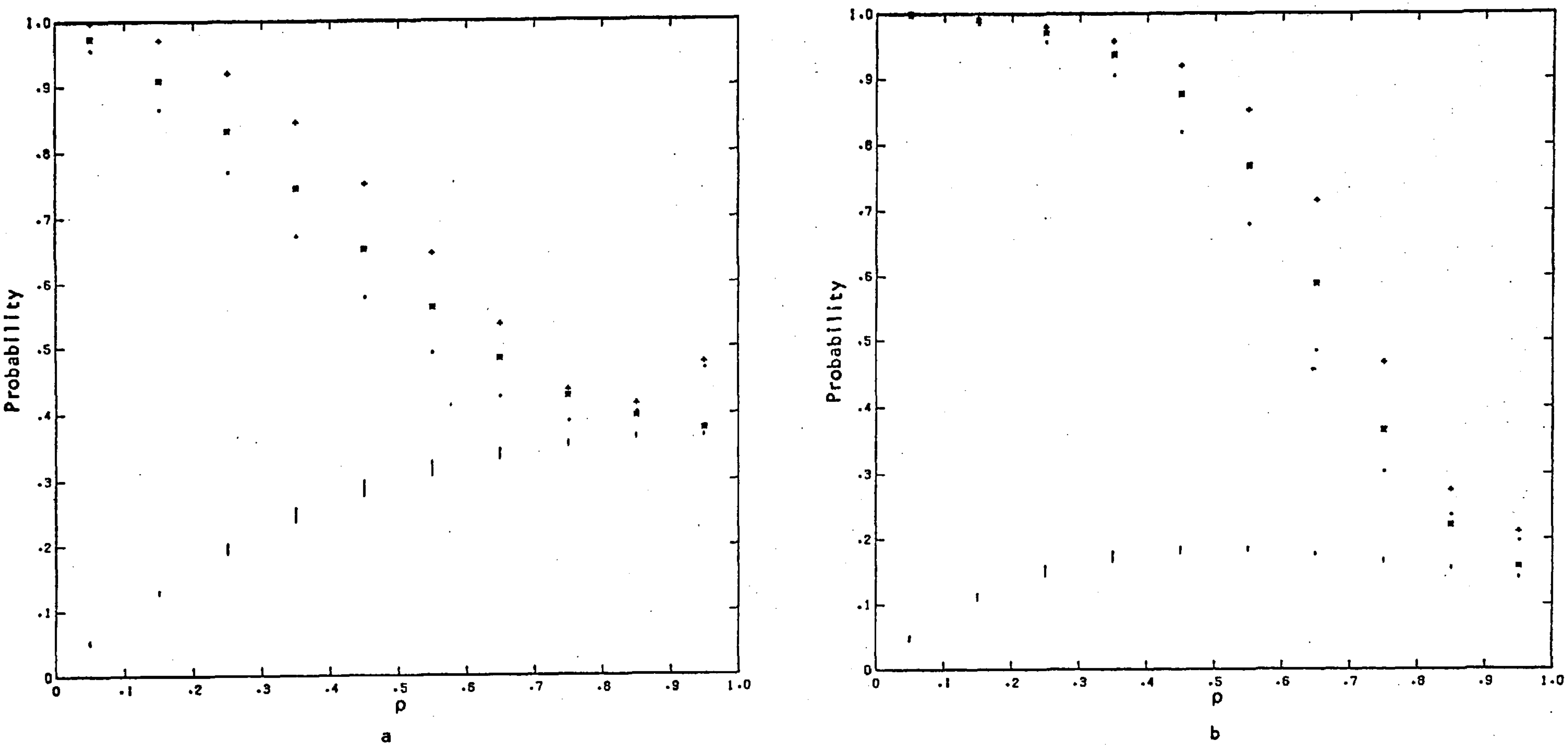


Fig. 3.3.1 The probabilities that (0,N) shut-down and non-shut-down queueing times individually exceed k times the mean queueing time in the non-shut-down queue when service times are D_1, D_2 or D_3 and the traffic intensity is ρ. Shut-down phases end when N customers are waiting. In (a) k=1 and N=2; In (b) k=2 and N=5.

\+ D_1 (Erlang) ■ D_2 (exponential) • D_3 (mixed exponential) | The range of probabilities for all three non-shut-down queues

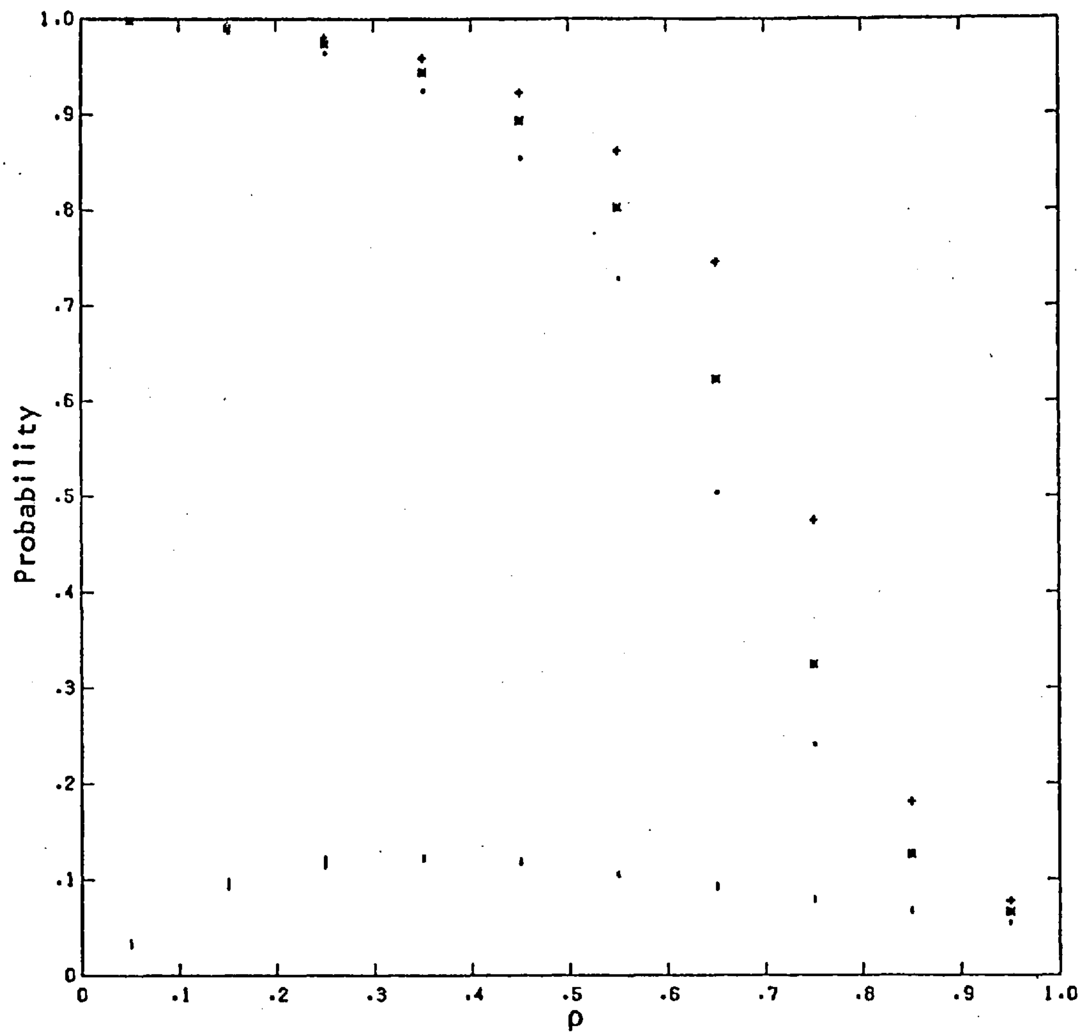


Fig. 3.3.1 c The probabilities that (0,N=8) shut-down and non-shut-down queueing times individually exceed three times the mean queueing time in the non-shut-down queue when service times are D_1, D_2 or D_3 and the traffic intensity is ρ. Shut-down phases end when 8 customers are waiting.

+ D_1 (Erlang) ■ D_2 (exponential) • D_3 (mixed exponential)

| The range of probabilities that queueing time in any of the three non-shut-down queues individually exceeds three times the mean queueing time.

these probabilities have been plotted in Figs. 3.3.1 a, b and c as the symbols, $|$, to indicate the range of $P_k(\rho)$ for the distributions D_1, D_2 and D_3.

Table 3.3.1 indicates several important aspects of the effect of (0,N) control on queueing time. For fixed ρ and D_i, $P_k(N,\rho)$, the probability of queueing longer than k times $E(W_q)$, increases with N. Since $E(W_q)$, mean queueing time in the corresponding queue without shut-down control, does not depend on N, the increase in $P_k(N,\rho)$ is an obvious consequence of requiring a greater number of customers to arrive before beginning a busy period. However, for fixed N and D_i, the table shows that $P_k(N,\rho)$ is decreasing in ρ. The reason for this decrease is discussed in §3.3.2. A consequence of this decreasing probability, however, is that the extra queueing time which shut-down control causes will be less additional inconvenience to customers if traffic conditions are already heavy.

Table 3.3.1 also shows that the probabilities $P_k(N,\rho)$ are generally smallest for the most variable service time distribution, D_3. This difference between the D_i's arises because $P_k(N,\rho)$ is defined as the probability of shut-down queueing times which are long relative to mean queueing time, $E(W_q)$, in the corresponding queue without shut-down control. By calculating, for fixed N and ρ, the probability of shut-down queueing times which are long relative to $E\{W_q(N)\}$, mean queueing time for the same shut-down queue, it was found that the shut-down control queueing time distributions determined by D_1, D_2 and D_3 do not differ significantly. The same calculations, which do not appear here, show that as N increases, so does the probability of queueing times at least as long as $E\{W_q(N)\}$; at the same time, however, shut-down control queueing times exceeding two and three times $E\{W_q(N)\}$ become rather less probable.

3.3.2 The effect of (0,V) control on queueing time

The Laplace transform of the queueing time distribution for (0,V) control is given by (3.2.10) and (3.2.11). Though (3.2.11) is a weighted sum of Laplace transforms, the weights depend on the distribution of the conditioning variable $N_{S(V)}$ and are independent of the transform argument. Thus, to determine the probability of a queueing time event, it is sufficient to evaluate the required probability for (0,N) control where N=1,2,... and then calculate a weighted sum of probabilities, truncating the sum when convergence is adequate. For D_1, D_2 and D_3, general expressions for the weights are, at worst, a linear combination of gamma distributions integrated on the interval $[0,V]$; these can be evaluated without difficulty.

Since the customer's perspective is equally important in both versions of shut-down control, we have calculated $P_k(V,\rho)$, the probability that queueing time in a (0,V) shut-down queue exceeds k times $E(W_q)$, mean queueing time in the corresponding queue without shut-down control. For each D_i, calculations were carried out for three values of ρ (0.25,0.55,0.85) and a range of values of V. The three values of ρ were chosen to represent light, moderate and heavy traffic conditions. Table 3.3.2 shows the calculated probabilities, approximated using gamma distributions. Due to rounding errors, the values of $P_k(V,\rho)$ for D_1 service times when $V \geq 14.0$ and $\rho=0.25$ are probably smaller than the exact probabilities.

The last row of Table 3.3.2 gives the mean values, $E(W_q)$, for corresponding queues without shut-down control; given D_i and ρ, k times $E(W_q)$ is the fixed length of queueing time used in calculating $P_k(\rho)$ and $P_k(V,\rho)$.

Some probabilities from Table 3.3.2 are plotted in Figs. 3.3.2 a, b and c. For fixed k and ρ, the band of probability between horizontal lines on each graph represents the range of $P_k(\rho)$ for D_1, D_2 and D_3.

Table 3.3.2 and Figs. 3.3.2 a, b and c show that unless ρ is reasonably large and V is quite small (two or three times the mean service time), shut-

V	k	ρ=0.25 D_1	D_2	D_3	ρ=0.55 D_1	D_2	D_3	ρ=0.85 D_1	D_2	D_3
	1	0.637	0.633	0.642	0.544	0.535	0.517	0.416	0.400	0.411
2.0	2	0.565	0.555	0.556	0.332	0.331	0.323	0.210	0.170	0.209
	3	0.495	0.488	0.487	0.200	0.200	0.199	0.081	0.073	0.085
	1	0.890	0.830	0.809	0.750	0.692	0.655	0.446	0.442	0.431
4.0	2	0.834	0.765	0.736	0.530	0.479	0.447	0.230	0.187	0.221
	3	0.775	0.704	0.673	0.356	0.318	0.296	0.094	0.080	0.093
	1	0.963	0.918	0.888	0.860	0.794	0.747	0.500	0.491	0.458
6.0	2	0.932	0.872	0.833	0.686	0.603	0.551	0.251	0.206	0.232
	3	0.897	0.826	0.782	0.515	0.437	0.391	0.108	0.088	0.101
	1	0.983	0.958	0.932	0.914	0.858	0.811	0.569	0.542	0.491
8.0	2	0.966	0.926	0.890	0.788	0.698	0.635	0.275	0.229	0.246
	3	0.945	0.894	0.850	0.644	0.542	0.479	0.124	0.097	0.109
	1	0.990	0.976	0.957	0.939	0.898	0.856	0.640	0.593	0.528
10.0	2	0.978	0.954	0.926	0.848	0.767	0.703	0.305	0.256	0.260
	3	0.965	0.931	0.895	0.736	0.629	0.555	0.141	0.107	0.118
	1	0.992	0.984	0.971	0.952	0.922	0.887	0.704	0.641	0.565
12.0	2	0.983	0.969	0.948	0.883	0.817	0.756	0.343	0.286	0.277
	3	0.973	0.952	0.923	0.796	0.697	0.621	0.160	0.120	0.128
	1	0.991	0.989	0.979	0.960	0.938	0.909	0.755	0.684	0.603
14.0	2	0.983	0.977	0.961	0.904	0.851	0.796	0.388	0.320	0.297
	3	0.976	0.964	0.942	0.835	0.750	0.675	0.182	0.134	0.138
	1	0.981	0.990	0.983	0.965	0.947	0.924	0.794	0.721	0.638
16.0	2	0.975	0.981	0.968	0.918	0.875	0.897	0.438	0.356	0.318
	3	0.968	0.970	0.953	0.861	0.789	0.719	0.207	0.151	0.148
	1	0.950	0.990	0.982	0.967	0.953	0.933	0.822	0.752	0.670
18.0	2	0.945	0.982	0.970	0.926	0.891	0.849	0.489	0.394	0.340
	3	0.939	0.973	0.957	0.877	0.817	0.754	0.236	0.170	0.160
	1	0.884	0.986	0.976	0.952	0.953	0.937	0.827	0.777	0.698
20.0	2	0.879	0.979	0.966	0.916	0.900	0.863	0.525	0.430	0.364
	3	0.874	0.972	0.955	0.873	0.834	0.779	0.261	0.191	0.172
$E(W_q)$		0.500	0.667	0.833	1.83	2.44	3.06	8.50	11.3	14.2

Table 3.3.2 Probabilities, $P_k(V,\rho)$, that (0,V) shut-down queueing times exceed k times $E(W_q)$, mean queueing time in a non-shut-down queue with identical traffic intensity, ρ, and D_1(Erlang), D_2(exponential) or D_3(mixed exponential) service times. Shut-down phases end when the virtual queueing time $\geq$V.

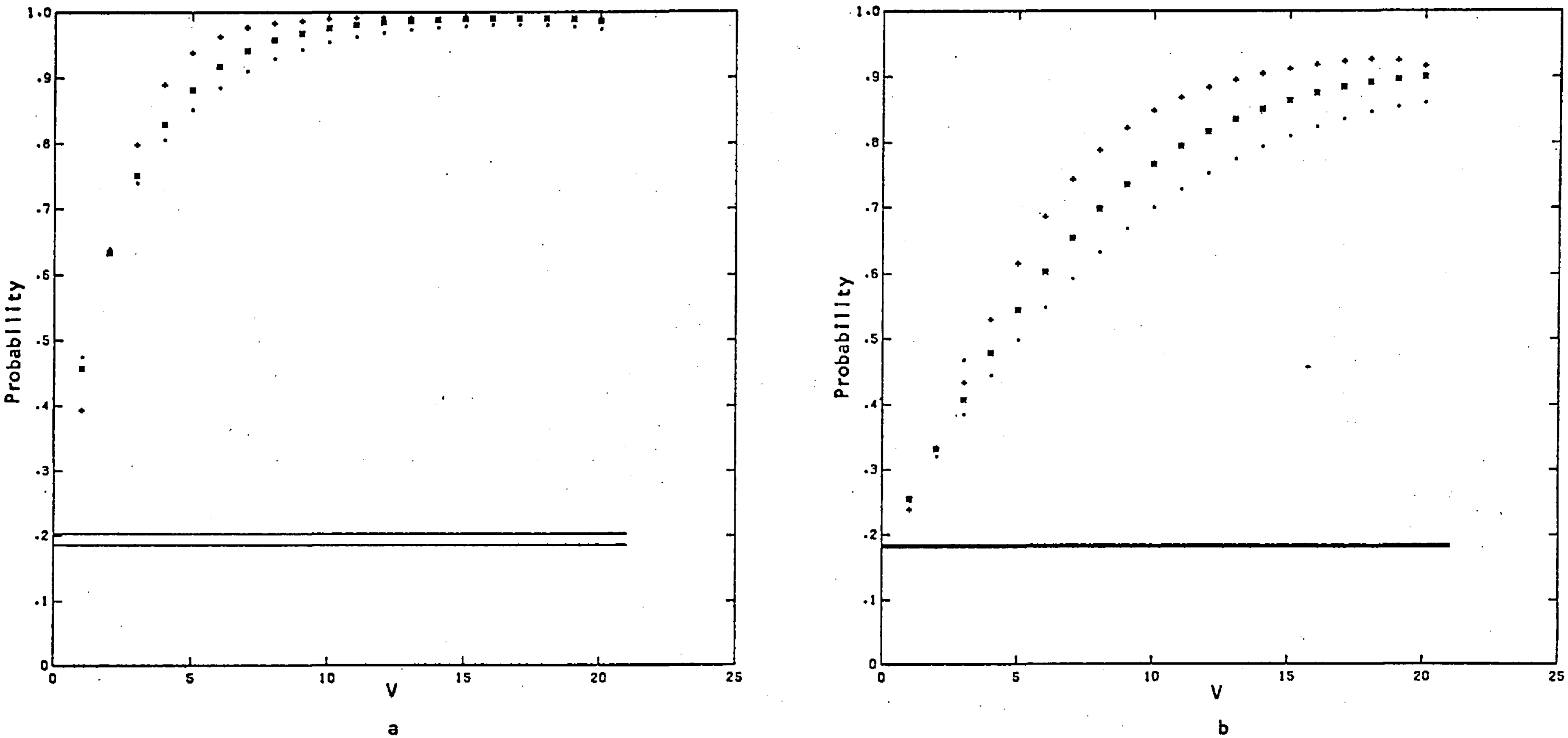


Fig. 3.3.2 The probabilities that (0,V) shut-down and non-shut-down queueing times individually exceed k times the mean queueing time in the non-shut-down queue when service times are D_1, D_2 or D_3 and the traffic intensity is ρ. Shut-down phases end when the virtual queueing time $\geq$ V. In (a) k=1 and ρ=0.25; In (b) k=2 and ρ=0.55.

+ D_1 (Erlang) ■ D_2 (exponential) • D_3 (mixed exponential) = The range of probabilities for all three non-shut-down queues.

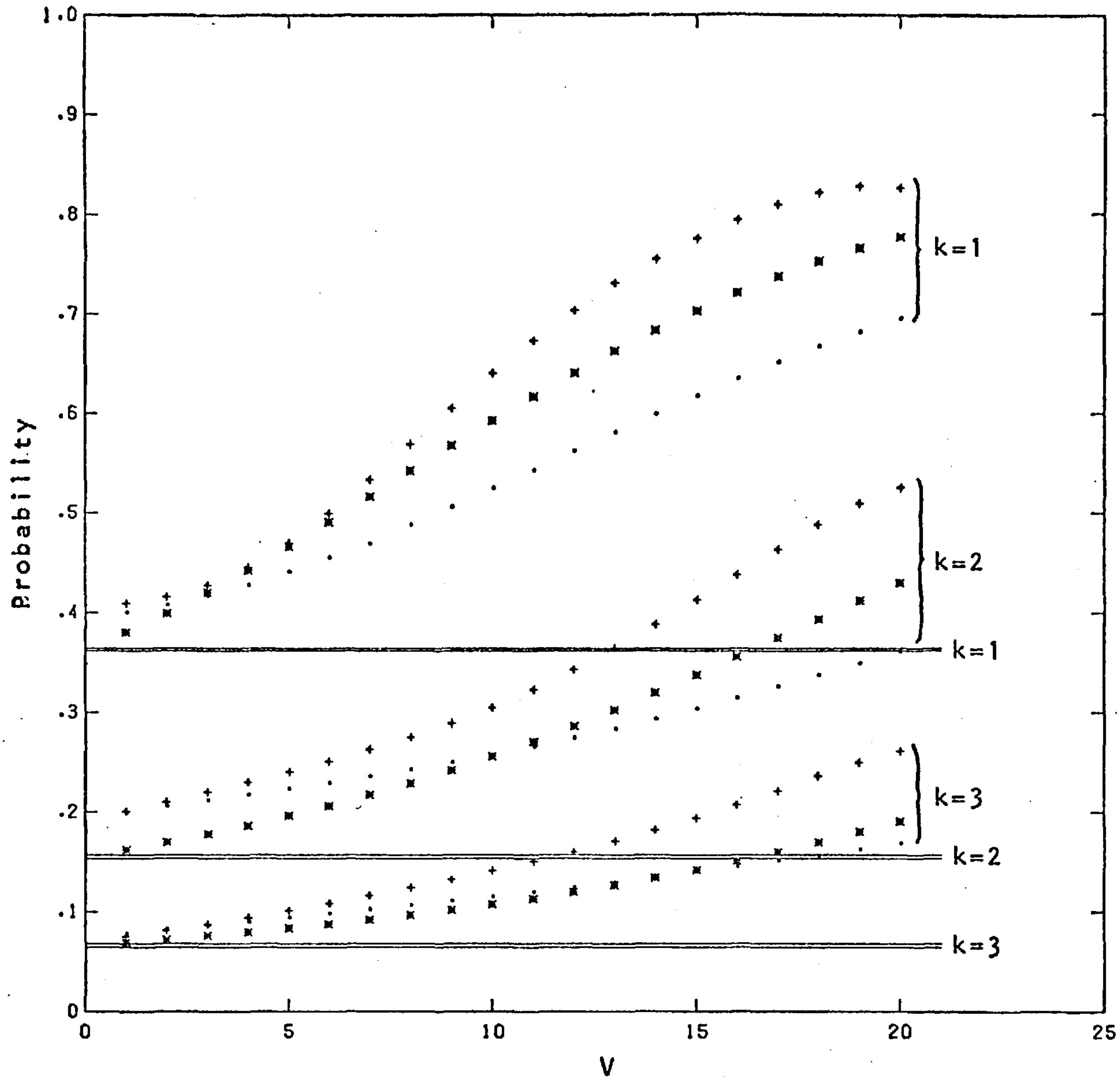


Fig. 3.3.2 c The probabilities that (0,V) shut-down and non-shut-down queueing times individually exceed k times the mean queueing time in the non-shut-down queue when service times are D_1, D_2 or D_3 and the traffic intensity ρ=0.85. Shut-down phases end when the virtual queueing time ≥ V.

+ D_1 (Erlang) ■ D_2 (exponential) D_3 (mixed exponential)

══ k= The range of probabilities that queueing times in all three non-shut-down queues exceed k times the mean queueing time (k=1,2,3).

down queueing time probabilities, $P_k(V,\rho)$, are often at least twice as large as $P_k(\rho)$, the probability of queueing for the same fixed length of time in a corresponding non-shut-down queue. When V is an order of magnitude larger than the mean service time, shut-down queueing times which are long relative to $E(W_q)$ are almost certain unless traffic conditions are already heavy, e.g. $\rho=0.85$.

It was noted in §3.3.1 that for fixed k, N and D_i, $P_k(N,\rho)$ is decreasing in ρ. Table 3.3.2 shows that when k, V and D_i are fixed, $P_k(V,\rho)$ is also a decreasing function of ρ. An explanation for this dependence is suggested below.

Since D_1, D_2 and D_3 have the same mean, changes in the traffic intensity are obtained by adjusting the rate of the Poisson arrival process. The conditional mean and variance of the queueing time distribution for shut-down phase arrivals, $E\{W_q(N)|C\epsilon S(N)\}$ and $Var\{W_q(N)|C\epsilon S(N)\}$, are both monotonic decreasing functions of λ. Conversely, $E(W_q)$, the mean queueing time in a corresponding queue without shut-down control, is an increasing function of λ. Therefore, the conditional probability that shut-down phase arrivals queue for longer than k times $E(W_q)$ decreases as λ increases. This effect is accentuated by the weighting factors which determine the unconditional probability $P_k(N,\rho)$. When ρ is small, the factor $1-\rho$ emphasizes the shut-down phase arrivals queueing time distribution; when ρ is fairly large, the weighting factor, ρ, emphasizes the contribution to $P_k(N,\rho)$ of the queueing time distribution for busy period arrivals. The mean of this latter conditional distribution is an increasing function of λ which always exceeds $E(W_q)$, but as $\rho\to 1$ the difference between the conditional mean and $E(W_q)$ becomes small relative to $E(W_q)$ for sensible values of N. Therefore, we conclude that $pr\{W_q(N) > kE(W_q)|C\epsilon B(N)\}$ is also a decreasing function of λ. Hence, as $\rho=2\lambda$ increases, the probability $P_k(N,\rho)$ decreases. Obviously, since $P_k(V,\rho)$ is a linear combination of $P_k(N,\rho)$'s and the weighting factors depend only on V and the service time distribution, $P_k(V,\rho)$, the probability of (0,V) queueing times which are long relative to $E(W_q)$, will be a decreasing function of ρ as well. Thus, the extra queueing time which either (0,N) or (0,V) control causes will be less additional inconvenience to customers if traffic conditions are already fairly heavy.

3.3.3 Choosing between (0,N) and (0,V) control

In deciding between (0,N) and (0,V) control in a given situation, the choice to be made would probably depend on several factors, but particularly on the pairs of values of N and V which define alternative versions of shut-down control. If we make the mean length of the shut-down phase the same for both (0,N) and (0,V) control then, for a given service time distribution with renewal function $H(\cdot)$, (0,i) and $(0,V_i)$ are alternative versions of shut-down control, where $N=i$, $V=V_i$ is a solution of the equation $N=1+H(V)$ $(i=2,3,\ldots)$.

Cox(1962, p.47) shows that $H(t)=\frac{t}{g_1}+\frac{g_2}{2g_1^2}-1+o(1)$. Thus, for each service time distribution D_i, we can approximate $1+H(V)$ by $\frac{1}{2}V+\frac{1}{8}g_2$; therefore,

$$V= 2N -\tfrac{1}{4}g_2 \qquad (3.3.1)$$

defines a set of alternative versions of shut-down control, $\{(0,n),(0,V_n)\}$ where $N=n$, $V=V_n$ satisfies (3.3.1) $(n=2,3,\ldots)$. The probabilities $P_k(n,\rho)$ and $P_k(V_n,\rho)$ $(n=2,3,\ldots)$ which were calculated in §§3.3.1 and (3.3.2) can then be used to choose between (0,N) and (0,V) control in a number of different queueing situations.

Table 3.3.3 gives values of $P_1(N,\rho)$ and $P_1(V,\rho)$ when V is defined by (3.3.1) for seven values of N $(2,\ldots,8)$ and the service time distributions D_1, D_2 and D_3; the three values of ρ (0.25,0.55,0.85) represent light, moderate and heavy traffic conditions, respectively. We suppose that, for the given set of alternative versions of shut-down control, the version with uniformly smaller values of $P_1(\cdot,\rho)$ for a given D_i and ρ is the better

$P_1(N,\rho)$			N		V	$P_1(V,\rho)$		
ρ=.25	ρ=.55	ρ=.85				ρ=.25	ρ=.55	ρ=.85
0.919	0.646	0.414		D_1	2.5	0.718	0.602	0.422
0.832	0.562	0.396	2	D_2	2.0	0.633	0.535	0.399
0.772	0.495	0.404		D_3	1.5	0.560	0.469	0.407
0.976	0.854	0.439		D_1	4.5	0.914	0.782	0.458
0.961	0.767	0.435	3	D_2	4.0	0.830	0.692	0.442
0.935	0.690	0.416		D_3	3.5	0.776	0.624	0.426
0.987	0.920	0.499		D_1	6.5	0.970	0.876	0.516
0.982	0.870	0.482	4	D_2	6.0	0.918	0.794	0.491
0.973	0.813	0.435		D_3	5.5	0.872	0.726	0.451
0.990	0.942	0.586		D_1	8.5	0.985	0.921	0.587
0.987	0.914	0.538	5	D_2	8.0	0.958	0.858	0.542
0.982	0.877	0.463		D_3	7.5	0.923	0.797	0.483
0.992	0.952	0.675		D_1	10.5	0.990	0.943	0.657
0.989	0.934	0.598	6	D_2	10.0	0.976	0.898	0.593
0.986	0.910	0.502		D_3	9.5	0.951	0.846	0.518
0.993	0.959	0.745		D_1	12.5	0.992	0.955	0.717
0.991	0.944	0.657	7	D_2	12.0	0.984	0.922	0.641
0.988	0.927	0.548		D_3	11.5	0.968	0.880	0.556
0.994	0.964	0.793		D_1	14.5	>0.992	0.964	0.766
0.992	0.952	0.708	8	D_2	14.0	0.989	0.938	0.684
0.990	0.937	0.598		D_3	13.5	0.977	0.904	0.593

Table 3.3.3 Probabilities, $P_1(N,\rho)$ and $P_1(V,\rho)$, that in alternative (0,N) and (0,V) shut-down queues queueing times exceed mean queueing time in a corresponding non-shut-down queue with traffic intensity, ρ, and D_1 (Erlang), D_2 (exponential) or D_3 (mixed exponential) service times.

choice in that specific queueing situation.

Table 3.3.3 shows that when ρ equals 0.25 and 0.55, $P_1(V,\rho) < P_1(N,\rho)$ for each D_i, though the difference between corresponding entries is not always significant. When traffic is heavy, e.g. ρ=0.85, neither (0,N) nor (0,V) control is clearly the better choice on the basis of the stated criterion. However, corresponding entries in the table for alternative versions of shut-down control are negligibly different when ρ=0.85. Therefore, Table 3.3.3 suggests that, from the customer's perspective, (0,V) control is probably a better choice than (0,N) control for many queueing situations. However, the advantages to customers of adopting (0,V) control instead of (0,N) control in heavily congested queueing systems are, at best, marginal.

In Chapters 4 and 5 we return to the problem of controlling congestion in the queue; therefore, we now relax special definitions and conventions which were a necessary part of the preceding discussion of shut-down control.

CHAPTER 4. Adaptive control of the service process

4.1 Linking the service process to the line size

Modern industrial processes frequently utilize feedback control procedures to ensure the quality of the finished product. By frequent or continuous monitoring of selected properties, process deviations which might affect product quality are detected. Corrective action is then automatically taken to prevent an unacceptable decline in the quality of the end-product.

Similar techniques ought to lend themselves naturally to the problem of controlling congestion in a queueing system. However, relatively few models have been suggested. Those which have appeared usually link the control action to the line size process. This association is a natural one since the length of the line is one indication of the amount of congestion in the system. Furthermore, the number of customers present may be one of the few properties of a queueing system which can be quickly evaluated or continuously monitored.

Control models which have been suggested generally require Markov assumptions. For the M/M/s queueing process, Moder & Phillips(1962) propose a control rule which permits the number of active servers to vary between the limits s_1 and s_2 $(s_1<s_2)$. Two control levels, n and N $(n<N)$ are selected, and initially s_1 servers are active. Each time the queue size increases from N-1 to N customers, an additional server is introduced unless s_2 servers are already busy. Conversely, each time the queue size drops from n+1 to n customers, a server is withdrawn unless only s_1 servers are active. The authors evaluate the usual equilibrium properties such as the state probability distribution, $E(L)$, $E(L_q)$ and the rate at which servers are activated. Some numerical computations illustrate the results obtained.

Unless channel start-up and running costs are negligible, however, the rule which Moder and Phillips propose is probably too sensitive to random fluctuations in the queue size. When an extra serving channel has just been opened, the coincidental arrival of another customer before at least one departure occurs is hardly sufficient reason to open yet another service channel.

A different approach to a similar situation has been suggested by Magazine(1971) who considers the same Markov queue with s servers but restricts the system capacity to M customers. At points equi-spaced in time a control decision is taken based on the number of customers present and the number of active servers, say $k \leq s$. Three different types of decisions are possible; additional servers, to a maximum of s-k, may be activated, surplus servers, to a maximum of k, may be withdrawn, or the system may be left unchanged. By assuming constant start-up, shut-down, and unit operating costs which are identical for each server, and convex customer holding costs, the author is able to use dynamic programming techniques to deduce the form of an optimal control policy. Though Magazine considers three different criteria for optimality — minimum expected operating cost discounted over a finite horizon, the same cost discounted over an infinite horizon, and minimum average operating cost — he shows that, in each case, the optimal control policy for this periodic review situation can always be characterized by a non-decreasing sequence of s integers. Regrettably, Magazine does not suggest a method for determining the optimal policy in a particular situation nor does he provide worked examples to illustrate the results.

In concurrent papers, Yadin & Naor(1967) and Gebhard(1967) have suggested an interesting technique called hysteresis control. The rudiments of the method are probably best understood by referring to Fig. 4.1.1. In the simplest case, two control thresholds, r and R ($r<R$) are selected.

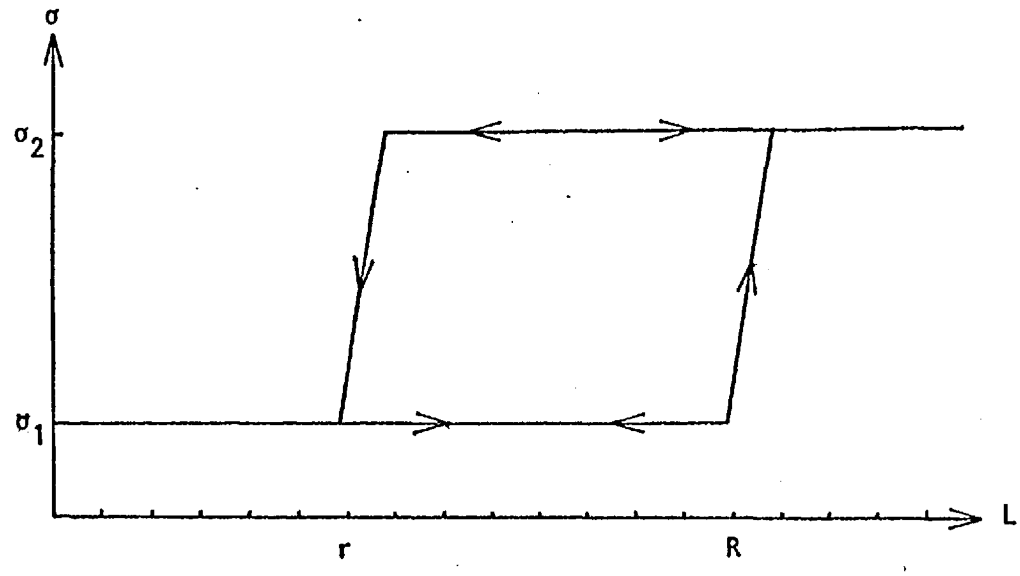


Fig. 4.1.1 Relation between line size, L, and the exponential service time rate parameter, σ, determining a hysteresis control pattern.

Customers may be served according to exponential distributions at one of two rates, σ_1 or σ_2, with $\sigma_1<\sigma_2$; the initial service rate is determined by the initial line size. If the line size increases to R, the service rate changes from σ_1 to σ_2; if the line length subsequently decreases to r customers, the service rate changes to σ_1 again. Gebhard(1967) analyzes both the unilevel (single control parameter, hence r+1=R) and bilevel hysteresis control models (r<R-1) for the M/M/1 queue in the steady state. Yadin & Naor(1967) consider a somewhat more general rule, again in the steady state. A total of k pairs of control levels (r_n,R_n) $(n=1,\ldots,k)$ are permitted by these authors; arrivals are Poisson and the k+1 distinct service time distributions are exponential with means $1/\mu_n$ $(n=1,\ldots,k+1)$ respectively, where $0<\mu_1<\cdots<\mu_{k+1}$. In each paper the equilibrium distribution for L and its usual associated properties are obtained.

More recently, Scott(1971) has suggested a model for hysteresis control of the Poisson arrival process of an M/G/1 queue. Neither unilevel nor bilevel hysteresis control of the service process for the M/G/1 queueing system appears to have been considered at any time.

4.2 Unilevel control of the service process

Unilevel service process control is an elementary analogue of modern industrial feedback control characterized by a single control parameter N. Customers are individually served at a single counter according to one of two service time distributions, $G_1(\cdot)$ or $G_2(\cdot)$. If L_t represents the number of customers in the queueing system at any time $t \geq 0$, then unilevel control rules specify that:

(i) when $L_t < N$ the customer at the service counter is served according to the distribution $G_1(\cdot)$,

(ii) when $L_t \geq N$ the customer at the service counter is served according to the distribution $G_2(\cdot)$,

(iii) if L_t increases from N-1 to N while a customer, C, is being served, immediately terminate service to C according to $G_1(\cdot)$ and begin a new service time for the same customer according to $G_2(\cdot)$.

Though rule (iii) might be considered unrealistic, it simplifies the analysis of the resulting line size process. Hence, decisions to change the service time distribution are made at arrival and service epochs.

We suppose that the probability density function $g_i(\cdot)$ corresponding to the service time distribution $G_i(\cdot)$ $(i=1,2)$ can be written in the form $g_i(\cdot)=\emptyset_i(\cdot)\mathcal{G}_i(\cdot)$, where $\mathcal{G}_i(\cdot)=1-G_i(\cdot)$, $(i=1,2)$, and that $\rho_2=\lambda\int_0^\infty tg_2(t)dt<1$. We will denote the Laplace transform of the probability density function $g_i(t)$ by $g_i^*(s)=\int_0^\infty e^{-st}g_i(t)dt$, $(i=1,2)$.

Since L_t is generally non-Markov, we augment the state space by adding the supplementary variable S_t representing the elapsed service time of the customer being served. The stochastic process (L_t,S_t) is Markov.

Let $p_o(t)=pr(L_t=0)$ $(t \geq 0)$ and $p_n(t,x)=pr(L_t=n,S_t=x)$ $(n=1,2,\ldots;\ x \geq 0;\ t \geq 0)$. The following time-dependent Kolmogorov forward differential equations can be obtained:

$$\frac{\partial}{\partial t} p_0(t) + \lambda p_0(t) = \int_0^\infty p_1(t,x)\phi_1(x)dx , \qquad (4.2.1)$$

$$\frac{\partial}{\partial t} p_1(t,x) + \frac{\partial}{\partial x} p_1(t,x) + \{\lambda+\phi_1(x)\} p_1(t,x) = 0 , \qquad (4.2.2)$$

$$\frac{\partial}{\partial t} p_j(t,x) + \frac{\partial}{\partial x} p_j(t,x) + \{\lambda+\phi_1(x)\} p_j(t,x) = \lambda p_{j-1}(t,x) \quad , (j=2,\cdots,N-1) \qquad (4.2.3)$$

$$\frac{\partial}{\partial t} p_N(t,x) + \frac{\partial}{\partial x} p_N(t,x) + \{\lambda+\phi_2(x)\} p_N(t,x) = 0 \quad , \qquad (4.2.4)$$

$$\frac{\partial}{\partial t} p_j(t,x) + \frac{\partial}{\partial x} p_j(t,x) + \{\lambda+\phi_2(x)\} p_j(t,x) = \lambda p_{j-1}(t,x) , (j=N+1,N+2,\ldots) \qquad (4.2.5)$$

Solutions to (4.2.1) to (4.2.5) must also satisfy the boundary equations

$$p_1(t,0) = \lambda p_0(t) + \int_0^\infty p_2(t,x)\phi_1(x)dx , \qquad (4.2.6)$$

$$p_j(t,0) = \int_0^\infty p_{j+1}(t,x)\phi_1(x)dx \quad , \quad (j=2,\cdots,N-2) \qquad (4.2.7)$$

$$p_{N-1}(t,0) = \int_0^\infty p_N(t,x)\phi_2(x)dx \quad , \qquad (4.2.8)$$

$$p_N(t,0) = \lambda \int_0^\infty p_{N-1}(t,x)dx + \int_0^\infty p_{N+1}(t,x)\phi_2(x)dx , \qquad (4.2.9)$$

$$p_j(t,0) = \int_0^\infty p_{j+1}(t,x)\phi_2(x)dx \quad , (j=N+1,N+2,\ldots) . \qquad (4.2.10)$$

Let p_o and $p_n(x)$ be the steady-state analogues of $p_o(t)$ and $p_n(t,x)$ ($n=1,2,\ldots;\ x \geq 0$). Equilibrium equations corresponding to (4.2.1) to (4.2.10) are given by

$$\lambda p_0 = \int_0^\infty p_1(x)\phi_1(x)dx \quad , \qquad (4.2.11)$$

$$\frac{\partial}{\partial x} p_1(x) + \{\lambda+\phi_1(x)\} p_1(x) = 0 \quad , \qquad (4.2.12)$$

$$\frac{\partial}{\partial x} p_j(x) + \{\lambda+\phi_1(x)\} p_j(x) = \lambda p_{j-1}(x) \quad , (j=2,\cdots,N-1) \qquad (4.2.13)$$

$$\frac{\partial}{\partial x} p_N(x) + \{\lambda+\phi_2(x)\} p_N(x) = 0 \quad , \qquad (4.2.14)$$

$$\frac{\partial}{\partial x} p_j(x) + \{\lambda+\phi_2(x)\} p_j(x) = \lambda p_{j-1}(x) \quad , \quad (j=N+1,N+2,\cdots) \qquad (4.2.15)$$

$$p_1(0) = \lambda p_0 + \int_0^\infty p_2(x)\phi_1(x)dx , \qquad (4.2.16)$$

$$p_j(0) = \int_0^\infty p_{j+1}(x)\phi_1(x)dx \quad , \quad (j=2,\cdots,N-2) \qquad (4.2.17)$$

$$p_{N-1}(0) = \int_0^\infty p_N(x)\,\phi_2(x)\,dx , \tag{4.2.18}$$

$$p_N(0) = \lambda \int_0^\infty p_{N-1}(x)\,dx + \int_0^\infty p_{N+1}(x)\,\phi_2(x)\,dx , \tag{4.2.19}$$

$$p_j(0) = \int_0^\infty p_{j+1}(x)\,\phi_2(x)\,dx , \qquad (j=N+1, N+2, \cdots) . \tag{4.2.20}$$

Solving (4.2.12) and (4.2.13) iteratively we obtain

$$p_j(x) = \sum_{n=0}^{j-1} p_{j-n}(0)\,\frac{(\lambda x)^n}{n!} e^{-\lambda x} \mathcal{G}_1(x) , \qquad (j=1,\cdots,N-1). \tag{4.2.21}$$

Define the probability generating functions $P_1(x;z)=\sum_{n=1}^{N-1} p_n(x)z^n$ and $P_2(x;z)=\sum_{n=N}^{\infty} p_n(x)z^n$ ($|z| \leqq 1$). Using $P_2(x;z)$ we can combine (4.2.14) and (4.2.15) in the single equation

$$\frac{\partial}{\partial x} P_2(x;z) = \left\{ \lambda z - \lambda - \phi_2(x) \right\} P_2(x;z)$$

which has the solution

$$P_2(x;z) = P_2(0;z)\, e^{-\lambda x(1-z)} \mathcal{G}_2(x) , \tag{4.2.22}$$

where $P_2(0;z)=\lim_{x\to 0+} P_2(x;z)$.

Similarly, we can combine (4.2.18), (4.2.19) and (4.2.20) in the equation

$$P_2(0;z) = \lambda p_{N-1} z^N - p_{N-1}(0)\, z^{N-1} + \frac{1}{z}\int_0^\infty P_2(x;z)\,\phi_2(x)\,dx, \tag{4.2.23}$$

where $p_j=\int_0^\infty p_j(x)dx$, $(j=1,2,\ldots)$. By substituting (4.2.22) in (4.2.23) and solving for $P_2(0;z)$ it follows that

$$P_2(0;z) = \frac{\lambda p_{N-1} z^{N+1} - p_{N-1}(0)\, z^N}{z - g_2^*(\lambda - \lambda z)} . \tag{4.2.24}$$

Therefore, once expressions for the $p_j(0)$'s $(j=1,\ldots,N-1)$ have been obtained, both $P_1(x;z)$ and $P_2(x;z)$ will be uniquely determined.

Using (4.2.21) in (4.2.11) and (4.2.16) we can show that

$$p_1(0) = \frac{\lambda p_0}{g_1^*(\lambda)} , \qquad p_2(0) = \frac{\lambda p_0}{\{g_1^*(\lambda)\}^2}\left\{ 1 - g_1^*(\lambda) + \lambda \frac{d}{d\lambda} g_1^*(\lambda) \right\} .$$

In general, values for $p_3(0),\ldots,p_{N-1}(0)$, unique to within p_0, may be obtained iteratively by solving the equation

$$\sum_{k=0}^{j} p_{j+1-k}(0)\left\{\delta_{1k} - \frac{(-\lambda)^k}{k!} g_1^{*(k)}(\lambda)\right\} = 0\,, \qquad (j=2,\ldots,N-2)$$

where $g_1^{*(k)}(\lambda)=\frac{d^k}{d\lambda^k} g_1^*(\lambda)$, $(k=0,\ldots,N-2)$ and δ_{1k} is the familiar Kronecker delta. By evaluating p_o, unique solutions for (4.2.11) to (4.2.20) will be determined.

Let $P_1(z)=\sum_{k=1}^{N-1} p_k z^k = \int_0^\infty P_1(x;z)dx$, $P_2(z)=\sum_{k=N}^{\infty} p_k z^k = \int_0^\infty P_2(x;z)dx$. The normalizing equation which determines p_o is therefore

$$p_o + P_1(1) + P_2(1) = 1\,. \tag{4.2.25}$$

By integrating (4.2.22) with respect to x and substituting for $P_2(0;z)$ we can show that $P_2(1)=p_{N-1}\frac{\rho_2}{1-\rho_2}$. Hence (4.2.25) becomes

$$p_o + p_{N-1}\frac{\rho_2}{1-\rho_2} + \sum_{j=1}^{N-1}\sum_{n=0}^{j-1} p_{j-n}(0)\frac{\lambda^n}{n!}J_n(\lambda) = 1 \quad , \tag{4.2.26}$$

where $J_n(\lambda)=\int_0^\infty x^n \mathcal{G}_1(x)e^{-\lambda x}dx$, $(n=0,\ldots,N-2)$.

Expressions for the marginal probability generating functions $P_1(z)$ and $P_2(z)$ are given by

$$P_1(z) = \sum_{k=1}^{N-1} z^k\left\{\sum_{n=0}^{k-1} p_{k-n}(0)\frac{\lambda^n}{n!}J_n(\lambda)\right\} \quad ,$$

$$P_2(z) = p_{N-1}z^N\frac{g_2^*(\lambda-\lambda z)-1}{z-g_2^*(\lambda-\lambda z)}\,.$$

Useful general expressions can be written for several interesting properties of the model. The mean line length is given by

$$E(L) = P_1'(1) + \frac{p_{N-1}}{1-\rho_2}\left\{N\rho_2 + \frac{1}{2}\frac{\lambda^2 g_{2,2}}{1-\rho_2}\right\}$$

where $g_{i,j}=\int_0^\infty t^j g_i(t)dt$ $(i=1,2;\ j=1,2,\ldots)$. The rate, σ, at which changes from $G_1(\cdot)$ to $G_2(\cdot)$ occur is given by

$$\begin{aligned}\sigma &= \lambda p_{N-1}\\ &= \int_0^\infty p_N(x)\phi_2(x)dx = p_N(0)\,g_2^*(\lambda)\ ;\end{aligned}$$

the total rate of changes in the service time distribution is 2σ. The long-run proportion of time during which customers are served according to $G_2(\cdot)$ is given by

$$\xi = P_2(1) = P_{N-1}\,\frac{\rho_2}{1-\rho_2}\ .$$

Similarly, if η is the long-run proportion of customers who are served according to $G_2(\cdot)$ we can show that

$$\eta = \xi + P_{N-1} = \frac{P_{N-1}}{1-\rho_2}$$

The following examples illustrate the application of unilevel service process control to a queueing process.

Example 4.2.1

Let $G_i(x)=1-e^{-\mu_i x}$ $(i=1,2)$ where $0<\mu_1<\mu_2$. This is the case which Gebhard(1967) treated. The results given below agree with expressions which he obtains in another way. Using the boundary equation solutions

$$p_1(0)=\lambda(1+\rho_1)p_0\ ,\qquad p_j(0)=\lambda\rho_1^j p_0\ ,\ (j=2,\cdots,N-1)$$

we can show that

$$P_1(z)=\frac{\rho_1 z-(\rho_1 z)^N}{1-\rho_1 z}\,p_0\ ,\qquad P_2(z)=\frac{\rho_1^{N-1}\rho_2 z^N}{1-\rho_2 z}\,p_0\ .$$

Hence

$$p_j=\rho_1^j p_0\ ,\ (j=1,\cdots,N-1)\ ,\ p_j=\rho_1^{N-1}\rho_2^{j-N+1}p_0\ ,\ (j=N,N+1,\cdots)$$

and

$$p_0=\frac{(1-\rho_1)(1-\rho_2)}{1-\rho_2-\rho_1^{N-1}(\rho_1-\rho_2)}\ .$$

Formulae for $E(L)$, σ, ξ, and η are given by

$$E(L)=p_0\left[\frac{\rho_1}{(1-\rho_1)^2}-\frac{\rho_1^{N-1}(\rho_1-\rho_2)}{(1-\rho_1)(1-\rho_2)}\left\{N-1+\frac{1-\rho_1\rho_2}{(1-\rho_1)(1-\rho_2)}\right\}\right],$$

$$\sigma=\lambda\rho_1^{N-1}p_0\ ,\qquad \xi=\frac{\rho_1^{N-1}\rho_2}{1-\rho_2}\,p_0\ ,\qquad \eta=\frac{\rho_1^{N-1}}{1-\rho_2}\,p_0\ .$$

Since we can replace μ_2 in the preceding example by $c\mu_1$ $(c>1)$, it follows that the assumption, $G_2(x)=G_1(cx)$, does not produce any special simplification of the general results.

By forming an equilibrium probability generating function for p_{N-1}, p_N,... in the general case, we see that

$$p_{N-1}z^{N-1} + P_2(z) = p_{N-1}z^{N-1}\frac{(z-1)\,g_2^*(\lambda-\lambda z)}{z-g_2^*(\lambda-\lambda z)} \; . \qquad (4.2.27)$$

This expression resembles the probability generating function for the equilibrium line size distribution in an M/G/1 queue [cf. Cox & Smith(1961, p.56)]. The similarity is due to rule (iii) and underlines the fact that transitions from state N-1 to state N initiate a change in the service time distribution.

Example 4.2.2

Let $G_1(x)=(1-p)(1-e^{-\mu_1 x}) + p(1-e^{-\nu_1 x})$ $(0<p<1;\ \mu_1,\ \nu_1>0)$

and $G_2(x)=\int_0^x \mu_2^2 y e^{-\mu_2 y}dy$ i.e. $G_1(\cdot)$ is a mixed exponential distribution and $G_2(\cdot)$ is a two-stage Erlang distribution. The equilibrium probability distribution for L is given by

$$p_j = p_0\left\{\frac{s_2-\mu_1-p(\nu_1-\mu_1)}{s_2-s_1}\left(\frac{\lambda}{s_1}\right)^j + \frac{\mu_1+p(\nu_1-\mu_1)-s_1}{s_2-s_1}\left(\frac{\lambda}{s_2}\right)^j\right\},\ (j=0,\cdots,N-1)$$

$$p_j = \frac{p_{N-1}}{t_2-t_1}\left\{t_2\left(\frac{\lambda}{t_1}\right)^{j-N+1} - t_1\left(\frac{\lambda}{t_2}\right)^{j-N+1}\right\}, \qquad (j=N,N+1,\cdots)$$

where s_1, s_2 are the roots of $x^2-(\nu_1+\mu_1+\lambda)x+\mu_1(\nu_1+\lambda)+p\lambda(\nu_1-\mu_1)=0$ and t_1, t_2 are the roots of $x^2-x(\lambda+2\mu_2)+\mu_2^2=0$. Expressions for p_0 and E(L) are given by

$$\frac{1}{p_0} = \frac{s_1s_2-s_1\{\mu_1+p(\nu_1-\mu_1)\}}{s_2-s_1}\,\frac{1-\left(\frac{\lambda}{s_1}\right)^N}{s_1-\lambda} + \frac{s_2\{\mu_1+p(\nu_1-\mu_1)\}-s_1s_2}{s_2-s_1}\,\frac{1-\left(\frac{\lambda}{s_2}\right)^N}{s_2-\lambda}$$
$$+\frac{\rho_2}{1-\rho_2}\left\{\frac{s_2-\mu_1-p(\nu_1-\mu_1)}{s_2-s_1}\left(\frac{\lambda}{s_1}\right)^{N-1} + \frac{\mu_1+p(\nu_1-\mu_1)-s_1}{s_2-s_1}\left(\frac{\lambda}{s_2}\right)^{N-1}\right\},$$

$$E(L) = p_0\left[\frac{s_1s_2-s_1\{\mu_1+p(\nu_1-\mu_1)\}}{s_2-s_1}\,\frac{(\lambda-s_1)N\left(\frac{\lambda}{s_1}\right)^N+\lambda\left\{1-\left(\frac{\lambda}{s_1}\right)^N\right\}}{(s_1-\lambda)^2}\right.$$
$$+\frac{s_2\{\mu_1+p(\nu_1-\mu_1)\}-s_1s_2}{s_2-s_1}\,\frac{(\lambda-s_2)N\left(\frac{\lambda}{s_2}\right)^N+\lambda\left\{1-\left(\frac{\lambda}{s_2}\right)^N\right\}}{(s_2-\lambda)^2}$$
$$\left.+\left\{N\rho_2+\frac{3\rho_2^2}{4(1-\rho_2)}\right\}\left\{\frac{s_2-\mu_1-p(\nu_1-\mu_1)}{(1-\rho_2)(s_2-s_1)}\left(\frac{\lambda}{s_1}\right)^{N-1} + \frac{\mu_1+p(\nu_1-\mu_1)-s_1}{(1-\rho_2)(s_2-s_1)}\left(\frac{\lambda}{s_2}\right)^{N-1}\right\}\right]$$

In §4.3 the results for this particular combination of service time distributions will be used in a numerical study.

The next example characterizes all service time distributions $G_2^+(\cdot)$ which, if substituted for $G_2(\cdot)$ in a unilevel control scheme, would reduce the mean number of customers in the system.

Example 4.2.3

Let $G_1(\cdot)$, $G_2(\cdot)$ and $G_2^+(\cdot)$ be three distinct service time distribution functions. Suppose that $G_2(\cdot)$ and $G_2^+(\cdot)$ have the same mean μ but variances σ^2 and σ^{+2} respectively. Since solutions for $p_j(0)$ $(j=1,\dots,N-1)$ depend only on $G_1(\cdot)$ and λ, the p_j's $(j=0,\dots,N-1)$ will be the same if we substitute $G_2^+(\cdot)$ for $G_2(\cdot)$. Let $E(L)$ and $E(L^+)$ represent the mean line size for the distribution pairs $G_1(\cdot)$, $G_2(\cdot)$ and $G_1(\cdot)$, $G_2^+(\cdot)$, respectively. Then

$$E(L) - E(L^+) = \frac{\lambda^2 P_{N-1}}{2(1-\rho_2)^2}\left(\sigma^2 - \sigma^{+2}\right) .$$

Therefore, provided $G_2^+(\cdot)$ is less dispersed than $G_2(\cdot)$, i.e. $\sigma^+<\sigma$, the mean number of customers is always less when $G_2^+(\cdot)$ is substituted for $G_2(\cdot)$. This is another consequence of rule (iii). Obviously, the decrease in $E(L)$ will be maximized for fixed N and ρ_2 if $\sigma^+=0$; that is, if service times are constant when the line size exceeds $N-1$.

4.3 The effect of unilevel control on the distribution of L

The theoretical results of §4.2 do not indicate how much the line size distribution is affected by unilevel control, nor the occasions when unilevel service process control can be implemented to distinct advantage. By investigating, numerically, the distribution of L in different situations, we shall try to explore these two questions.

Three features of unilevel control can be adjusted; these are the control threshold parameter, N, and the two service time distributions $G_1(\cdot)$ and $G_2(\cdot)$. Since service time distributions may be largely determined by other considerations, we will concentrate on the relation between N and the distribution of L, regarding the choice of $G_1(\cdot)$ and $G_2(\cdot)$ as a question of secondary importance.

Unacceptable levels of congestion are often associated with very long-tailed distributions for properties of the queueing system such as line length and queueing time. To determine characteristics of the line size distribution under unilevel control and to gauge the effect of unilevel control on a congested queueing system, we will compare tail probabilities for the equilibrium distribution of line length in different sets of circumstances. To standardize comparisons for different distributions of L under unilevel control we will use the mean value of each line size distribution as the respective unit of scale. Therefore, we calculate the probability that line length exceeds integral multiples of its mean value, i.e. $R_k(N;\rho_1,\rho_2)=\mathrm{pr}\{L>kE(L)\}$ $(k=1,2,\ldots;\ N=2,3,\ldots)$. Since the traffic intensities ρ_1, ρ_2 also indicate the relative level of congestion, we will consider various combinations of ρ_1 and ρ_2.

We choose $G_1(\cdot)$ and $G_2(\cdot)$ to represent a range of distributions. In §§2.2 and 3.3 three distributions — a two-stage Erlang(D_1), an exponential (D_2) and a mixed exponential distribution(D_3) — were used in this way. We cannot consider all possible combinations of $G_1(\cdot)$ and $G_2(\cdot)$ from this triplet; however, three sensible choices are the pairs when $G_1(\cdot)$ and $G_2(\cdot)$ have the same mathematical form, e.g. both exponential, etc. A fourth choice is the pair with $G_1(\cdot)$ a mixed exponential distribution and $G_2(\cdot)$ a two-stage Erlang distribution, since Erlang service times are more regular than mixed exponential service times. The four combinations, then, are

$$C_1:\quad g_1(x)=\mu_1^2xe^{-\mu_1x}\quad,\qquad g_2(x)=\mu_2^2xe^{-\mu_2x}\quad,$$

$$C_2:\quad g_2(x)=\nu_1e^{-\nu_1x}\quad,\qquad g_2(x)=\nu_2e^{-\nu_2x}\quad,$$

$$C_3:\ g_1(x)=(1-p_1)\alpha_1 e^{-\alpha_1 x}+p_1\beta_1 e^{-\beta_1 x},\ g_2(x)=(1-p_2)\alpha_2 e^{-\alpha_2 x}+p_2\beta_2 e^{-\beta_2 x},$$
$$0<p_1,p_2<1,$$

$$C_4:\ g_1(x)=(1-q_1)\gamma_1 e^{-\gamma_1 x}+q_1\delta_1 e^{-\delta_1 x},\ 0<q_1<1,\ \ g_2(x)=\gamma_2^2 x e^{-\gamma_2 x}.$$

For each combination C_i $(i=1,2,3,4)$, pairs of values (ρ_1,ρ_2) in differing ratios to each other can only be obtained by adjusting the parameters of the distributions. For fixed λ, the values of μ_i, ν_i $(i=1,2)$ in C_1 and C_2 are determined if ρ_1 and ρ_2 are fixed. The same is not true of the mixed exponential distribution. As in §§2.2, 2.4 and 3.3 we require that $p_i=\frac{1}{2}$, $\alpha_i=3\beta_i$ $(i=1,2)$, $q_1=\frac{1}{2}$ and $\gamma_1=3\delta_1$; γ_2 is determined if λ, ρ_2 are fixed.

Theoretical expressions for the equilibrium distribution of L for the combinations C_2 and C_4 are given in Examples 4.2.1 and 4.2.2 respectively. Results for the other two combinations appear below. For C_1 and C_3 the expression for E(L) has been omitted; this can be derived from the equilibrium distributions

$$C_1:\qquad P_j = P_0\frac{\mu_1^2}{\lambda(s_1-s_2)}\left\{\left(\frac{\lambda}{s_2}\right)^{j+1}-\left(\frac{\lambda}{s_1}\right)^{j+1}\right\},\qquad (j=0,\dots,N-1)$$
$$P_j = \frac{P_{N-1}}{t_2-t_1}\left\{t_2\left(\frac{\lambda}{t_1}\right)^{j-N+1}-t_1\left(\frac{\lambda}{t_2}\right)^{j-N+1}\right\},\qquad (j=N,N+1,\dots)$$

$$\frac{1}{P_0}=1+\frac{\mu_1^2}{s_1-s_2}\left[\frac{\left(\frac{\lambda}{s_2}\right)-\left(\frac{\lambda}{s_2}\right)^N}{s_2-\lambda}-\frac{\left(\frac{\lambda}{s_1}\right)-\left(\frac{\lambda}{s_1}\right)^N}{s_1-\lambda}+\frac{2}{\mu_2(1-\rho_2)}\left\{\left(\frac{\lambda}{s_2}\right)^N-\left(\frac{\lambda}{s_1}\right)^N\right\}\right]$$

where s_1, s_2 are the roots of $x^2-x(\lambda+2\mu_1)+\mu_1^2=0$ and t_1, t_2 are the roots of $x^2-x(\lambda+2\mu_2)+\mu_2^2=0$,

$$C_3:\ P_j = P_0\left\{\frac{s_2'-\alpha_1-p_1(\beta_1-\alpha_1)}{s_2'-s_1'}\left(\frac{\lambda}{s_1'}\right)^j+\frac{\alpha_1+p_1(\beta_1-\alpha_1)-s_1'}{s_2'-s_1'}\left(\frac{\lambda}{s_2'}\right)^j\right\},\ (j=0,\dots,N-1)$$
$$P_j=\frac{P_{N-1}}{t_2'-t_1'}\left[\left\{\beta_2+\lambda-t_1'-p_2(\beta_2-\alpha_2)\right\}\left(\frac{\lambda}{t_1'}\right)^{j-N+1}+\left\{p_2(\beta_2-\alpha_2)-\beta_2-\lambda+t_2'\right\}\left(\frac{\lambda}{t_2'}\right)^{j-N+1}\right],$$
$$(j=N,N+1,\dots)$$
$$\frac{1}{P_0}=\frac{\rho_2}{1-\rho_2}\left\{\frac{s_2'-\alpha_1-p_1(\beta_1-\alpha_1)}{s_2'-s_1'}\left(\frac{\lambda}{s_1'}\right)^{N-1}+\frac{\alpha_1+p_1(\beta_1-\alpha_1)-s_1'}{s_2'-s_1'}\left(\frac{\lambda}{s_2'}\right)^{N-1}\right\}$$
$$+\frac{s_1's_2'-s_1'\{\alpha_1+p_1(\beta_1-\alpha_1)\}}{s_2'-s_1'}\,\frac{1-\left(\frac{\lambda}{s_1'}\right)^N}{s_1'-\lambda}+\frac{s_2'\{\alpha_1+p_1(\beta_1-\alpha_1)\}-s_1's_2'}{s_2'-s_1'}\,\frac{1-\left(\frac{\lambda}{s_2'}\right)^N}{s_2'-\lambda}.$$

where s_1', s_2' are the roots of $x^2-(\beta_1+\alpha_1+\lambda)x+\alpha_1(\beta_1+\lambda)+p_1\lambda(\beta_1-\alpha_1)=0$ and t_1', t_2' are the roots of $x^2-(\beta_2+\alpha_2+\lambda)x+\alpha_2(\beta_2+\lambda)+p_2\lambda(\beta_2-\alpha_2)=0$.

The probabilities $R_k(N;\rho_1,\rho_2)$ $(k=1,2,3;\ N=2,\ldots,10)$ have been calculated for different values of ρ_1 and ρ_2; the results of the calculations are given in Tables 4.3.1 a, b, c and d. The mean values, E(L), which were used in the calculations are also given. To represent a range of queueing situations, eight different traffic intensity combinations, (ρ_1, ρ_2), were considered for each C_i, these being (1.5,0.95), (1.5,0.55), (1.1,0.95), (1.1,0.55), (0.9,0.15), (0.9,0.55), (0.5,0.05) and (0.5,0.25). The values $\rho_1=1.5$ and $\rho_1=1.1$ represent systems which could not achieve an equilibrium without unilevel control. Each of these values is combined with $\rho_2=0.95$ and $\rho_2=0.55$ typifying heavy and moderate traffic conditions, respectively. Similarly, $\rho_2=0.25$ and $\rho_2=0.05$ represent light and very light traffic; these values would probably be a reasonable choice for ρ_2 only if traffic is moderate or light when fewer than N customers are present, i.e. $\rho_1 \leqq 0.5$.

Under the conditions shown in Tables 4.3.1 c and d, combinations C_3 and C_4 satisfy the requirements of Example 4.2.3. Since an Erlang distribution is underdispersed with respect to any mixed exponential distribution having the same mean, entries for E(L) in Table 4.3.1 d are uniformly smaller than corresponding entries in Table 4.3.1 c.

A general feature of the probabilities in Tables 4.3.1 a, b, c and d is the short-tailed character of each distribution considered. For all values of N, ρ_1 and ρ_2 except $\rho_1=0.5$, the probability, $R_3(N;\rho_1,\rho_2)$, that line length exceeds three times the mean value of the same distribution, is frequently much less than 0.1. Similarly, unless $\rho_1=0.5$, $R_2(N;\rho_1,\rho_2)$, the probability that a line exceeds twice its average length, is usually less than 0.15. Since $\rho_1=0.5$ appears to be exceptional, we consider this case separately.

E(L)										$R_k(N;\rho_1,\rho_2)$							
ρ_1=1.5		ρ_1=1.1		ρ_1=0.9		ρ_1=0.5		N	k	ρ_1=1.5		ρ_1=1.1		ρ_1=0.9		ρ_1=0.5	
ρ_2=.55	ρ_2=.95	ρ_2=.55	ρ_2=.95	ρ_2=.15	ρ_2=.55	ρ_2=.05	ρ_2=.25			ρ_2=.55	ρ_2=.95	ρ_2=.55	ρ_2=.95	ρ_2=.15	ρ_2=.55	ρ_2=.05	ρ_2=.25
									1	0.452	0.359	0.416	0.380	0.565	0.391	0.372	0.429
1.73	15.1	1.57	15.0	0.665	1.46	0.382	0.554	2	2	0.104	0.129	0.096	0.137	0.085	0.19	0.371	0.107
									3	0.022	0.046	0.045	0.049	0.085	0.042	0.019	0.107
									1	0.401	0.378	0.333	0.367	0.362	0.517	0.454	0.474
2.46	15.9	2.10	15.4	1.12	1.83	0.594	0.695	3	2	0.092	0.127	0.077	0.132	0.054	0.139	0.146	0.178
									3	0.009	0.043	0.017	0.044	0.006	0.030	0.146	0.045
									1	0.374	0.374	0.511	0.381	0.495	0.389	0.482	0.490
3.26	16.7	2.65	15.9	1.55	2.19	0.719	0.774	4	2	0.040	0.117	0.065	0.128	0.038	0.105	0.190	0.203
									3	0.004	0.037	0.014	0.043	0.005	0.023	0.059	0.073
									1	0.360	0.372	0.447	0.371	0.573	0.458	0.493	0.496
4.12	17.6	3.24	16.5	1.95	2.53	0.788	0.816	5	2	0.018	0.109	0.057	0.125	0.186	0.081	0.207	0.212
									3	0.001	0.034	0.006	0.039	0.003	0.018	0.079	0.084
									1	0.352	0.371	0.563	0.362	0.443	0.505	0.497	0.498
5.02	18.5	3.86	17.1	2.33	2.86	0.824	0.838	6	2	0.008	0.101	0.051	0.114	0.143	0.132	0.214	0.216
									3	0.000	0.030	0.002	0.036	0.003	0.014	0.086	0.089
									1	0.632	0.370	0.517	0.380	0.495	0.409	0.499	0.499
5.96	19.5	4.49	17.7	2.69	3.18	0.842	0.849	7	2	0.008	0.101	0.047	0.111	0.113	0.107	0.217	0.218
									3	0.000	0.026	0.001	0.033	0.000	0.011	0.090	0.091
									1	0.628	0.369	0.483	0.373	0.398	0.443	0.500	0.500
6.91	20.5	5.15	18.4	3.03	3.47	0.851	0.854	8	2	0.004	0.094	0.020	0.109	0.091	0.159	0.218	0.218
									3	0.000	0.023	0.000	0.030	0.000	0.009	0.091	0.091
									1	0.625	0.369	0.570	0.367	0.435	0.470	0.500	0.500
7.88	21.5	5.84	19.1	3.34	3.75	0.855	0.857	9	2	0.002	0.088	0.019	0.101	0.146	0.131	0.218	0.219
									3	0.000	0.020	0.000	0.028	0.000	0.008	0.091	0.092
									1	0.624	0.369	0.543	0.388	0.464	0.393	0.500	0.500
8.86	22.4	6.54	19.8	3.63	4.02	0.857	0.858	10	2	0.001	0.082	0.009	0.099	0.121	0.110	0.219	0.219
									3	0.000	0.017	0.000	0.025	0.001	0.006	0.092	0.092

Table 4.3.1 a Probabilities, $R_k(N;\rho_1,\rho_2)$, that line length, L, in a unilevel control queue with control threshold N exceeds k times its mean value, E(L). When $L<N(L\geq N)$ service times are Erlang(Erlang) with traffic intensity $\rho_1(\rho_2)$, i.e. combination C_1

E(L)								N	k	$R_k(N;\rho_1,\rho_2)$							
ρ_1=1.5		ρ_1=1.1		ρ_1=0.9		ρ_1=0.5				ρ_1=1.5		ρ_1=1.1		ρ_1=0.9		ρ_1=0.5	
ρ_2=.55	ρ_2=.95	ρ_2=.55	ρ_2=.95	ρ_2=.15	ρ_2=.55	ρ_2=.05	ρ_2=.25			ρ_2=.55	ρ_2=.95	ρ_2=.55	ρ_2=.95	ρ_2=.15	ρ_2=.55	ρ_2=.05	ρ_2=.25
									1	0.423	0.365	0.390	0.361	0.514	0.367	0.345	0.400
1.71	19.4	1.58	19.1	0.605	1.48	0.363	0.533	2	2	0.128	0.138	0.118	0.136	0.077	0.202	0.345	0.100
									3	0.039	0.049	0.065	0.051	0.077	0.061	0.017	0.100
									1	0.367	0.376	0.309	0.366	0.334	0.487	0.433	0.455
2.35	19.9	2.04	19.4	1.04	1.81	0.590	0.697	3	2	0.111	0.135	0.093	0.138	0.050	0.147	0.149	0.182
									3	0.019	0.048	0.028	0.049	0.008	0.045	0.149	0.046
									1	0.337	0.371	0.472	0.372	0.467	0.374	0.469	0.478
3.07	20.6	2.55	19.7	1.47	2.16	0.745	0.811	4	2	0.056	0.126	0.079	0.133	0.240	0.113	0.203	0.217
									3	0.009	0.045	0.024	0.048	0.005	0.034	0.070	0.087
									1	0.581	0.368	0.412	0.361	0.549	0.447	0.485	0.489
3.86	21.4	3.10	20.1	1.88	2.52	0.845	0.887	5	2	0.053	0.125	0.069	0.129	0.183	0.090	0.227	0.234
									3	0.005	0.041	0.011	0.046	0.004	0.027	0.098	0.106
									1	0.561	0.366	0.521	0.370	0.434	0.499	0.493	0.495
4.70	22.3	3.68	20.5	2.28	2.87	0.909	0.933	6	2	0.028	0.118	0.062	0.126	0.145	0.134	0.239	0.242
									3	0.001	0.038	0.006	0.045	0.003	0.022	0.112	0.116
									1	0.549	0.364	0.476	0.361	0.490	0.414	0.496	0.497
5.58	23.2	4.28	21.0	2.66	3.21	0.947	0.962	7	2	0.015	0.112	0.056	0.123	0.118	0.111	0.244	0.246
									3	0.001	0.034	0.005	0.042	0.003	0.018	0.119	0.120
									1	0.541	0.363	0.558	0.373	0.405	0.452	0.498	0.499
6.49	24.1	4.90	21.6	3.03	3.54	0.970	0.978	8	2	0.015	0.106	0.052	0.121	0.097	0.093	0.247	0.248
									3	0.000	0.031	0.003	0.041	0.000	0.016	0.122	0.128
									1	0.536	0.362	0.524	0.366	0.446	0.483	0.499	0.499
7.41	25.1	5.53	22.2	3.38	3.85	0.983	0.987	9	2	0.008	0.101	0.027	0.118	0.159	0.144	0.249	0.249
									3	0.000	0.028	0.001	0.038	0.000	0.013	0.123	0.124
									1	0.524	0.362	0.496	0.379	0.478	0.414	0.500	0.500
8.36	26.1	6.19	22.8	3.71	4.16	0.991	0.993	10	2	0.005	0.095	0.026	0.116	0.135	0.123	0.249	0.250
									3	0.000	0.025	0.001	0.036	0.000	0.011	0.124	0.124

Table 4.3.1 b Probabilities, $R_k(N;\rho_1,\rho_2)$, that line length, L, in a unilevel control queue with control threshold N exceeds k times its mean value, E(L). When $L<N$ ($L \geq N$) service times are exponential (exponential) with traffic intensity ρ_1 (ρ_2), i.e. combination C_2.

E(L)								N	k	$R_k(N;\rho_1,\rho_2)$							
ρ_1=1.5		ρ_1=1.1		ρ_1=0.9		ρ_1=0.5				ρ_1=1.5		ρ_1=1.1		ρ_1=0.9		ρ_1=0.5	
ρ_2=.55	ρ_2=.95	ρ_2=.55	ρ_2=.95	ρ_2=.15	ρ_2=.55	ρ_2=.05	ρ_2=.25			ρ_2=.55	ρ_2=.95	ρ_2=.55	ρ_2=.95	ρ_2=.15	ρ_2=.55	ρ_2=.05	ρ_2=.25
									1	0.407	0.369	0.374	0.364	0.483	0.351	0.325	0.379
1.77	23.6	1.63	23.3	0.571	1.53	0.343	0.514	2	2	0.146	0.138	0.134	0.142	0.072	0.126	0.325	0.095
									3	0.055	0.054	0.083	0.055	0.072	0.077	0.016	0.095
									1	0.345	0.374	0.290	0.363	0.615	0.459	0.415	0.437
2.33	24.0	2.02	23.2	0.982	1.80	0.569	0.679	3	2	0.124	0.140	0.104	0.141	0.310	0.149	0.147	0.179
									3	0.047	0.052	0.040	0.055	0.047	0.056	0.147	0.045
									1	0.312	0.368	0.440	0.364	0.440	0.355	0.455	0.466
3.00	24.5	2.49	23.4	1.40	2.12	0.739	0.812	4	2	0.069	0.132	0.143	0.142	0.226	0.115	0.205	0.221
									3	0.016	0.049	0.033	0.053	0.006	0.043	0.075	0.093
									1	0.533	0.364	0.383	0.368	0.522	0.429	0.475	0.481
3.74	25.2	3.00	23.6	1.80	2.47	0.862	0.912	5	2	0.065	0.131	0.076	0.137	0.176	0.158	0.235	0.243
									3	0.010	0.047	0.018	0.054	0.005	0.035	0.110	0.119
									1	0.511	0.361	0.487	0.372	0.418	0.483	0.486	0.489
4.53	26.0	3.55	23.9	2.20	2.82	0.950	0.983	6	2	0.038	0.124	0.068	0.139	0.142	0.131	0.250	0.255
									3	0.006	0.043	0.016	0.052	0.004	0.029	0.128	0.134
									1	0.497	0.374	0.445	0.363	0.475	0.407	0.259	0.262
5.36	26.9	4.12	24.3	2.59	3.17	1.01	1.03	7	2	0.037	0.124	0.062	0.136	0.117	0.110	0.139	0.141
									3	0.002	0.041	0.009	0.051	0.003	0.024	0.072	0.075
									1	0.487	0.372	0.524	0.370	0.518	0.448	0.264	0.266
6.23	27.8	4.71	24.7	2.97	3.52	1.05	1.06	8	2	0.023	0.118	0.057	0.133	0.190	0.095	0.144	0.146
									3	0.001	0.038	0.005	0.048	0.003	0.021	0.078	0.080
									1	0.480	0.371	0.490	0.362	0.441	0.480	0.267	0.268
7.12	28.7	5.31	25.1	3.34	3.86	1.08	1.09	9	2	0.014	0.113	0.054	0.130	0.163	0.148	0.148	0.149
									3	0.001	0.034	0.005	0.047	0.000	0.018	0.082	0.083
									1	0.476	0.370	0.554	0.371	0.476	0.418	0.269	0.269
8.03	29.6	5.93	25.7	3.70	4.19	1.09	1.10	10	2	0.009	0.108	0.051	0.128	0.140	0.129	0.149	0.150
									3	0.000	0.033	0.003	0.046	0.000	0.016	0.084	0.084

Table 4.3.1 c Probabilities, $R_k(N;\rho_1,\rho_2)$, that line length, L, in a unilevel control queue with control threshold N exceeds k times its mean value, E(L). When $L<N(L\geq N)$ service times are mixed exponential(mixed exponential) with traffic intensity $\rho_1(\rho_2)$, i.e. combination C_3.

E(L)										$R_k(N;\rho_1,\rho_2)$							
ρ_1=1.5		ρ_1=1.1		ρ_1=0.9		ρ_1=0.5		N	k	ρ_1=1.5		ρ_1=1.1		ρ_1=0.9		ρ_1=0.5	
ρ_2=.55	ρ_2=.95	ρ_2=.55	ρ_2=.95	ρ_2=.15	ρ_2=.55	ρ_2=.05	ρ_2=.25			ρ_2=.55	ρ_2=.95	ρ_2=.55	ρ_2=.95	ρ_2=.15	ρ_2=.55	ρ_2=.05	ρ_2=.25
									1	0.407	0.379	0.374	0.374	0.483	0.351	0.325	0.379
1.52	14.9	1.40	14.7	0.565	1.31	0.342	0.498	2	2	0.093	0.136	0.183	0.135	0.072	0.171	0.325	0.379
									3	0.043	0.049	0.040	0.048	0.072	0.082	0.016	0.095
									1	0.345	0.369	0.528	0.358	0.615	0.459	0.415	0.437
2.12	15.5	1.84	15.0	0.978	1.64	0.569	0.671	3	2	0.079	0.133	0.142	0.129	0.310	0.123	0.147	0.179
									3	0.017	0.045	0.031	0.046	0.047	0.058	0.147	0.045
									1	0.568	0.363	0.440	0.369	0.440	0.355	0.455	0.466
2.81	16.2	2.34	15.4	1.39	2.00	0.738	0.809	4	2	0.072	0.122	0.118	0.133	0.223	0.095	0.205	0.221
									3	0.007	0.041	0.012	0.045	0.004	0.021	0.075	0.093
									1	0.533	0.359	0.545	0.383	0.522	0.429	0.475	0.481
3.56	17.0	2.88	15.9	1.80	2.37	0.862	0.910	5	2	0.031	0.113	0.103	0.129	0.176	0.158	0.235	0.243
									3	0.003	0.035	0.010	0.043	0.003	0.017	0.110	0.119
									1	0.511	0.381	0.487	0.372	0.418	0.483	0.486	0.489
4.36	17.9	3.43	16.4	2.20	2.74	0.950	0.982	6	2	0.030	0.112	0.092	0.125	0.142	0.132	0.250	0.255
									3	0.001	0.033	0.004	0.039	0.003	0.014	0.128	0.134
									1	0.497	0.379	0.445	0.389	0.475	0.407	0.259	0.262
5.19	18.8	4.01	17.0	2.59	3.11	1.01	1.03	7	2	0.014	0.104	0.040	0.122	0.117	0.110	0.139	0.141
									3	0.000	0.028	0.002	0.038	0.002	0.012	0.072	0.075
									1	0.487	0.377	0.524	0.380	0.518	0.448	0.264	0.266
6.06	19.7	4.61	17.6	2.97	3.46	1.05	1.06	8	2	0.006	0.097	0.037	0.111	0.190	0.171	0.144	0.146
									3	0.000	0.025	0.002	0.035	0.002	0.010	0.078	0.080
									1	0.636	0.376	0.490	0.372	0.441	0.480	0.267	0.268
6.96	20.6	5.22	18.2	3.34	3.81	1.08	1.09	9	2	0.006	0.090	0.034	0.109	0.163	0.148	0.148	0.149
									3	0.000	0.023	0.001	0.032	0.000	0.009	0.082	0.083
									1	0.630	0.376	0.554	0.391	0.476	0.418	0.269	0.269
7.87	21.6	5.85	18.8	3.70	4.14	1.09	1.10	10	2	0.003	0.084	0.032	0.107	0.140	0.129	0.149	0.150
									3	0.000	0.020	0.000	0.029	0.000	0.008	0.084	0.084

Table 4.3.1 d Probabilities, $R_k(N;\rho_1,\rho_2)$, that line length, L, in a unilevel control queue with control threshold N exceeds k times its mean value, E(L). When $L<N(L\geq N)$ service times are mixed exponential(Erlang) with traffic intensity $\rho_1(\rho_2)$, i.e. combination C_4.

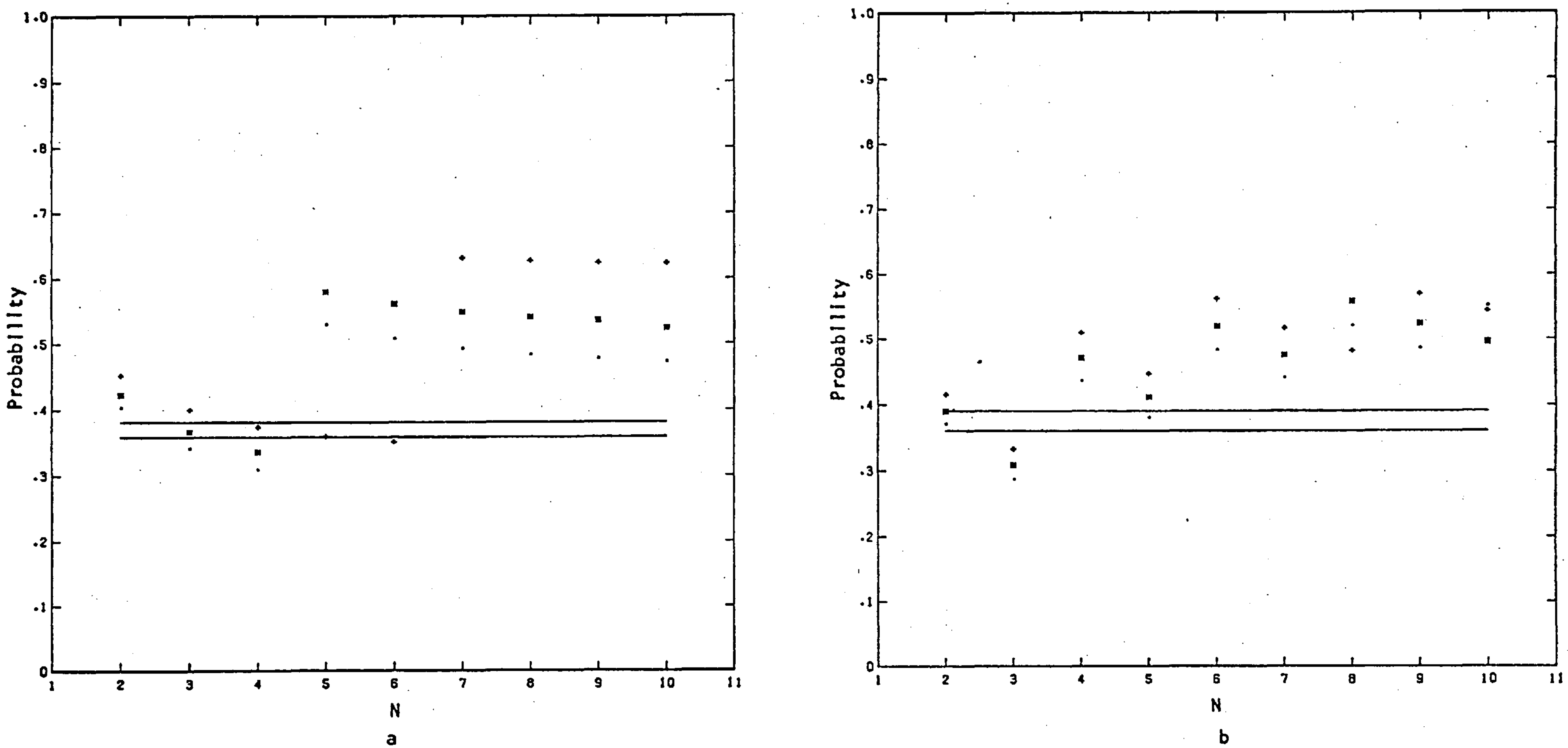


Fig. 4.3.1 Probabilities, $R_k(N;\rho_1,\rho_2)$, that the line length, L, exceeds k times its mean value in a unilevel control queue with control threshold N and service time combination C_i (i=1,2,3). In (a) k=1, ρ_1=1.5, ρ_2=0.55; in (b) k=1, ρ_1=1.1, ρ_2=0.55.

+ C_1 (service times are Erlang(Erlang) when L<N(L≥N))   ▪ C_2 (service times are exponential(exponential) when L<N(L≥N))

• C_3 (service times are mixed exponential(mixed exponential) when L<N(L≥N))

═ The range of $R_k(N;\rho_1,\rho_2)$ for all C_i's when ρ_2=0.95; k and ρ_1 are as given in (a) and (b).

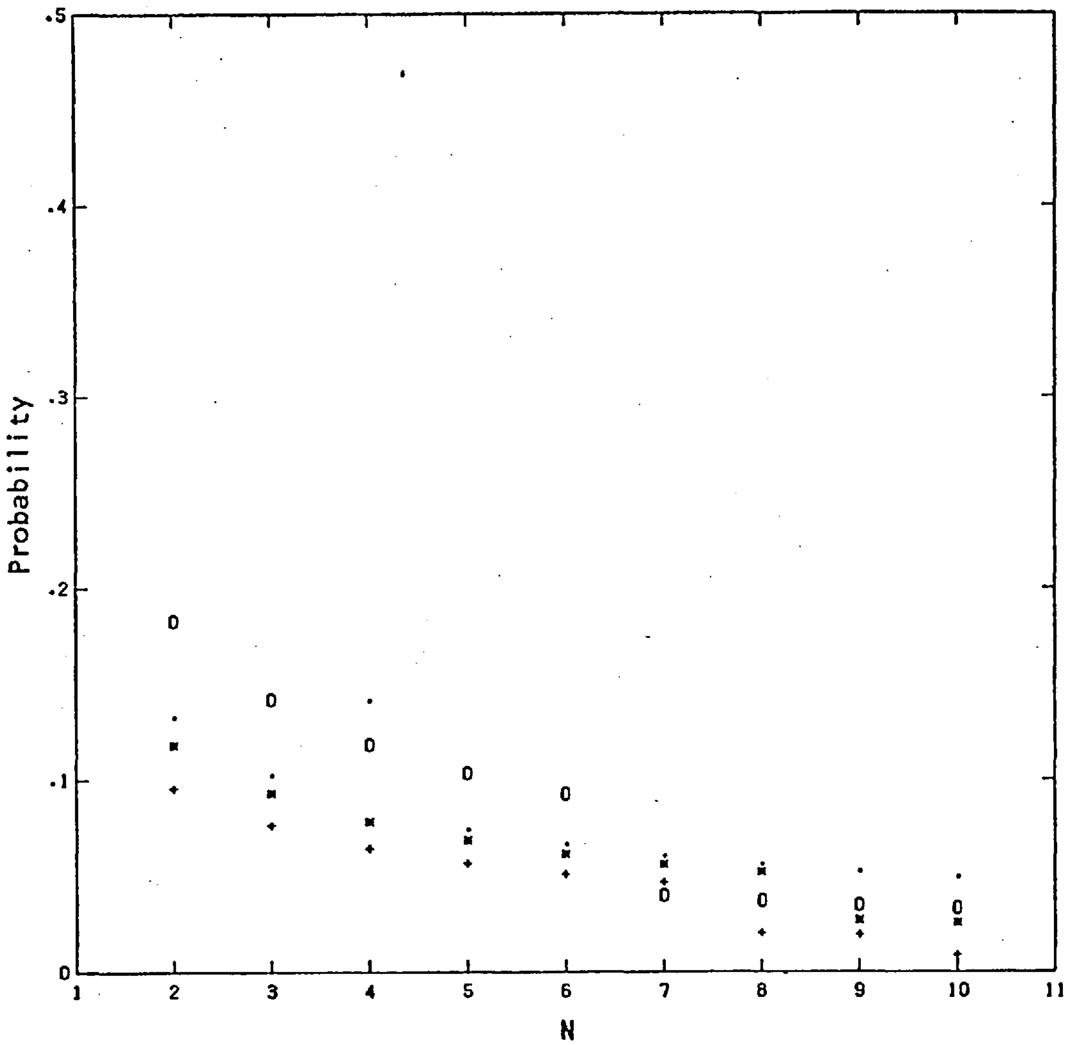


Fig 4.3.1 c The probabilities, $R_2(N;\rho_1,\rho_2)$, that the line length, L, exceeds twice its mean value in a unilevel control queue with control threshold N and service time combination C_i ($i=1,2,3,4$) when $\rho_1=1.1$ and $\rho_2=0.55$

\+ C_1 {service times are Erlang(Erlang) when L<N(L≥N)}

■ C_2 {service times are exponential(exponential) when L<N(L≥N)}

• C_3 {service times are mixed exponential(mixed exponential) when L<N(L≥N)}

0 C_4 {service times are mixed exponential(Erlang) when L<N(L≥N)}

For fixed ρ_1 and ρ_2, graphs of probabilities from Tables 4.3.1 a, b, c and d have been plotted against N (cf. Figs. 4.3.1 and 4.3.2). The probabilities for C_4 frequently coincide with those of C_3 (cf. Tables 4.3.1 c and d); therefore, probabilities for the former distribution combination have been plotted only when C_4 is noticeably different from C_3. Consider first Figure 4.3.1.

The plotted points show the probabilities $R_k(N;\rho_1,\rho_2=0.55)$ when $\rho_1>1$ and k equals 1 or 2. The horizontal band on Figs. 4.3.1 a and b encloses the range of $R_1(N;\rho_1,\rho_2=0.95)$ for all C_i's and all N, ρ_1 being fixed, thereby indicating the approximate long-run proportion of time that a line exceeds its average length in any of the prescribed circumstances. For fixed k and N, the tables show that if traffic is always heavy, i.e. $\rho_1>1$, $\rho_2=0.95$, the standardized probabilities, $R_k(N;\rho_1,\rho_2)$, do not vary much among the C_i's. When $\rho_1>1$ and $\rho_2=0.55$, individual distribution combinations are more easily distinguished (cf. Fig 4.3.1).

Interpreting the probabilities $R_k(N;\rho_1,\rho_2)$ when $\rho_2\neq0.95$ is made more difficult because E(L), the respective unit of scale for each line size distribution, need not take integral values; L, however, is a discrete random variable. One example of the difficulty which this causes is particularly prominent whenever $\rho_1/\rho_2\geqq2$ $(\rho_1\neq0.5)$. The tables show that for fixed ρ_1,ρ_2 and C_i, unit increases in N usually cause E(L) to increase by less than unity. If $E(L_j)$ is the mean line length when N=j, the apparent relation between $R_1(N;\rho_1,\rho_2)$ and N (decreasing as N increases) is abruptly reversed between $R_1(j;\rho_1,\rho_2)$ and $R_1(j+1;\rho_1,\rho_2)$ if $p\leqq E(L_j)<E(L_{j+1})<p+1$ for some integer p (see, for example, Table 4.3.1 a when $\rho_1=1.5$, $\rho_2=0.55$, j=6, p=5 or Table 4.3.1 c when $\rho_1=1.1$, $\rho_2=0.55$, j=3, p=2). If the sample space for the line size process was continuous rather than discrete, or if a continuous approximation to the distribution of L could be devised, this difficulty obviously would not arise. However, taking the above difficulty of interpretation into account, Tables 4.3.1 a, b, c and d and

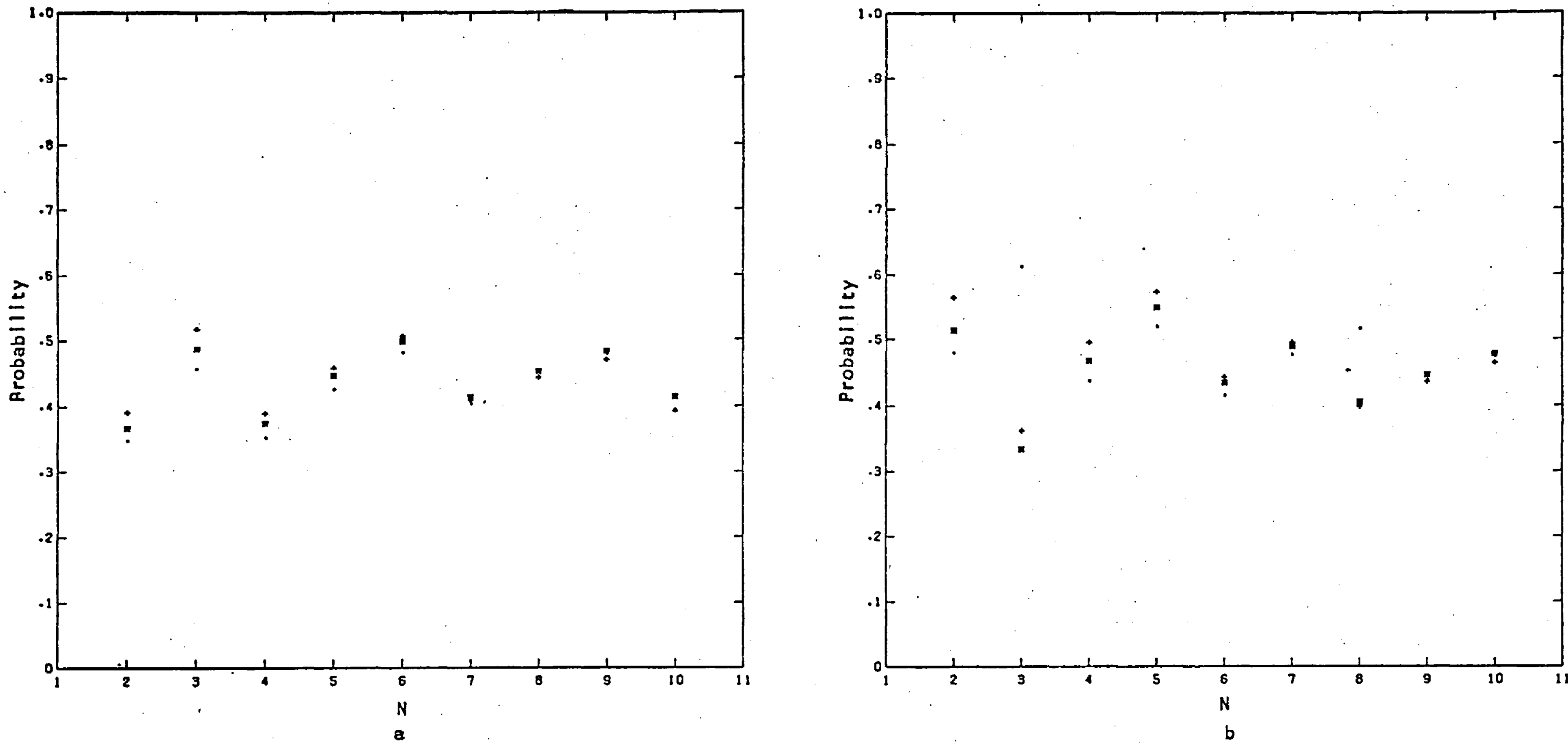


Fig. 4.3.2 Probabilities, $R_k(N;\rho_1,\rho_2)$, that the line length, L, exceeds k times its mean value in a unilevel control queue with control threshold N and service time combination C_i $(i=1,2,3)$. In (a) $k=1$, $\rho_1=0.9$, $\rho_2=0.55$; In (b) $k=1$, $\rho_1=0.9$, $\rho_2=0.15$.

\+ C_1 (service times are Erlang(Erlang) when $L<N$ $(L\geq N)$)

■ C_2 (service times are exponential(exponential) when $L<N$ $(L\geq N)$)

• C_3 (service times are mixed exponential(mixed exponential) when $L<N$ $(L\geq N)$)

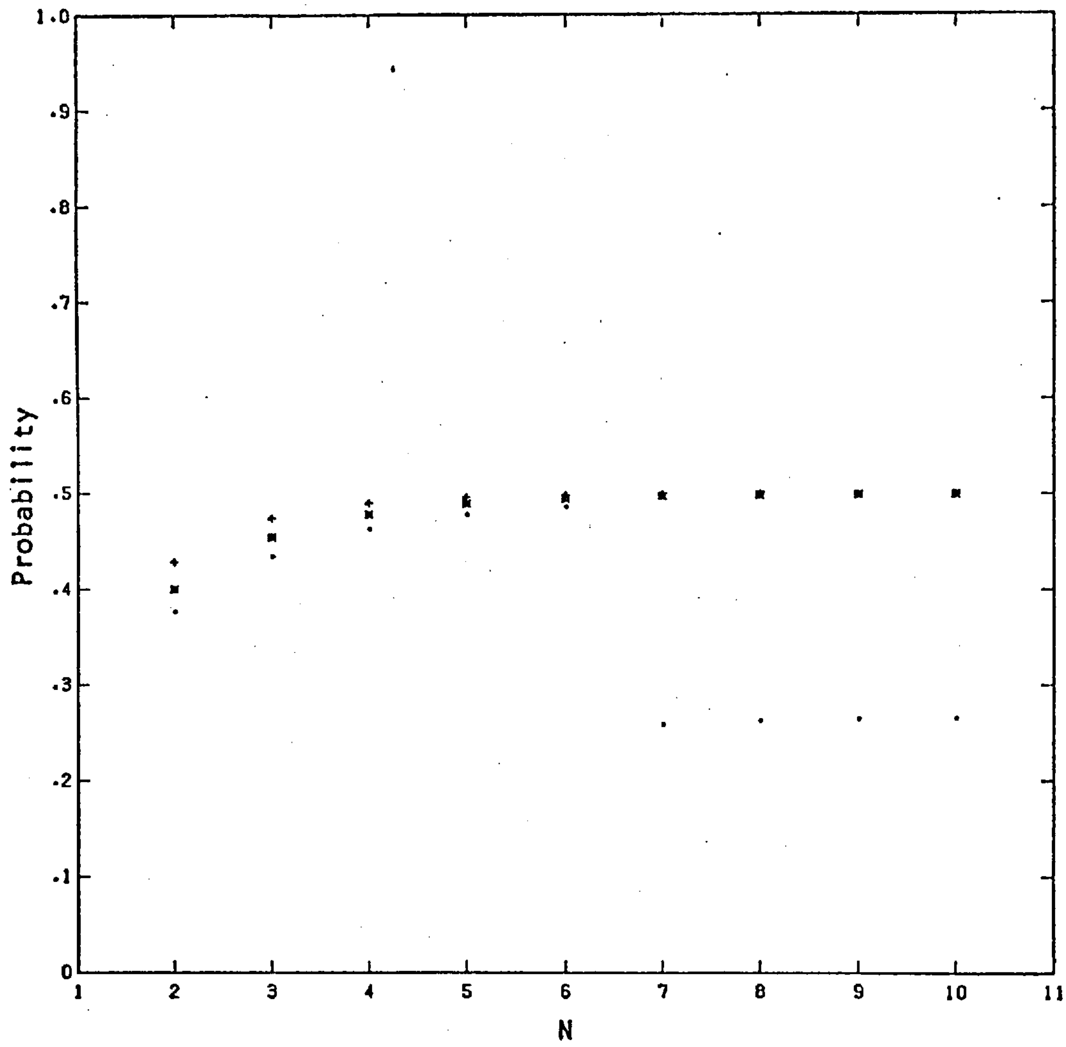


Fig. 4.3.2 c The probabilities, $R_1(N;\rho_1,\rho_2)$, that the line length, L, exceeds its mean value in a unilevel control queue with control threshold N and service time combination C_i $(i=1,2,3)$ when $\rho_1=0.5$ and $\rho_2=0.25$.

+ C_1 {service times are Erlang(Erlang) when $L<N(L\geq N)$}

■ C_2 {service times are exponential(exponential) when $L<N(L\geq N)$}

• C_3 {service times are mixed exponential(mixed exponential) when $L<N(L\geq N)$}

Figs. 4.3.1, 4.3.2 suggest that $R_1(N;\rho_1,\rho_2)$ decreases as N increases. The effect of the abrupt changes mentioned earlier is clearly illustrated in Fig. 4.3.1 a. For each C_i, the general decrease in $R_1(N;\rho_1,\rho_2)$ combines with one sudden increase to partition the control levels 2,...,10, naturally, into two distinct sets on which $R_1(N;\rho_1,\rho_2)$ is a decreasing function of N.

Since the inequality $p \leqq kE(L_j) < kE(L_{j+1}) < p+1$ for some integer p is satisfied less frequently when k is 2 or 3, it is more apparent from Tables 4.3.1 a, b, c and d that, for fixed ρ_1,ρ_2 and C_i ($\rho_1 \neq 0.5$), $R_2(N;\rho_1,\rho_2)$ and $R_3(N;\rho_1,\rho_2)$ both decrease as N increases (cf. Fig. 4.3.1 c).

The case $\rho_1=0.5$ remains to be interpreted. Two aspects of the results for this case require explanation (cf. Fig. 4.3.2 c). The probabilities $R_k(N;\rho_1=0.5,\rho_2)$ converge very quickly as N increases. If $\rho_1=0.5$, a change from $G_1(\cdot)$ to $G_2(\cdot)$ occurs infrequently unless N is typically 2 or 3; thus, nearly every customer's service time is an observation from $G_1(\cdot)$. For non-unilevel control queues with traffic intensity 0.5 and service times specified by the distribution $G_1(\cdot)$ from combination C_i, simple calculations show that lines exceeding k times the value of E(L) given in the table for C_i (k=1,2,3) occur with probabilities which are approximately the tabulated limiting values.

The tabulated distributions also show that when N is seven or more, the $R_k(N;\rho_1=0.5,\rho_2)$'s for C_1 and C_2 are considerably larger than the corresponding values for C_3 and C_4. The mean values for these cases reflect the fact that the choice of $G_1(\cdot)$ for C_3 and C_4 (mixed exponential) is overdispersed with respect to the choice of $G_1(\cdot)$ for C_1 and C_2 (Erlang or exponential, respectively). The differences in E(L) among the C_i's is quite small. However, since L is a discrete random variable, the definition of $R_k(N;\rho_1,\rho_2)$ exaggerates small differences in E(L) among similar line size distributions when the respective mean values happen to bracket an integer. This is the single reason, in this case, for the very different values of $R_k(N;\rho_1=0.5,\rho_2)$ $(k=1,2,3)$ given in the tables.

To explore further the effect of unilevel control on the distribution of the line length we now compare line size distributions in similar circumstances for queues with and without unilevel control. We begin by defining an overall traffic intensity, ρ, for any unilevel control queue.

It is a well-known property of the M/G/1 queue that if p_o is the equilibrium probability that the system is empty, $1-p_o$ is the long-run proportion of time that the server is busy. This property is one which unilevel control does not affect. Therefore, we define ρ, the overall traffic intensity in a unilevel control queue, to be $1-p_o$. According to (4.2.26), p_o depends on N, ρ_1 and ρ_2, the three essential features of unilevel control; hence $\rho=\rho(N;\rho_1,\rho_2)$.

We also assume that when unilevel control is introduced into a queueing system the mathematical form of the existing service time distribution is retained in choosing $G_1(\cdot)$ and $G_2(\cdot)$. With this assumption and the above definition of $\rho=\rho(N;\rho_1,\rho_2)$ we can specify queueing systems without unilevel control which correspond to those with distribution combinations C_1, C_2 and C_3 for all values of N, ρ_1 and ρ_2; C_4 is excluded because $G_1(\cdot)$ and $G_2(\cdot)$ have different mathematical forms. Call the service time distributions in these queueing systems without unilevel control C_1', C_2' and C_3', respectively, where C_1': $g(x)=\mu^2xe^{-\mu x}$, C_2': $g(x)=\nu e^{-\nu x}$, and C_3': $g(x)=(1-p)\alpha e^{-\alpha x}+p\beta e^{-\beta x}$, $0<p<1$. To ensure the closest correspondence between C_3 and C_3' we fix $p=\frac{1}{2}$ and $\alpha=3\beta$.

Call L^- the line size process in a queueing system without unilevel control. A unilevel control queue with service time distribution combination C_i and parameters N, ρ_1 and ρ_2 will be said to correspond to a queueing system without unilevel control if the latter system has service time distribution C_i' and traffic intensity $\rho=\rho(N;\rho_1,\rho_2)$, the overall traffic intensity in the given unilevel control queue. To determine how unilevel control affects the line size distribution in a queueing system, we now calculate $R_k(\rho)$, the probability that line length, L^-, in a non-unilevel control queue with traffic intensity ρ exceeds k times mean line length, E(L), in the corresponding unilevel control queue with overall traffic intensity $\rho=\rho(N;\rho_1,\rho_2)$, i.e. $R_k(\rho)=pr\{L^->kE(L)|\rho=\rho(N;\rho_1,\rho_2)\}$.

For each unilevel control situation previously considered (cf. Tables 4.3.1 a, b and c), values of $R_k(\rho)$ for the corresponding queue without unilevel control have been calculated. The results are given in Tables 4.3.2 a, b and c. The correspondence between $R_k(\rho)$ and $R_k(N;\rho_1,\rho_2)$ for fixed C_i' and C_i (i=1,2,3) is indicated by identifying entries in Tables 4.3.1 and 4.3.2 by the same values of N, ρ_1 and ρ_2. The overall traffic intensity, $\rho=\rho(N;\rho_1,\rho_2)$, is also given in Table 4.3.2.

Entry by entry comparison of Tables 4.3.1 and 4.3.2 shows the effect of unilevel control on the line size distributions for corresponding queueing systems. Figs. 4.3.3 a, b and c illustrate some of these comparisons; for fixed k, N, ρ_1 and ρ_2, each graph shows the respective ranges, for all C_i and C_i' (i=1,2,3) of the probabilities $R_k(N;\rho_1,\rho_2)$ and $R_k(\rho)$.

Comparing corresponding tables for $R_k(N;\rho_1,\rho_2)$ and $R_k(\rho)$ underlines the short-tailed aspect of the line size distribution for unilevel control in contrast to the line size distribution for the queue without unilevel control. For corresponding queues, unilevel control often produces two and three-fold reductions in the probability of lines exceeding the same length, k times E(L); larger reductions can also be found. By providing faster service when the line is longer than N-1, unilevel control acts automatically to control the line length. This automatic action modifies the distribution of line length by redistributing much of the probability originally associated with lines longer than N amongst the states 0,..., N-1. For unilevel control situations with $\rho_1>1$, lines with N-2, N-1 and N customers are probably the most frequently occurring in the system; the effect, in this case, is very similar to that of industrial feedback control.

	$\rho(N;\rho_1,\rho_2)$										$R_k(\rho)$							
	ρ_1=1.5		ρ_1=1.1		ρ_1=0.9		ρ_1=0.5				ρ_1=1.5		ρ_1=1.1		ρ_1=0.9		ρ_1=0.5	
	ρ_2=.55	ρ_2=.95	ρ_2=.55	ρ_2=.95	ρ_2=.15	ρ_2=.55	ρ_2=.05	ρ_2=.25			ρ_2=.55	ρ_2=.95	ρ_2=.55	ρ_2=.95	ρ_2=.15	ρ_2=.55	ρ_2=.05	ρ_2=.25
										1	0.644	0.607	0.538	0.505	0.565	0.468	0.372	0.429
	0.821	0.976	0.757	0.966	0.565	0.710	0.372	0.429	2	2	0.385	0.376	0.262	0.251	0.284	0.302	0.372	0.157
										3	0.229	0.233	0.182	0.125	0.284	0.124	0.117	0.157
										1	0.719	0.761	0.532	0.553	0.448	0.559	0.454	0.474
	0.911	0.987	0.836	0.972	0.696	0.770	0.454	0.474	3	2	0.562	0.576	0.332	0.314	0.282	0.284	0.178	0.195
										3	0.389	0.435	0.207	0.171	0.176	0.143	0.178	0.077
										1	0.789	0.849	0.635	0.607	0.543	0.464	0.482	0.490
	0.953	0.993	0.880	0.976	0.760	0.804	0.482	0.490	4	2	0.650	0.718	0.382	0.364	0.267	0.261	0.202	0.209
										3	0.536	0.607	0.272	0.219	0.186	0.147	0.081	0.086
										1	0.848	0.906	0.624	0.636	0.603	0.509	0.493	0.496
	0.974	0.996	0.908	0.980	0.797	0.826	0.493	0.496	5	2	0.736	0.819	0.424	0.411	0.334	0.239	0.212	0.215
										3	0.639	0.744	0.288	0.259	0.184	0.144	0.087	0.089
										1	0.894	0.942	0.690	0.662	0.499	0.543	0.497	0.498
	0.985	0.998	0.927	0.983	0.821	0.841	0.497	0.498	6	2	0.810	0.886	0.461	0.446	0.297	0.274	0.216	0.217
										3	0.734	0.836	0.308	0.300	0.177	0.138	0.090	0.091
										1	0.939	0.964	0.685	0.701	0.535	0.461	0.499	0.499
	0.992	0.999	0.941	0.985	0.838	0.853	0.499	0.499	7	2	0.878	0.931	0.495	0.489	0.266	0.245	0.218	0.218
										3	0.821	0.898	0.330	0.340	0.132	0.130	0.091	0.091
										1	0.958	0.978	0.687	0.723	0.455	0.484	0.500	0.500
	0.995	0.999	0.951	0.987	0.850	0.861	0.500	0.500	8	2	0.916	0.958	0.493	0.530	0.239	0.268	0.218	0.219
										3	0.876	0.937	0.354	0.381	0.125	0.121	0.092	0.092
										1	0.972	0.987	0.732	0.744	0.479	0.503	0.500	0.500
	0.997	0.999	0.960	0.989	0.859	0.868	0.500	0.500	9	2	0.944	0.974	0.527	0.560	0.262	0.237	0.219	0.219
										3	0.917	0.961	0.379	0.421	0.118	0.112	0.092	0.092
										1	0.982	0.992	0.736	0.774	0.498	0.433	0.500	0.500
	0.998	0.999	0.966	0.990	0.866	0.873	0.500	0.500	10	2	0.964	0.985	0.533	0.596	0.233	0.211	0.219	0.219
										3	0.946	0.977	0.405	0.459	0.132	0.103	0.092	0.092

Table 4.3.2 a Probabilities, $R_k(\rho)$, that in a non-unilevel control queue with traffic intensity, ρ, and C_1' (Erlang) service times, the line length, L^-, exceeds k times mean line length in the corresponding unilevel control queue with overall traffic intensity $\rho=\rho(N;\rho_1,\rho_2)$.

$\rho(N;\rho_1,\rho_2)$										$R_k(\rho)$							
ρ_1=1.5		ρ_1=1.1		ρ_1=0.9		ρ_1=0.5		N	k	ρ_1=1.5		ρ_1=1.1		ρ_1=0.9		ρ_1=0.5	
ρ_2=.55	ρ_2=.95	ρ_2=.55	ρ_2=.95	ρ_2=.15	ρ_2=.55	ρ_2=.05	ρ_2=.25			ρ_2=.55	ρ_2=.95	ρ_2=.55	ρ_2=.95	ρ_2=.15	ρ_2=.55	ρ_2=.05	ρ_2=.25
									1	0.592	0.519	0.504	0.411	0.514	0.444	0.345	0.400
0.769	0.968	0.710	0.957	0.514	0.667	0.345	0.400	2	2	0.350	0.278	0.254	0.177	0.265	0.296	0.345	0.160
									3	0.207	0.145	0.180	0.076	0.265	0.132	0.119	0.160
									1	0.651	0.653	0.495	0.461	0.422	0.533	0.433	0.455
0.867	0.979	0.791	0.962	0.650	0.730	0.433	0.455	3	2	0.489	0.427	0.310	0.221	0.274	0.284	0.187	0.207
									3	0.318	0.279	0.194	0.102	0.178	0.151	0.187	0.094
									1	0.711	0.746	0.594	0.507	0.518	0.455	0.469	0.478
0.918	0.986	0.841	0.967	0.720	0.769	0.469	0.478	4	2	0.511	0.557	0.352	0.257	0.373	0.269	0.220	0.229
									3	0.427	0.421	0.249	0.130	0.193	0.159	0.103	0.109
									1	0.809	0.817	0.582	0.534	0.581	0.504	0.485	0.489
0.948	0.991	0.873	0.971	0.763	0.796	0.485	0.489	5	2	0.655	0.674	0.388	0.293	0.338	0.254	0.235	0.240
									3	0.530	0.551	0.258	0.161	0.197	0.161	0.114	0.117
									1	0.844	0.870	0.647	0.574	0.495	0.542	0.493	0.495
0.967	0.994	0.897	0.974	0.791	0.815	0.493	0.495	6	2	0.713	0.761	0.418	0.329	0.310	0.293	0.243	0.245
									3	0.602	0.666	0.270	0.194	0.194	0.159	0.119	0.121
									1	0.877	0.908	0.639	0.597	0.535	0.474	0.496	0.497
0.978	0.996	0.914	0.977	0.812	0.830	0.496	0.497	7	2	0.769	0.827	0.446	0.365	0.286	0.270	0.246	0.247
									3	0.689	0.754	0.312	0.223	0.188	0.154	0.122	0.123
									1	0.904	0.935	0.687	0.632	0.468	0.500	0.498	0.499
0.986	0.997	0.928	0.979	0.827	0.841	0.498	0.499	8	2	0.830	0.877	0.472	0.400	0.265	0.250	0.248	0.249
									3	0.750	0.822	0.324	0.258	0.150	0.148	0.124	0.124
									1	0.927	0.955	0.682	0.652	0.495	0.521	0.499	0.499
0.991	0.998	0.938	0.982	0.839	0.850	0.499	0.499	9	2	0.868	0.913	0.466	0.433	0.292	0.272	0.249	0.249
									3	0.805	0.873	0.339	0.288	0.145	0.142	0.124	0.125
									1	0.945	0.969	0.682	0.683	0.517	0.462	0.500	0.500
0.994	0.999	0.947	0.984	0.848	0.857	0.500	0.500	10	2	0.899	0.939	0.492	0.466	0.268	0.249	0.250	0.250
									3	0.850	0.911	0.354	0.318	0.138	0.134	0.125	0.125

Table 4.3.2 b Probabilities, $R_k(\rho)$, that in a non-unilevel control queue with traffic intensity, ρ, and C_2^1 (exponential) service times, the line length, $\bar{L}$, exceeds k times mean line length in the corresponding unilevel control queue with overall traffic intensity $\rho=\rho(N;\rho_1,\rho_2)$.

$\rho(N;\rho_1,\rho_2)$										$R_k(\rho)$							
$\rho_1=1.5$		$\rho_1=1.1$		$\rho_1=0.9$		$\rho_1=0.5$		N	k	$\rho_1=1.5$		$\rho_1=1.1$		$\rho_1=0.9$		$\rho_1=0.5$	
$\rho_2=.55$	$\rho_2=.95$	$\rho_2=.55$	$\rho_2=.95$	$\rho_2=.15$	$\rho_2=.55$	$\rho_2=.05$	$\rho_2=.25$			$\rho_2=.55$	$\rho_2=.95$	$\rho_2=.55$	$\rho_2=.95$	$\rho_2=.15$	$\rho_2=.55$	$\rho_2=.05$	$\rho_2=.25$
									1	0.567	0.472	0.484	0.370	0.483	0.430	0.325	0.379
0.739	0.962	0.680	0.950	0.483	0.638	0.325	0.379	2	2	0.345	0.226	0.257	0.145	0.253	0.206	0.325	0.160
									3	0.213	0.111	0.189	0.057	0.253	0.143	0.119	0.160
									1	0.616	0.582	0.473	0.395	0.615	0.510	0.415	0.437
0.836	0.973	0.758	0.954	0.615	0.699	0.415	0.437	3	2	0.461	0.342	0.303	0.166	0.401	0.283	0.190	0.210
									3	0.346	0.201	0.195	0.070	0.268	0.159	0.190	0.105
									1	0.668	0.669	0.565	0.432	0.495	0.443	0.455	0.466
0.892	0.980	0.810	0.958	0.688	0.740	0.455	0.466	4	2	0.507	0.451	0.402	0.196	0.363	0.272	0.226	0.236
									3	0.385	0.309	0.242	0.086	0.198	0.168	0.117	0.125
									1	0.765	0.743	0.554	0.469	0.559	0.492	0.475	0.481
0.926	0.986	0.845	0.962	0.734	0.770	0.475	0.481	5	2	0.599	0.561	0.370	0.223	0.336	0.322	0.246	0.251
									3	0.469	0.423	0.247	0.109	0.205	0.172	0.132	0.136
									1	0.798	0.802	0.616	0.506	0.485	0.531	0.486	0.489
0.949	0.990	0.871	0.966	0.765	0.791	0.486	0.489	6	2	0.647	0.651	0.396	0.259	0.315	0.302	0.257	0.260
									3	0.548	0.528	0.285	0.132	0.205	0.173	0.142	0.143
									1	0.831	0.854	0.608	0.527	0.525	0.473	0.263	0.265
0.964	0.993	0.891	0.969	0.788	0.808	0.492	0.494	7	2	0.719	0.732	0.420	0.288	0.296	0.284	0.145	0.147
									3	0.604	0.626	0.291	0.157	0.202	0.171	0.082	0.083
									1	0.860	0.890	0.655	0.560	0.557	0.501	0.266	0.267
0.975	0.995	0.907	0.972	0.806	0.821	0.496	0.497	8	2	0.761	0.793	0.442	0.317	0.331	0.267	0.148	0.149
									3	0.674	0.706	0.299	0.179	0.197	0.167	0.084	0.084
									1	0.886	0.917	0.649	0.579	0.497	0.524	0.268	0.269
0.982	0.996	0.919	0.975	0.820	0.832	0.497	0.498	9	2	0.801	0.842	0.463	0.345	0.308	0.291	0.150	0.150
									3	0.724	0.773	0.330	0.206	0.163	0.162	0.085	0.085
									1	0.908	0.938	0.687	0.609	0.521	0.473	0.269	0.270
0.987	0.997	0.929	0.977	0.831	0.841	0.499	0.499	10	2	0.837	0.880	0.483	0.374	0.287	0.272	0.151	0.151
									3	0.771	0.827	0.339	0.234	0.159	0.156	0.086	0.086

Table 4.3.2 c   Probabilities, $R_k(\rho)$, that in a non-unilevel control queue with traffic intensity , ρ, and C_3^1 (mixed exponential) service times, the line length, L^-, exceeds k times mean line length in the corresponding unilevel control queue with overall traffic intensity $\rho=\rho(N;\rho_1,\rho_2)$.

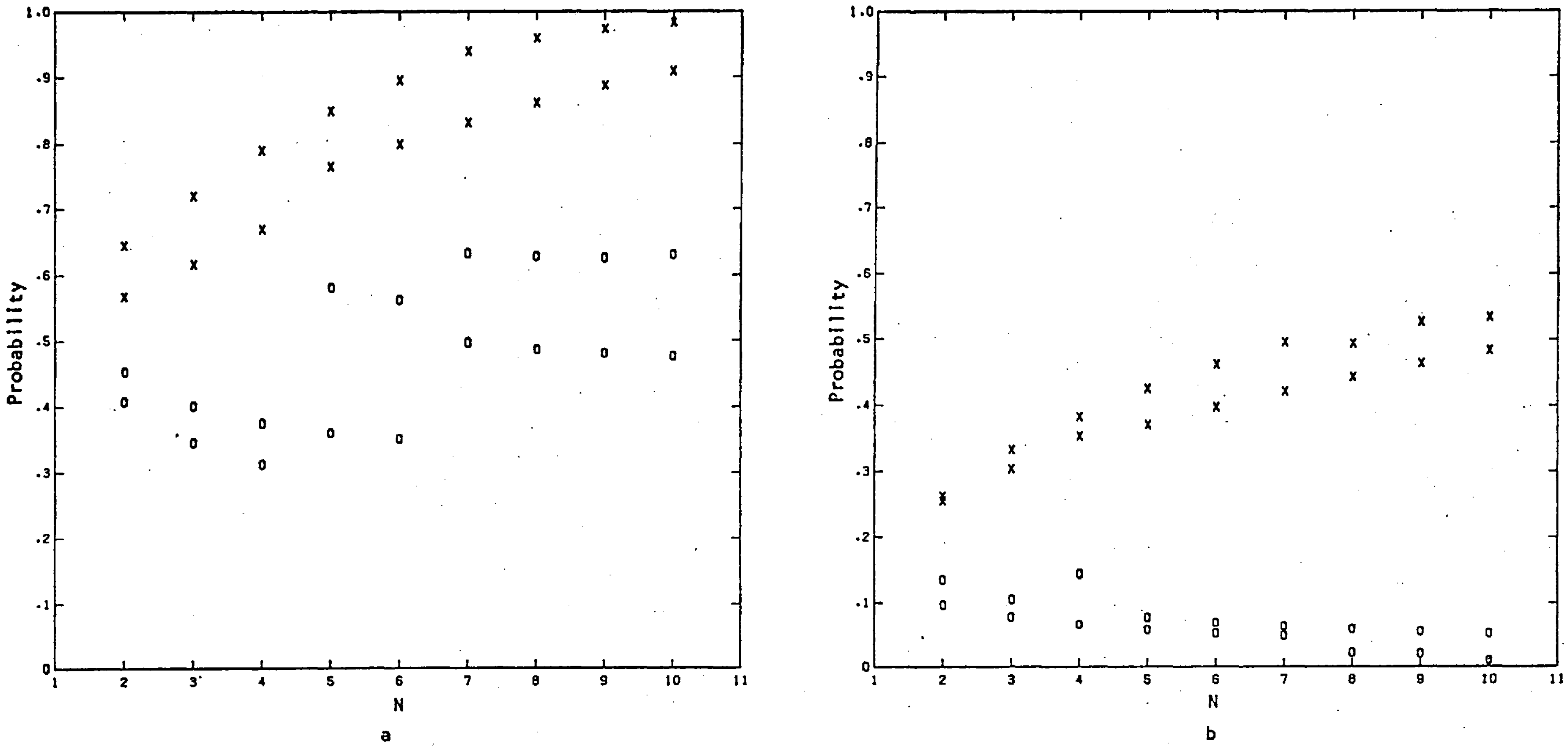


Fig. 4.3.3 The range of probabilities that line length, L or L^-, exceeds k times mean unilevel control line length; L $\left(L^-\right)$ is the line length in a unilevel control (non-unilevel control) queue with service time combination C_i (service times C_i'), parameters N, ρ_1, ρ_2 and traffic intensity $\rho(N;\rho_1,\rho_2)$ $\left(\rho=\rho(N;\rho_1,\rho_2)\right)$, for $i=1,2,3$. In (a) $k=1$, $\rho_1=1.5$, $\rho_2=0.55$; In (b) $k=2$, $\rho_1=1.1$ and $\rho_2=0.55$.

X
X The range of probabilities for unilevel control queues

0
0 The range of probabilities for non-unilevel control queues

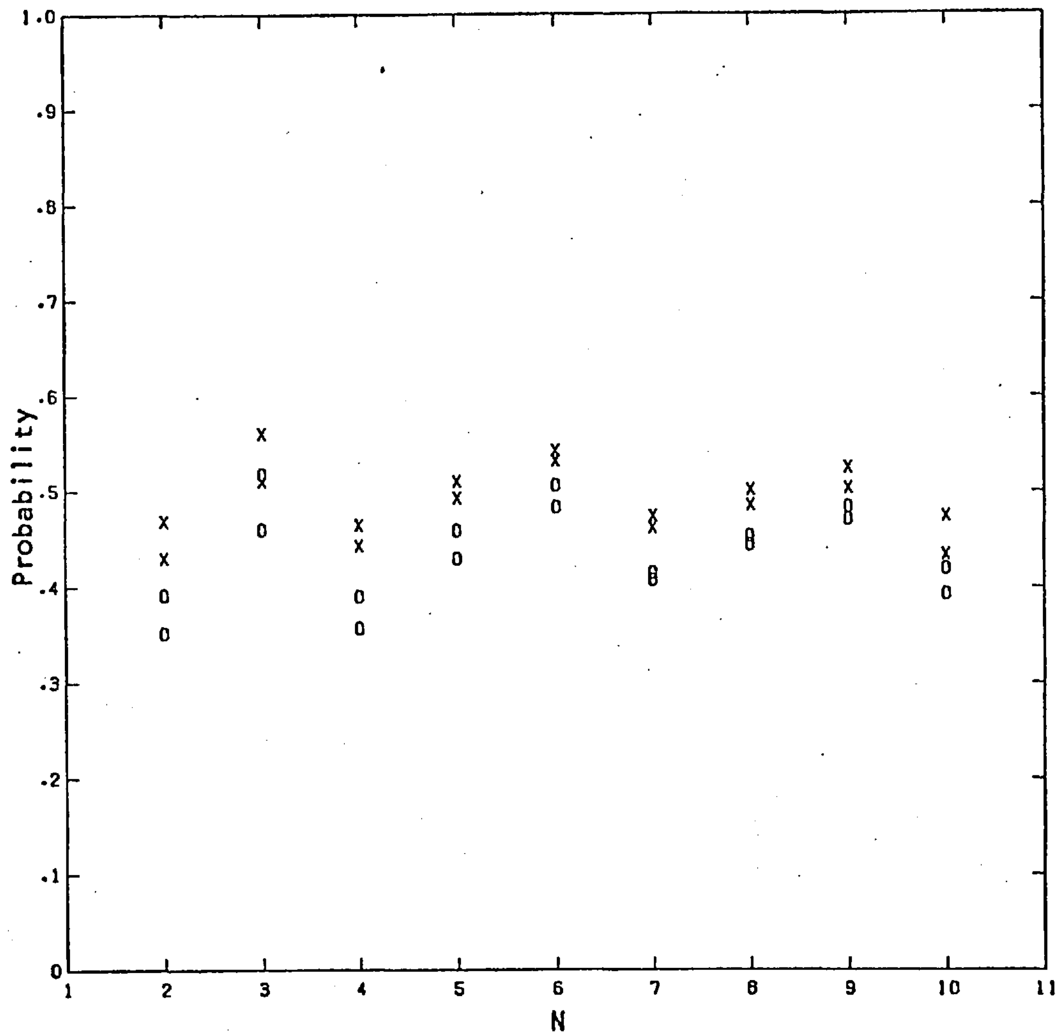


Fig. 4.3.3 c The range of probabilities that line length, L or L^-, exceeds mean unilevel control line length; L (L^-) is the line length in a unilevel control (non-unilevel control) queue with service time combination C_i (service times C_i'), parameters N, $\rho_1=0.9$, $\rho_2=0.55$ and traffic intensity $\rho(N;\rho_1,\rho_2)$ $(\rho=\rho(N;\rho_1,\rho_2))$, for i=1,2 or 3.

X The range of probabilities for unilevel control queues with service
X time combinations C_1, C_2 or C_3.

O The range of probabilities for non-unilevel control queues with service
O times C_1', C_2' or C_3'.

When both ρ_1 and ρ_2 are less than unity, the overall effect of unilevel control on the distribution of line size in a queueing system is less marked (cf. Fig. 4.3.3 c). However, comparing Tables 4.3.1 and 4.3.2 in this case shows that $R_k(N;\rho_1,\rho_2)$ is always less than $R_k(\rho)$ for the corresponding queue without unilevel control. When $\rho_1=0.9$, lines exceeding two and three times E(L) occur at least twice as frequently in queueing systems without unilevel control as do lines of the same length in corresponding unilevel control queues. However, lines longer than E(L) appear to be only slightly less probable for unilevel control than for the corresponding queue without unilevel control. The entries in Tables 4.3.1 and 4.3.2 for $\rho_1=0.5$ show that in these moderate traffic conditions, unle-unless N is typically 2 or 3, unilevel control hardly affects the line size distribution.

The results indicate that in queueing situations which are heavily or very heavily congested, unilevel service process control is an effective means of reducing mean line length and the probability of lines which are long relative to E(L), the reduced mean line length. In particular, unilevel control can be used to manage queueing situations which otherwise cannot be expected to reach an equilibrium. There is some evidence (cf. Figs. 4.3.1 a and b) to suggest that when service times vary considerably, unilevel control is more effective in regulating the line length. This distinction is less important, however, than the degree to which changes in the line size distribution are affected by the choice of N. For fixed ρ_1, ρ_2 and C_i the mean line length, E(L), is smallest when N=2; however, the disruptive effects of changing from slow to faster service probably decrease as N increases. In any practical situation both aspects of unilevel control should be considered in choosing the control threshold.

4.4 Bilevel hysteresis control of the service process

If the cost of changing from one service time distribution to another is substantial, bilevel hysteresis control could be more suited to the situation than unilevel control. According to strict bilevel control rules, a change from $G_2(\cdot)$ to $G_1(\cdot)$ cannot follow a change from $G_1(\cdot)$ to $G_2(\cdot)$ unless two or more service completions have occurred. This enforced delay between changes from slow to faster service and back to slower service again is one of the features distinguishing bilevel hysteresis control from unilevel control. It follows that bilevel control requires two distinct control parameters, one to determine changes from $G_1(\cdot)$ to $G_2(\cdot)$ and the second to indicate changes from $G_2(\cdot)$ to $G_1(\cdot)$ (cf. Fig. 4.1.1). Since bilevel hysteresis control is a generalization of unilevel control, we can use the discussion of §4.2 as a guide in analyzing a similar model for bilevel hysteresis control of the M/G/1 queueing process. Unless otherwise indicated, the notation and assumptions of §4.2 will not be changed.

Under bilevel hysteresis control, customers may be served according to one of two service time distributions, $G_1(\cdot)$ or $G_2(\cdot)$; the decision points of the control process are the arrival and service epochs. The choice of service time distribution depends both on the number of customers present and on the immediate history of the process. Two control parameters, r and R ($r<R$), are required. The following rules determine the service time distribution for the customer currently in service:

(i) when $L_t \leq r$ the customer is served according to the distribution $G_1(\cdot)$,

(ii) when $L_t \geq R$ the customer is served according to the distribution $G_2(\cdot)$,

(iii) when $r<L_t<R$ the service time distribution is not changed

(iv) if L_t increases from R-1 to R while a customer, C, is being served according to the distribution $G_1(\cdot)$, immediately terminate service to C and begin a new service time for the same customer according to the distribution $G_2(\cdot)$.

As in the case of unilevel control, rule (iv) serves to simplify the analysis of the resulting line size process.

Since L_t is generally non-Markov we redefine the state of the system by adjoining two supplementary variables, S_t and Z_t. The former supplementary variable, S_t, is the elapsed service time at time t; Z_t is an indicator function which takes the value j when $G_j(\cdot)$ (j=1,2) is the service time distribution in use at time t.

Let $p_o(t)=pr(L_t=0)$ $(t\geq 0)$, $p_n(t,x;1)=pr(L_t=n,S_t=x;Z_t=1)$ $(n=1,\ldots,R-1;$ $x\geq 0;\ t\geq 0)$ and $p_n(t,x;2)=pr(L_t=n,S_t=x;Z_t=2)$ $(n=r+1,r+2,\ldots;\ x\geq 0;\ t\geq 0)$. Time-dependent Kolmogorov forward differential equations for bilevel hysteresis control are given by

$$\frac{\partial}{\partial t}p_0(t)+\lambda p_0(t)=\int_0^\infty p_1(t,x;1)\phi_1(x)\,dx\ , \qquad (4.4.1)$$

$$\frac{\partial}{\partial t}p_1(t,x;1)+\frac{\partial}{\partial x}p_1(t,x;1)+\{\lambda+\phi_1(x)\}p_1(t,x;1)=0\ , \qquad (4.4.2)$$

$$\frac{\partial}{\partial t}p_j(t,x;1)+\frac{\partial}{\partial x}p_j(t,x;1)+\{\lambda+\phi_1(x)\}p_j(t,x;1)=\lambda p_{j-1}(t,x;1)\ ,\ (j=2,\ldots,R-1) \qquad (4.4.3)$$

$$\frac{\partial}{\partial t}p_{r+1}(t,x;2)+\frac{\partial}{\partial x}p_{r+1}(t,x;2)+\{\lambda+\phi_2(x)\}p_{r+1}(t,x;2)=0\ , \qquad (4.4.4)$$

$$\frac{\partial}{\partial t}p_j(t,x;2)+\frac{\partial}{\partial x}p_j(t,x;2)+\{\lambda+\phi_2(x)\}p_j(t,x;2)=\lambda p_{j-1}(t,x;2)\ ,\ (j=r+2,r+3,\ldots) \qquad (4.4.5)$$

Solutions for (4.4.1) to (4.4.5) must also satisfy the boundary conditions

$$p_1(t,0;1)=\lambda p_0(t)+\int_0^\infty p_2(t,x;1)\phi_1(x)\,dx\ , \qquad (4.4.6)$$

$$p_j(t,0;1)=\int_0^\infty p_{j+1}(t,x;1)\phi_1(x)\,dx\ ,\ (j=2,\ldots,r-1,r+1,\ldots,R-2) \qquad (4.4.7)$$

$$p_r(t,0;1)=\int_0^\infty p_{r+1}(t,x;1)\phi_1(x)\,dx+\int_0^\infty p_{r+1}(t,x;2)\phi_2(x)\,dx\ , \qquad (4.4.8)$$

$$p_{R-1}(t,0;1)=0\ , \qquad (4.4.9)$$

$$p_j(t,0;2)=\int_0^\infty p_{j+1}(t,x;2)\phi_2(x)\,dx\ ,\ (j=r+1,\ldots,R-1,R+1,R+2,\ldots) \qquad (4.4.10)$$

$$p_R(t,0;2)=\lambda\int_0^\infty p_{R-1}(t,x;1)\,dx+\int_0^\infty p_{R+1}(t,x;2)\phi_2(x)\,dx. \qquad (4.4.11)$$

Let p_o, $p_n(x;1)$ $(n=1,\ldots,R-1;\ x\geq 0)$ and $p_n(x;2)$ $(n=r+1,r+2,\ldots;\ x\geq 0)$ be the steady-state analogues of $p_o(t)$, $p_n(t,x;1)$ and $p_n(t,x;2)$ respectively. Equilibrium equations corresponding to (4.4.1) to (4.4.11) are given by

$$\lambda p_o = \int_0^\infty p_1(x;1)\,\phi_1(x)\,dx\ , \tag{4.4.12}$$

$$\frac{\partial}{\partial x} p_1(x;1) + \{\lambda + \phi_1(x)\}\, p_1(x;1) = 0, \tag{4.4.13}$$

$$\frac{\partial}{\partial x} p_j(x;1) + \{\lambda + \phi_1(x)\}\, p_j(x;1) = \lambda p_{j-1}(x;1)\ , \qquad (j=2,\cdots,R-1) \tag{4.4.14}$$

$$\frac{\partial}{\partial x} p_{r+1}(x;2) + \{\lambda + \phi_2(x)\}\, p_{r+1}(x;2) = 0 \tag{4.4.15}$$

$$\frac{\partial}{\partial x} p_j(x;2) + \{\lambda + \phi_2(x)\}\, p_j(x;2) = \lambda p_{j-1}(x;2)\ ,\ (j=r+2,r+3,\ldots) \tag{4.4.16}$$

$$p_1(0;1) = \lambda p_o + \int_0^\infty p_2(x;1)\,\phi_1(x)\,dx, \tag{4.4.17}$$

$$p_j(0;1) = \int_0^\infty p_{j+1}(x;1)\,\phi_1(x)\,dx\ ,\ (j=2,\cdots,r-1,r+1,\cdots,R-2) \tag{4.4.18}$$

$$p_r(0;1) = \int_0^\infty p_{r+1}(x;1)\,\phi_1(x)\,dx + \int_0^\infty p_{r+1}(x;2)\,\phi_2(x)\,dx\ , \tag{4.4.19}$$

$$p_{R-1}(0;1) = 0\ , \tag{4.4.20}$$

$$p_j(0;2) = \int_0^\infty p_{j+1}(x;2)\,\phi_2(x)\,dx\ ,\ (j=r+1,\cdots,R-1,R+1,R+2,\cdots) \tag{4.4.21}$$

$$p_R(0;2) = \lambda\int_0^\infty p_{R-1}(x;1)\,dx + \int_0^\infty p_{R+1}(x;2)\,\phi_2(x)\,dx\ . \tag{4.4.22}$$

The general solution for (4.4.13) and (4.4.14) is

$$p_j(x;1) = \sum_{n=0}^{j-1} p_{j-n}(0;1)\,\frac{(\lambda x)^n}{n!}\,e^{-\lambda x}\,\mathcal{G}_1(x)\ ,\quad (j=1,\cdots,R-1)\ . \tag{4.4.23}$$

Define the probability generating functions $P_1(x;z)=\sum_{k=1}^{R-1} p_k(x;1)z^k$ and $P_2(x;z)=\sum_{k=r+1}^{\infty} p_k(x;2)z^k$ $(|z|\leq 1)$. Using $P_2(x;z)$ we can combine (4.4.15) and (4.4.16) in the single equation

$$\frac{\partial}{\partial x} P_2(x;z) = \{\lambda z - \lambda - \phi_2(x)\}\, P_2(x;z)$$

which has the solution

$$P_2(x;z) = P_2(0;z)\, e^{-\lambda x(1-z)}\, \mathcal{G}_2(x) \quad , \tag{4.4.24}$$

where $P_2(0;z)=\lim_{x\to 0+} P_2(x;z)$. Similarly, combine (4.4.21) and (4.4.22) in the single equation

$$P_2(0;z) = \frac{1}{z}\int_0^\infty P_2(x;z)\,\phi_2(x)\,dx - z^r\int_0^\infty p_{r+1}(x;2)\,\phi_2(x)\,dx + \lambda p_{R-1}(1)\, z^R , \tag{4.4.25}$$

where $p_j(i)=\int_0^\infty p_j(x;i)\,dx \begin{bmatrix} i=1;\ j=1,\ldots,R-1 \\ i=2;\ j=r+1,r+2,\ldots \end{bmatrix}$. By substituting (4.4.24) in (4.4.25), we can solve for $P_2(0;z)$; hence,

$$P_2(0;z) = \frac{\lambda p_{R-1}(1)\, z^{R+1} - p_{r+1}(0;2)\, g_2^*(\lambda)\, z^{r+1}}{z - g_2^*(\lambda-\lambda z)} \quad . \tag{4.2.26}$$

Since the system is in equilibrium, it follows that

$$\lambda\int_0^\infty p_{R-1}(x;1)\,dx = \int_0^\infty p_{r+1}(x;2)\,\phi_2(x)\,dx \quad .$$

Therefore,

$$P_2(x;z) = \lambda p_{R-1}(1)\,\frac{z^{R+1} - z^{r+1}}{z - g_2^*(\lambda-\lambda z)}\, e^{-\lambda x(1-z)}\, \mathcal{G}_2(x) \ . \tag{4.4.27}$$

By substituting for $p_1(x;1)$, $p_2(x;1)$ in (4.4.12) and (4.4.17) we can show that

$$p_1(0;1) = \frac{\lambda p_0}{g_1^*(\lambda)} \quad , \qquad p_2(0;1) = \frac{\lambda p_0}{\{g_1^*(\lambda)\}^2}\left\{1 - g_1^*(\lambda) + \lambda\frac{d}{d\lambda}g_1^*(\lambda)\right\} .$$

By substituting (4.4.23) in (4.4.18) we obtain the equation

$$\sum_{k=0}^{j} p_{j+1-k}(0;1)\left\{\delta_{1k} - \frac{(-\lambda)^k}{k!}\, g_1^{*(k)}(\lambda)\right\} = 0 \ , \qquad (j=2,\cdots,r-1)$$

which may be solved iteratively to obtain expressions for $p_3(0;1),\ldots,$ $p_r(0;1)$ which are unique to within p_o. To obtain expressions for $p_{r+2}(0;1)$, $\ldots,p_{R-1}(0;1)$ in terms of $p_{r+1}(0;1)$ and p_o, iteratively solve

$$\sum_{k=0}^{j} p_{j+1-k}(0;1)\left\{\delta_{1k} - \frac{(-\lambda)^k}{k!}\, g_1^{*(k)}(\lambda)\right\} = 0 \ , \qquad (j=r+1,\cdots,R-2) \ .$$

By (4.4.20) $p_{R-1}(0;1)=0$. Hence, $p_{r+1}(0;1)$ can be determined to within p_o

and the expressions for $p_{r+2}(0;1),\ldots,p_{R-2}(0;1)$ can be simplified as well.

To obtain unique solutions for (4.4.12) to (4.4.22) a normalizing condition is required. Let $P_i(z)=\int_0^\infty P_i(x;z)dx$ $(i=1,2)$. The relation which determines the unique solution of (4.4.12) to (4.4.22) is given by

$$p_o+\sum_{k=1}^{R-1}\sum_{n=0}^{k-1}p_{k-n}(0;1)\,\frac{\lambda^n}{n!}\,J_n(\lambda)+p_{R-1}(1)\,\frac{(R-r)\rho_2}{1-\rho_2}=1 \quad , \qquad (4.4.28)$$

where $J_n(\lambda)=\int_0^\infty x^n g_1(x)e^{-\lambda x}dx$, $(n=0,\ldots,R-2)$. The left hand side of (4.4.28) is a linear function of p_o; by solving (4.4.28) for p_o, unique solutions for the steady-state equations (4.4.12) to (4.4.22) will be determined. Hence the marginal probability generating functions for the states $(j;1)$ $(j=1,\ldots,R-1)$ and $(k;2)$ $(k=r+1,r+2,\ldots)$ are

$$P_1(z)=\sum_{k=1}^{R-1}z^k\sum_{n=0}^{k-1}p_{k-n}(0;1)\,\frac{\lambda^n}{n!}\,J_n(\lambda) \quad ,$$

$$P_2(z)=p_{R-1}(1)\,\frac{z^{R+1}-z^{r+1}}{z-g_2^*(\lambda-\lambda z)}\;\frac{1-g_2^*(\lambda-\lambda z)}{1-z} \quad ,$$

respectively.

General expressions can be written for several properties of the equilibrium process. For example, the mean line length, $E(L)$, is given by

$$E(L)=P_1'(1)+\frac{1}{2}\,\frac{(R-r)\,p_{R-1}(1)}{1-\rho_2}\left\{(R+r+1)\rho_2+\frac{\lambda^2 g_{2,2}}{1-\rho_2}\right\} .$$

The rate, σ, at which the service time distribution changes from $G_1(\cdot)$ to to $G_2(\cdot)$, or vice versa, is equal to

$$\sigma=\lambda\,p_{R-1}(1)=p_{r+1}(0;2)\,g_2^*(\lambda) \quad .$$

This relation was used to obtain (4.4.27). If ξ is the long-run proportion of time that customers are served according to $G_2(\cdot)$, then

$$\xi=P_2(1)=\frac{(R-r)\rho_2}{1-\rho_2}\,p_{R-1}(1) \quad .$$

Since the arrival process is Poisson and customers arrive or depart singly, a result due to Khintchine(1932) which is quoted by Cox & Miller (1965, p.269) guarantees that the equilibrium line size probability distribution imbedded in the continuous time process at departure epochs is identical to the equilibrium distribution in continuous time. Therefore, the probabilities $p_n(j)$ approximate the proportion of customers who leave behind a total of n customers and the service process operating at level j. However, no customer can depart and leave the system in the state (L=R-1;Z=1). It follows that $p_{R-1}(1)$ must approximate the proportion of customers whose service is interrupted by a change from $G_1(\cdot)$ to $G_2(\cdot)$. Since the system is in equilibrium, $p_{R-1}(1)$ also approximates the proportion of customers whose departure causes a change from $G_2(\cdot)$ to $G_1(\cdot)$. Therefore, if η is the proportion of customers whose service times are independent observations from the distribution $G_2(\cdot)$,

$$\eta = \xi + p_{R-1}(1) = p_{R-1}(1)\,\frac{\rho_2(R-r-1)+1}{1-\rho_2}$$

Similarly, since p_o is the long-run proportion of time that the server is idle, the overall traffic intensity, ρ, is equal to $1-p_o$.

The following examples illustrate the application of bilevel hysteresis control to different queueing situations.

Example 4.4.1

Let $G_i(x)=1-e^{-\mu_i x}$ $(i=1,2)$ and $0<\mu_1<\mu_2$. The solutions to the boundary equations are

$$p_1(0;1) = \lambda(1+\rho_1)p_0 \quad , \quad p_j(0;1) = \lambda\rho_1^j p_0 \quad , (j=2,\cdots,r)$$

$$p_j(0;1) = \lambda p_0 \rho_1^j \,\frac{1-\rho_1^{R-j-1}}{1-\rho_1^{R-r}} \quad , \quad (j=r+1,\cdots,R-1) \; .$$

where $\rho_1 = \frac{\lambda}{\mu_1}$. Hence,

$$p_j(1) = p_0\rho_1^j \quad (j=1,\cdots,r) \quad , \quad p_j(1) = p_0\,\frac{\rho_1^j-\rho_1^R}{1-\rho_1^{R-r}} \quad (j=r+1,\cdots,R-1),$$

$$p_j(2) = p_0 A_1 \frac{\rho_2-\rho_2^{j+1-r}}{1-\rho_2^{R-r}} \quad (j=r+1,\cdots,R), \quad p_j(2) = p_0 A_1 \rho_2^{j+1-R} \quad (j=R+1,R+2,\cdots) \; .$$

where $A_1 = \rho_1^{R-1}\,\frac{1-\rho_1}{1-\rho_1^{R-r}}\,\frac{1-\rho_2^{R-r}}{1-\rho_2}$, and p_o is defined by the equation

$$\frac{1}{p_0} = \frac{1}{1-\rho_1} - \frac{\rho_1^{R-1}(\rho_1-\rho_2)(R-r)}{(1-\rho_1^{R-r})(1-\rho_2)} \; .$$

Formulae for $E(L)$, σ, and η are therefore given by

$$E(L) = P_0\left[\frac{\rho_1}{(1-\rho_1)^2} - \frac{\rho_1^{R-1}(\rho_1-\rho_2)(R-r)}{(1-\rho_1^{R-r})(1-\rho_2)}\left\{\frac{1}{2}(R+r-1) + \frac{1-\rho_1\rho_2}{(1-\rho_1)(1-\rho_2)}\right\}\right] ,$$

$$\sigma = \lambda P_0 \frac{\rho_1^{R-1}(1-\rho_1)}{1-\rho_1^{R-r}} , \qquad \eta = P_0 \frac{\rho_1^{R-1}(1-\rho_1)}{1-\rho_1^{R-r}} \frac{\rho_2(R-r-1)+1}{1-\rho_2} .$$

This example corresponds to the case treated by Gebhard(1967) and the above expressions for the equilibrium probabilities, etc. are identical to results which Gebhard obtains by solving steady-state equations for this particular Markov queueing process.

Note that by setting $r+1=N=R$, bilevel hysteresis control reduces to unilevel control and the expressions in Example 4.4.1 are identical to the results of Example 4.2.1.

To illustrate, briefly, a few of the differences between bilevel hysteresis and unilevel adaptive control, line size distributions for several combinations of r, R are presented in Table 4.4.1. The final column in the table gives the distribution of L for unilevel control. Both $G_1(\cdot)$ and $G_2(\cdot)$ are assumed to be exponential distributions.

The next example considers the effect of bilevel hysteresis control on the switching rate, σ.

Example 4.4.2

Suppose that $G_1(x)=1-e^{-\mu x}$ and $G_2(\cdot)$ is arbitrary with $\rho_2<\rho_1$. Let σ be the switching rate for a bilevel hysteresis control model with control levels r and R ($r<R-1$) and let σ' be the corresponding rate for a unilevel control model with control threshold R. According to Examples 4.4.1 and 4.2.1

$$\sigma = \lambda\rho_1^{R-1} P_0 \frac{1-\rho_1}{1-\rho_1^{R-r}} , \qquad \sigma' = \lambda\rho_1^{R-1} P_0'$$

where

$$\frac{1}{P_0} = \frac{1}{1-\rho_1} - \frac{\rho_1^{R-1}(R-r)(\rho_1-\rho_2)}{(1-\rho_1^{R-r})(1-\rho_2)} , \qquad \frac{1}{P_0'} = \frac{1}{1-\rho_1} - \frac{\rho_1^{R-1}(\rho_1-\rho_2)}{(1-\rho_1)(1-\rho_2)} .$$

Then $\sigma<\sigma'$ and $\xi'<\xi$, where ξ' and ξ are evaluated for the same unilevel and bilevel hysteresis control models, respectively. Thus, if server operating costs are higher for the distribution $G_2(\cdot)$, a decrease in the

<table>
<tr><th>(r,R)</th><th>(2,8)</th><th>(3,8)</th><th>(4,8)</th><th>(5,8)</th><th>(6,8)</th><th>(7,8)</th></tr>
<tr><td>p_o</td><td>0.053</td><td>0.039</td><td>0.030</td><td>0.023</td><td>0.018</td><td>0.014</td></tr>
<tr><td>$p_1(1)$</td><td>0.080</td><td>0.059</td><td>0.045</td><td>0.034</td><td>0.027</td><td>0.021</td></tr>
<tr><td>$p_2(1)$</td><td>0.119</td><td>0.089</td><td>0.067</td><td>0.052</td><td>0.040</td><td>0.032</td></tr>
<tr><td>$p_3(1)$</td><td>0.114</td><td rowspan="2">0.133</td><td rowspan="2">0.101</td><td rowspan="2">0.077</td><td rowspan="2">0.061</td><td rowspan="2">0.048</td></tr>
<tr><td>$p_3(2)$</td><td>0.024</td></tr>
<tr><td>$p_4(1)$</td><td>0.105</td><td>0.123</td><td rowspan="2">0.151</td><td rowspan="2">0.116</td><td rowspan="2">0.091</td><td rowspan="2">0.072</td></tr>
<tr><td>$p_4(2)$</td><td>0.037</td><td>0.028</td></tr>
<tr><td>$p_5(1)$</td><td>0.092</td><td>0.108</td><td>0.132</td><td rowspan="2">0.174</td><td rowspan="2">0.136</td><td rowspan="2">0.108</td></tr>
<tr><td>$p_5(2)$</td><td>0.044</td><td>0.044</td><td>0.035</td></tr>
<tr><td>$p_6(1)$</td><td>0.073</td><td>0.085</td><td>0.105</td><td>0.138</td><td rowspan="2">0.204</td><td rowspan="2">0.162</td></tr>
<tr><td>$p_6(2)$</td><td>0.048</td><td>0.052</td><td>0.053</td><td>0.045</td></tr>
<tr><td>$p_7(1)$</td><td>0.044</td><td>0.051</td><td>0.063</td><td>0.083</td><td>0.123</td><td rowspan="2">0.244</td></tr>
<tr><td>$p_7(2)$</td><td>0.051</td><td>0.057</td><td>0.064</td><td>0.070</td><td>0.067</td></tr>
<tr><td>$p_8(2)$</td><td>0.052</td><td>0.059</td><td>0.070</td><td>0.084</td><td>0.105</td><td>0.134</td></tr>
<tr><td>$p_9(2)$</td><td>0.029</td><td>0.033</td><td>0.038</td><td>0.046</td><td>0.058</td><td>0.074</td></tr>
<tr><td>$p_{10}(2)$</td><td>0.016</td><td>0.018</td><td>0.021</td><td>0.025</td><td>0.032</td><td>0.041</td></tr>
<tr><td>$p_{11}(2)$</td><td>0.009</td><td>0.010</td><td>0.012</td><td>0.014</td><td>0.017</td><td>0.022</td></tr>
<tr><td>$pr(L\geq 12)$</td><td>0.010</td><td>0.012</td><td>0.013</td><td>0.019</td><td>0.021</td><td>0.028</td></tr>
<tr><td>$E(L)$</td><td>4.43</td><td>4.78</td><td>5.17</td><td>5.59</td><td>6.03</td><td>6.48</td></tr>
<tr><td>σ</td><td>0.044</td><td>0.051</td><td>0.063</td><td>0.083</td><td>0.123</td><td>0.244</td></tr>
<tr><td>η</td><td>0.363</td><td>0.363</td><td>0.370</td><td>0.385</td><td>0.423</td><td>0.541</td></tr>
</table>

Table 4.4.1 Equilibrium marginal probability distributions for the line length, L, in six different hysteresis control queues with control levels (r,R) and traffic intensity 1.5(0.55) for slow(fast) exponential service. Changes from slow(1) to faster(2) service occur at rate σ, and a proportion, η, of customers receive faster service.

switching frequency and presumably in the switching costs as well will be partly offset by increased serving costs.

It was suggested that rule (iv) simplifies the analysis of bilevel control. However, if rule (iv) is omitted, we can still determine the equilibrium line size distribution. In this case, the decision points of the process are service epochs. The analysis follows similar lines, but is complicated by the possibility that when the state of the process is $(R,x;1)$, transitions through the states $(R+1,y;1)$, $(R+2,z;1)$ $(x<y<z)$, etc. may occur before the next departure leaves the system in the state $(j,0;2)$ $(j\geq R)$. Thus, deleting rule (iv) increases the number of possible states, but as the results of the next example illustrate, the difference between the solutions for the two models is probably negligible in most situations.

Example 4.4.3

Let $G_i(x)=1-e^{-\mu_i x}$ $(i=1,2)$ and $0<\mu_1<\mu_2$. When rule (iv) is deleted the equilibrium distribution of L is given by

$$p_k(1) = p_0\rho_1^k \quad (k=1,\cdots,r) \quad , \quad p_k(1) = p_0\,\frac{\rho_1^k-\rho_1^{R+1}}{1-\rho_1^{R-r+1}} \quad (k=r+1,\cdots,R-1)\,,$$

$$p_k(1) = p_0\,\frac{\rho_1^k}{(1+\rho_1)^{k-R}}\,\frac{1-\rho_1}{1-\rho_1^{R-r+1}} \quad (k=R,R+1,\dots)\,,\quad p_k(2) = p_0 B_1\,\frac{\rho_2-\rho_2^{k-r+1}}{1-\rho_2} \quad (k=r+1,\dots,R),$$

$$p_k(2) = p_0 B_1\,\frac{\rho_2}{1-\rho_2}\left\{\frac{\rho_2^{k-R+1}}{\rho_2+\rho_1\rho_2-\rho_1} - \rho_2^{k-r} - \frac{\rho_1^{k-R+1}(1-\rho_2)}{(1+\rho_1)^{k-R}(\rho_2+\rho_1\rho_2-\rho_1)}\right\} \quad (k=R+1,R+2,\dots),$$

where $B_1 = \dfrac{\rho_1^R(1-\rho_1)}{1-\rho_1^{R-r+1}}$, and p_0 satisfies the equation

$$\frac{1}{p_0} = \frac{1}{1-\rho_1} - \frac{(\rho_1-\rho_2)\,\rho_1^R\,(R-r+\rho_1)}{(1-\rho_2)(1-\rho_1^{R-r+1})} \quad .$$

Example 4.4.1 gives the solution for the same service time distributions when rule (iv) is not deleted.

Similar comments apply to the results of §4.2.

Whether rule (iv) is retained or deleted, many factors will undoubtedly influence the choice of control parameters in any practical application of bilevel hysteresis control. The conclusions of §4.3 are only a guide in making such a decision.

CHAPTER 5. Generalized hysteresis control of the service process

5.1 2k-level hysteresis control

A natural, though somewhat less practical, generalization of both the unilevel and bilevel control models discussed in Chapter 4 is one with 2k control levels. The most general formulation of 2k-level hysteresis control involves k+1 service time distributions $G_j(\cdot)$ $(j=1,\ldots,k+1)$ and k+2 control level pairs, (r_n,R_n) $(n=0,\ldots,k+1)$, with $0=r_o<R_o=1<r_1<R_1<\cdots<r_k<R_k<r_{k+1}=R_{k+1}=\infty$. Figure 5.1.1 shows a possible configuration when k=2.

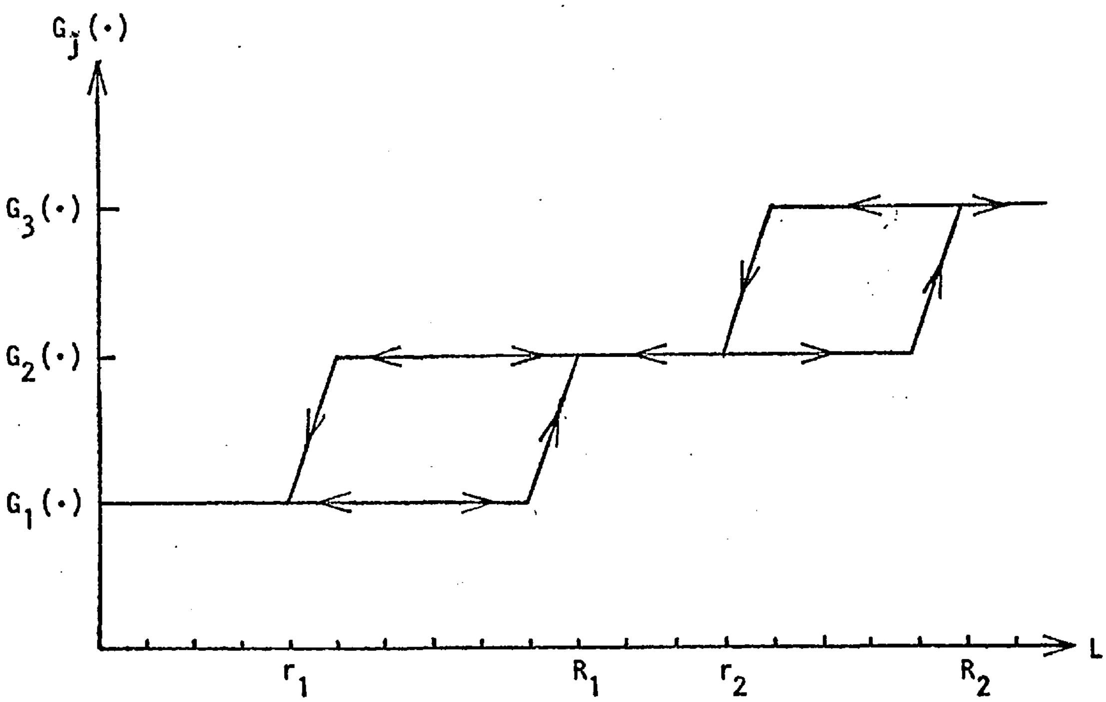


Fig. 5.1.1 Relation between line size, L, and service time distributions, $G_j(\cdot)$ (j=1,2,3), in generalized hysteresis control.

In their introductory paper on hysteresis control, Yadin & Naor(1967) assume arrivals are Poisson and service times at each control level are independent, exponentially distributed. Using the methods of Chapter 4, the more general 2k-level case when the $G_j(\cdot)$ are arbitrary can be solved. In the present chapter, the principle results are outlined.

The decision points of 2k-level hysteresis control are the arrival and service epochs. The choice of service time distribution depends both on the number of customers present and on the recent history of the process. If L_t is the line size process at any time t, the following rules determine the service time distribution at each decision point:

(i) when $R_{j-1} \leq L_t \leq r_j$, the customer is served according to the distribution $G_j(\cdot)$ $(j=1,\ldots,k+1)$,

(ii) when $r_j < L_t < R_j$, the service time distribution remains unchanged, $(j=1,\ldots,k)$,

(iii) if, while a customer, C, is being served according to the distribution $G_j(\cdot)$, L_t increases from R_j-1 to R_j, immediately terminate service to C and begin a new service time for the same customer according to the distribution $G_{j+1}(\cdot)$ $(j=1,\ldots,k)$.

Only rule (iii) causes a customer's service time to be interrupted.

We assume that $G_i(\cdot)$ has derivative $g_i(\cdot)=\phi_i(\cdot)\mathcal{G}_i(\cdot)$ with Laplace transform $g_i^*(s)=\int_0^\infty e^{-st}g_i(t)dt$, $(i=1,\ldots,k+1)$ and that $\rho_{k+1}=\lambda\int_0^\infty t g_{k+1}(t)dt<1$.

Redefine the state of the process, in equilibrium, as the triplet $(L,S;Z)$; L is the line size process, S is the elapsed service time of the customer in service and Z takes the value i if $G_i(\cdot)$ was chosen at the last decision point $(i=1,\ldots,k+1)$. If $p_0=pr(L=0)$ and $p_n(x;j)=pr(L=n,S=x;Z=j)$ $(n=r_{j-1}+1,\ldots,R_j-1;\ j=1,\ldots,k+1;\ x\geq 0)$ then steady-state equations for the probability distribution of $(L,S;Z)$ are given by

$$\lambda p_0 = \int_0^\infty p_1(x;1)\,\phi_1(x)\,dx \quad , \tag{5.1.1}$$

$$\frac{\partial}{\partial x} p_{r_j+1}(x;j+1) + \{\lambda+\phi_{j+1}(x)\}\, p_{r_j+1}(x;j+1) = 0 \quad , \quad (j=0,\ldots,k) \tag{5.1.2}$$

$$\frac{\partial}{\partial x} p_n(x;j+1) + \{\lambda+\phi_{j+1}(x)\}\, p_n(x;j+1) = \lambda p_{n-1}(x;j+1) \quad , \quad \begin{pmatrix} n=r_j+2,\ldots,R_{j+1}-1; \\ j=0,\ldots,k \end{pmatrix} \tag{5.1.3}$$

$$p_1(0;1) = \lambda p_0 + \int_0^\infty p_2(x;1)\,\phi_1(x)\,dx \quad , \tag{5.1.4}$$

$$p_n(0;j+1) = \int_0^\infty p_{n+1}(x;j+1)\,\phi_{j+1}(x)\,dx, \quad \begin{bmatrix} n=r_j+1,\ldots,R_j-1,R_j+1,\ldots,r_{j+1}-1,r_{j+1}+1,\ldots,R_{j+1}-2; \\ j=0,\ldots,k \end{bmatrix} \tag{5.1.5}$$

$$p_{r_j}(0;j) = \int_0^\infty p_{r_j+1}(x;j+1)\,\phi_{j+1}(x)\,dx + \int_0^\infty p_{r_j+1}(x;j)\,\phi_j(x)\,dx, \quad (j=1,\ldots,k) \tag{5.1.6}$$

$$p_{R_j-1}(0;j) = 0, \qquad (j=1,\ldots,k) \tag{5.1.7}$$

$$p_{R_j}(0;j+1) = \lambda\int_0^\infty p_{R_j-1}(x;j)\,dx + \int_0^\infty p_{R_j+1}(x;j+1)\,\phi_{j+1}(x)\,dx, \quad (j=1,\ldots,k) \tag{5.1.8}$$

Let $P_{j+1}(x;z) = \sum_{n=r_j+1}^{R_{j+1}-1} p_n(x;j+1)z^n$ $(|z|\leq 1)$ $(j=0,\ldots,k)$. Solutions to (5.1.2) and (5.1.3) for $j=0,\ldots,k-1$ are given by

$$P_{j+1}(x;z) = \sum_{n=r_j+1}^{R_{j+1}-1} z^n \left\{ \sum_{m=0}^{n-r_j-1} p_{n-m}(0;j+1)\,\frac{(\lambda x)^m}{m!}\,e^{-\lambda x}\,\mathcal{G}_{j+1}(x) \right\}. \tag{5.1.9}$$

When $j=k$ we can combine (5.1.2) and (5.1.3) in the single equation

$$\frac{\partial}{\partial x} P_{k+1}(x;z) = \left\{ \lambda z - \lambda - \phi_{k+1}(x) \right\} P_{k+1}(x;z)$$

which has the solution

$$P_{k+1}(x;z) = P_{k+1}(0;z)\,e^{-\lambda x(1-z)}\,\mathcal{G}_{k+1}(x), \tag{5.1.10}$$

where $P_{k+1}(0;z)=\lim_{x\to 0+} P_{k+1}(x;z)$. We can also combine (5.1.5) and (5.1.8) as

$$P_{k+1}(0;z) = \frac{1}{z}\int_0^\infty P_{k+1}(x;z)\,\phi_{k+1}(x)\,dx - z^{r_k}\int_0^\infty p_{r_k+1}(x;k+1)\,\phi_{k+1}(x)\,dx + \lambda p_{R_k-1}(k)\,z^{R_k}, \tag{5.1.11}$$

where $p_n(j+1)=\int_0^\infty p_n(x;j+1)dx$, $(n=r_j+1,\ldots,R_{j+1}-1;\ j=0,\ldots,k)$. By substituting (5.1.10) in (5.1.11), we can solve the resulting equation for $P_{k+1}(0;z)$; thus

$$P_{k+1}(0;z) = \frac{\lambda p_{R_k-1}(k)\,z^{R_k+1} - z^{r_k+1}\int_0^\infty p_{r_k+1}(x;k+1)\,\phi_{k+1}(x)\,dx}{z - g^*_{k+1}(\lambda-\lambda z)}. \tag{5.1.12}$$

Since the system is in equilibrium, $\lambda p_{R_k-1}(k) = \int_0^\infty p_{r_k+1}(x;k+1)\not{0}_{k+1}(x)dx$; hence

$$P_{k+1}(0;z) = \lambda p_{R_k-1}(k) \frac{z^{R_k+1} - z^{r_k+1}}{z - g^*_{k+1}(\lambda-\lambda z)} \quad . \tag{5.1.13}$$

Unique solutions for (5.1.1) to (5.1.8) will be determined if unique expressions for $p_n(0;j+1)$ $(n=r_j+1,\ldots,R_{j+1}-1;\ j=0,\ldots,k-1)$ can be derived.

We begin by obtaining expressions for $p_1(0;1),\ldots,p_{R_1-1}(0;1)$ which are unique to within p_o and then sketch the details of a general procedure for evaluating $p_n(0;j+1)$ $(n=r_j+1,\ldots,R_{j+1}-1;\ j=1,\ldots,k-1)$.

Substitute solutions for $p_1(x;1)$ and $p_2(x;1)$ in (5.1.1) and (5.1.4) to show that

$$p_1(0;1) = \frac{\lambda p_o}{g_1^*(\lambda)} \quad , \quad p_2(0;1) = \frac{\lambda p_o}{\{g_1^*(\lambda)\}^2}\left\{1 - g_1^*(\lambda) + \lambda \frac{d}{d\lambda} g_1^*(\lambda)\right\}.$$

In general, we can obtain expressions for $p_3(0;1),\ldots,p_{r_1}(0;1)$ in terms of p_o by solving, iteratively,

$$\sum_{m=0}^{n} p_{n+1-m}(0;1)\left\{\delta_{1m} - \frac{(-\lambda)^m}{m!} g_1^{*(m)}(\lambda)\right\} = 0 \quad , \quad (n=2,\ldots,r_1-1) \tag{5.1.14}$$

where $g_1^{*(m)}(\lambda) = \frac{d^m}{d\lambda^m} g_1^*(\lambda)$, $(m=0,\ldots,R_1-2)$. If we substitute for $p_{r_1+1}(x;1)$ and $p_{r_1+1}(x;2)$ in (5.1.6) we obtain

$$p_{r_1}(0;1) = p_{r_1+1}(0;1) g_1^*(\lambda) + p_{r_1+1}(0;2) g_2^*(\lambda) + \sum_{m=1}^{r_1} p_{r_1+1-m}(0;1) \frac{(-\lambda)^m}{m!} g_1^{*(m)}(\lambda) \quad , \tag{5.1.15}$$

which can be solved for $p_{r_1+1}(0;2)$, say, in terms of $p_{r_1+1}(0;1)$ and p_o. Continued iterative solving of (5.1.14) for $n=r_1+1,\ldots,R_1-2$ generates expressions for $p_{r_1+2}(0;1),\ldots,p_{R_1-1}(0;1)$ in terms of $p_{r_1+1}(0;1)$ and p_o. But according to (5.1.7), $p_{R_1-1}(0;1)=0$. Using this equation for $p_{R_1-1}(0;1)$, we can evaluate $p_{r_1+1}(0;1)$ in terms of p_o and hence simplify the expressions for $p_{r_1+1}(0;1),\ldots,p_{R_1-2}(0;1)$; we can also solve (5.1.15) for $p_{r_1+1}(0;2)$. Hence, values for $p_1(0;1),\ldots,p_{R_1-1}(0;1)$,

$p(0;2)_{r_1+1}$ which are unique to within p_o have been obtained. This expression for $p(0;2)_{r_1+1}$ is the initial solution for level 2.

In general, solving the jth level boundary equations ($j=1,\ldots,k$) also determines the value of $p(0;j+1)_{r_j+1}$, the initial solution for the $(j+1)$th level. With this initial solution, an equation derived from (5.1.5) [cf. (5.1.14)] can be solved iteratively for $p(0;j+1)_{r_j+2},\ldots,p(0;j+1)_{R_j}$ in terms of p_o. Next, substitute for $p(0;j+1)_{R_j}$, $p(j)_{R_j-1}$ and $p(x;j+1)_{R_j+1}$ in (5.1.8) to obtain an expression for $p(0;j+1)_{R_j+1}$ in terms of p_o. Continued iterative solving of the equation derived from (5.1.5) then generates solutions for $p(0;j+1)_{R_j+2},\ldots,p(0;j+1)_{r_{j+1}}$ which are unique to within p_o, and expressions for $p(0;j+1)_{r_{j+1}+2},\ldots,p(0;j+1)_{R_{j+1}-1}$ which can be written in terms of $p(0;j+1)_{r_{j+1}+1}$ and p_o. But according to (5.1.7), $p(0;j+1)_{R_{j+1}-1}=0$. This equation determines a solution for $p(0;j+1)_{r_{j+1}+1}$ in terms of p_o and hence $p(0;j+1)_{r_{j+1}+2},\ldots,p(0;j+1)_{R_{j+1}-2}$ can also be evaluated to within p_o. Finally, the initial solution for level j+2 can be obtained by substituting in (5.1.6) and solving the resulting equation for $p(0;j+2)_{r_{j+1}+1}$.

Thus, beginning with an initial solution for level j+1, it is possible to solve all the (j+1)-level boundary equations, determining $p(0;j+1)_{r_j+1},\ldots,$ $p(0;j+1)_{R_{j+1}-1}$ and the initial solution for level j+2 in terms of p_o. The procedure is identical for $j=1,\ldots,k-1$. Hence solutions for $p(0;j+1)_{r_j+1},\ldots,$ $p(0;j+1)_{R_{j+1}-1}$ ($j=0,\ldots,k-1$) which are unique to within p_o can be obtained.

To obtain unique solutions for (5.1.1) to (5.1.8) we require a normalizing equation. Let $P_{j+1}(z)=\sum_{n=r_j+1}^{R_{j+1}-1} p_n(j+1)z^n=\int_0^\infty P_{j+1}(x;z)dx \quad (j=0,\ldots,k)$. Since $p_0+\sum_{j=1}^{k+1} P_j(1)=1$, therefore

$$p_0+\sum_{j=0}^{k-1}\sum_{n=r_j+1}^{R_{j+1}-1}\sum_{m=0}^{n-r_j-1} p_{n-m}(0;j+1)\frac{\lambda^m}{m!}\psi_{j+1,m}(\lambda)+p_{R_k-1}(k)\frac{(R_k-r_k)\rho_{k+1}}{1-\rho_{k+1}}=1, \quad (5.1.16)$$

where $\psi_{j+1,m}(\lambda)=\int_0^\infty x^m \mathcal{S}_{j+1}(x)e^{-\lambda x}dx \quad (m=0,\ldots,R_{j+1}-r_j-2;\ j=0,\ldots,k-1)$. The left hand side of (5.1.16) is linear in p_0; hence p_0 is uniquely specified by (5.1.16) and unique solutions for (5.1.1) to (5.1.8) can be obtained.

The marginal probability generating function for the states $(n;j+1)$ $(n=r_j+1,\ldots,R_{j+1}-1)$, i.e. the $(j+1)$th level, is

$$P_{j+1}(z)=\sum_{n=r_j+1}^{R_{j+1}-1} z^n \sum_{m=0}^{n-r_j-1} p_{n-m}(0;j+1)\frac{\lambda^m}{m!}\psi_{j+1,m}(\lambda), \quad (j=0,\ldots,k-1)$$

$$=p_{R_k-1}(k)\frac{z^{R_k+1}-z^{r_k+1}}{z-g^*_{k+1}(\lambda-\lambda z)}\,\frac{1-g^*_{k+1}(\lambda-\lambda z)}{1-z}, \quad (j=k).$$

The marginal steady-state probabilities, p_0, $p_n(j+1)$, can be used to obtain expressions for the usual properties of the equilibrium queueing system. These expressions are simple generalizations of the formulae derived in §4.4 for corresponding properties of bilevel control. Therefore we omit them from this outline.

Obviously, by making the substitution $r_j=N_j=R_j-1$ for some values of j in the preceding discussion, the solution procedure can be simplified since (5.1.5) is then only defined for $n \geqq N_j+1$. Similarly, (5.1.6) and (5.1.7) are together superseded by the equation $p_{N_j-1}(0;j)=\int_0^\infty p_{N_j}(x;j+1)\phi_{j+1}(x)dx$. When $r_j=N_j=R_j-1$ for all values of j, the resulting model is a k-level analogue of the unilevel control model discussed in §§4.2 and 4.3.

The following simple example concludes this discussion of generalized hysteresis control.

Example 5.1.1

Let $G_j(x)=1-e^{-\mu_j x}$, $(j=1,\dots,k+1)$ where $0<\mu_1<\cdots<\mu_{k+1}$. Define $\rho_j=\frac{\lambda}{\mu_j}$ and assume that $\rho_{k+1}<1$. Solutions for the boundary equations are given by

$$P_n(0;j+1)=\lambda P_0 A_j \frac{1-\rho_{j+1}^{n-r_j+1}}{1-\rho_{j+1}^{R_j-r_j}}, \quad (n=r_j+1,\dots,R_j;\ j=0,\dots,k)$$

$$P_n(0;j+1)=\lambda P_0 A_j \rho_{j+1}^{n+1-R_j}, \quad (n=R_j+1,\dots,r_{j+1};\ j=0,\dots,k)$$

$$P_n(0;j+1)=\lambda P_0 A_j \frac{\rho_{j+1}^{n-R_j+1}-\rho_{j+1}^{R_{j+1}-R_j}}{1-\rho_{j+1}^{R_{j+1}-r_{j+1}}}, \quad (n=r_{j+1}+1,\dots,R_{j+1}-1;\ j=0,\dots,k-1)$$

where $A_j=\prod_{s=1}^{j}\rho_s^{R_s-R_{s-1}}\left(\frac{1-\rho_{s+1}^{R_s-r_s}}{1-\rho_s^{R_s-r_s}}\,\frac{1-\rho_s}{1-\rho_{s+1}}\right)$, $(j=1,\dots,k)$ and $A_0\equiv 1$. Hence

$$P_n(j+1)=P_0 A_j \frac{\rho_{j+1}-\rho_{j+1}^{n+1-r_j}}{1-\rho_{j+1}^{R_j-r_j}}, \quad (n=r_j+1,\dots,R_j;\ j=0,\dots,k)$$

$$P_n(j+1)=P_0 A_j \rho_{j+1}^{n+1-R_j}, \quad (n=R_j+1,\dots,r_{j+1};\ j=0,\dots,k)$$

$$P_n(j+1)=P_0 A_j \frac{\rho_{j+1}^{n+1-R_j}-\rho_{j+1}^{R_{j+1}-R_j+1}}{1-\rho_{j+1}^{R_{j+1}-r_{j+1}}}, \quad (n=r_{j+1}+1,\dots,R_{j+1}-1;\ j=0,\dots,k-1).$$

The value of p_0 is determined by the equation

$$\frac{1}{P_0}=1+\sum_{j=0}^{k}A_j\rho_{j+1}\frac{R_j-r_j}{1-\rho_{j+1}^{R_j-r_j}}-\sum_{j=0}^{k-1}A_j\rho_{j+1}^{R_{j+1}-R_j+1}\frac{R_{j+1}-r_{j+1}}{1-\rho_{j+1}^{R_{j+1}-r_{j+1}}}.$$

If we set k=1, the results of Example 5.1.1 reduce to those obtained in Example 4.4.1.

CHAPTER 6. Concluding remarks

6.1 An alternative to optimal control

The question of explicitly optimizing the control of single server queueing systems has been largely ignored in preceding chapters except in §§2.1 and 2.2. There it was necessary to choose a service time, $\emptyset$, which divides customers into "short" and "long" classes. The particular value, $\emptyset^*$, which was selected is one which minimizes mean queueing time.

Optimal control of queueing processes is not a neglected subject in the literature. Various authors whose work has already been mentioned have tried to explore this question [cf. Heyman(1968), Bell(1971) and Yechiali(1971)]. Most authors adopt one of two approaches.

One may postulate a queueing system with a specified cost structure involving items such as holding, serving, start-up and shut-down costs and a finite list of possible actions in each situation. This approach usually requires the use of dynamic or Markov renewal programming techniques to determine the form of optimal policies for different planning horizons, with and without cost discounting over time. For examples of this method, see Heyman(1968), Bell(1971), Yadin & Zacks(1971) and Crabill(1972).

On the other hand, one may prescribe a control policy of a particular form for a given queueing system. In this case, the effect of the prescribed control policy on various system features is usually determined in terms of average values. Optimal control is then introduced as the problem of selecting control parameters in order to minimize costs or maximize revenues as determined by a postulated cost framework. This approach is exemplified by the work of Yadin & Naor(1963), Moder & Phillips(1962) and Gebhard(1967).

A few authors adopt a third approach to the problem of optimal control of queueing processes. This involves applying traditional mathematical methods in order to optimize a particular aspect of a queueing process. Thus, Shapiro(1965) uses the second method of Lyapunov to minimize the mean squared deviation of waiting time from a predetermined standard. Man(1973) utilizes the Pontryagin maximum principle to determine a dynamic operating policy in a time-dependent M/M/s queue with N-s places for queueing customers. The optimal control policy which Man derives regulates the customer arrival rate in order to minimize the mean squared excess of customers over servers in a specified finite interval of time.

The development of Chapters 2-5 has not followed any one of these three common approaches. In applications of queueing theory [cf. Lee (1966)] the problems that arise do not appear to require rigorous, optimal solutions for conceptual models; however, practical, operational solutions are obviously necessary. Ideally, developments in queueing theory should arise as new problems are met. When this is not the case, theoretical advances ought to be supplemented by indications of their appropriateness and applicability in various situations.

For this reason, no attempt has been made to optimize the methods suggested in Chapters 2-5. It might be possible to define a general cost structure and, within that frame of reference, determine which of the various control techniques optimizes a selected objective criterion. Instead, attention has concentrated on some ways in which information about the present, or perhaps future, state of an M/G/1 queueing system can be used to manage congestion. By a series of numerical studies an attempt has been made to determine the likely effects of the suggested methods on existing queueing processes.

Various qualitative conclusions are another result of these same numerical studies. In each case, there appears to be some evidence, occasionally quite conclusive, that the suggested control methods are most effective in managing congestion when the service time distribution is more dispersed than the exponential distribution. Conversely, when service times tend to be regular, the various methods considered appear to be less effective. Obviously, for a fixed arrival pattern and queue discipline, the degree to which a system is congested will very much depend on the service time distribution. Since the suggested control methods — with the notable exception of shut-down control — tend to impose a greater regularity on the system than had existed previously, so the differential effect of those same control methods on more congested systems is greater.

As an alternative to theories of optimal control, then, specific changes in the basic features of a queueing process have been suggested. Theoretical treatments of the results of these changes are supported by quantitative evidence in specific cases indicating qualitative effects in more general situations. Lee(1966) demonstrates conclusively that "applications involve much bending and twisting of the theoretical models". This suggests that results which offer insight into simple schemes for managing congestion are probably of greater practical importance than theoretical solutions for optimizing the control of a given queueing system.

6.2 Some outstanding problems

No mention was made in Chapters 4 or 5 of the equilibrium distribution of W_q, the queueing time. Since service times under hysteresis control depend on the line size, customers' queueing times are partly determined by the pattern of subsequent arrivals. In most cases, this dependence makes the equilibrium queueing time distribution difficult to analyze. However, if customers are served in order of arrival, the Laplace transform of the distribution of W_q for unilevel control can be derived by the following argument, provided the control threshold, N, equals 2.

Clearly, the probability distribution of W_q will be of the form $p_o+(1-p_o)v(x)$; p_o is the equilibrium probability that the line size is zero and $v(x)$ is the conditional probability density function of positive queueing times. Thus $E(e^{-sW_q})=p_o+(1-p_o)v^*(s)$, where $v^*(s)=\int_{0+}^{\infty} e^{-sx}v(x)dx$.

Let C be any customer who joins the queue during a busy period. Since N=2, the service time distribution for all customers preceding C in the queue will be $G_2(\cdot)$. When C's service time begins the number of customers behind him in the queue, i.e. who arrived during C's queueing time, is j with probability $p_{j+1}/(1-p_o)$ $(j=0,1,\ldots)$; $p_k/(1-p_o)$ is the equilibrium probability that the line size, L, is k $(k=1,2,\ldots)$, given that $L>0$. Since arrivals are Poisson

$$\frac{p_{j+1}}{1-p_0} = \int_{0+}^{\infty} \frac{(\lambda x)^j}{j!} e^{-\lambda x} v(x)\,dx \quad ,$$

and so

$$\begin{aligned}\frac{1}{1-p_0}\sum_{j=0}^{\infty} p_{j+1} z^j &= \sum_{j=0}^{\infty}\left\{\int_{0+}^{\infty}\frac{(\lambda x z)^j}{j!} e^{-\lambda x} v(x)\,dx\right\} \\ &= v^*(\lambda-\lambda z) \quad , \qquad (|z|\le 1).\end{aligned}$$

But by (4.2.27) we know that for N=2, $\sum_{j=1}^{\infty} p_j z^j = p_1 z\frac{(z-1)g_2^*(\lambda-\lambda z)}{z-g_2^*(\lambda-\lambda z)}$. Hence,

$$v^*(\lambda-\lambda z) = \frac{p_1}{1-p_0}\,\frac{(z-1)\,g_2^*(\lambda-\lambda z)}{z-g_2^*(\lambda-\lambda z)} \quad . \tag{6.2.1}$$

To determine $v^*(s)$, set $s=\lambda-\lambda z$ in (6.2.1). Then

$$v^*(s) = \frac{p_1}{1-p_0}\,\frac{s g_2^*(s)}{s-\lambda+\lambda g_2^*(s)} \quad ,$$

giving

$$E\left(e^{-sW_q}\right) = p_0 + p_1\,\frac{s g_2^*(s)}{s-\lambda+\lambda g_2^*(s)} \quad .$$

When N>2 the number of customers who arrive while C is queueing may cause a change in the service time distribution, either from $G_1(\cdot)$ to $G_2(\cdot)$ or vice versa. Since C's queueing time is determined by the service times of customers preceding him in the queue, it follows that C's queueing time depends on both the number of customers who follow C in the queue and their arrival pattern. This greatly complicates the problem of determining the equilibrium queueing time distribution in general. If the distribution of W_q could be determined when N>2, numerical studies such as those in §§2.2, 3.3 and 4.3 could be used to quantify the effects of unilevel control on queueing time.

Similar comments apply to the problem of determining the queueing time distribution for bilevel hysteresis control. An expression for the general form of this distribution is unlikely to be obtained before the analogous unilevel control problem has been solved.

Excluding a few authors whose work has been mentioned, almost no one has considered the problem of controlling multi-server queueing processes. No doubt this is partly due to the considerable analytical difficulties which are encountered whenever the queueing process is non-Markov. There is the added problem, however, of devising multi-server control schemes which are both sensible and practicable. Kingman(1962) has shown that if customers are indistinguishable from the point of view of service time, and if we exclude the possibility of an idle server while other customers are queueing, then service in order of arrival minimizes the queueing time variance. Until some means of using further information about the state of a multi-server queueing process is devised, full server availability combined with service in order of arrival is probably the most effective solution to this very practical problem.

REFERENCES

Abramowitz, M. & Stegun, I.A. (Eds.) (1964). Handbook of Mathematical Functions. New York: Dover.

Adler, I. & Naor, P. (1969). Social Optimization Versus Self-Optimization in Waiting Lines. Israel Institute of Technology, Faculty of Management & Industrial Engineering. Haifa.

Bell, C.E. (1971). Characterization and Computation of Optimal Policies for Operating an M/G/1 Queueing System With Removable Server. Ops. Res. 19, 208-218.

Cobham, A. (1954). Priority Assignment in Waiting Line Problems. J. Ops. Res. Soc. Am. 2, 70-76.

Cobham, A. (1955). Priority Assignment — A Correction. J. Ops. Res. Soc. Am. 3, 547.

Cohen, J.W. (1969). The Single Server Queue. Amsterdam: North-Holland.

Conway, R.W. & Maxwell, W.L. (1962). Network Dispatching by the Shortest-Operation Discipline. Ops. Res. 10, 51-73.

Cox, D.R. (1962). Renewal Theory. London: Methuen.

Cox, D.R. & Miller, H.D. (1965). The Theory of Stochastic Processes. London: Methuen.

Cox, D.R. & Smith, W.L. (1961). Queues. London: Methuen.

Crabill, T.B. (1972). Optimal Control of a Service Facility With Variable Exponential Service Times and Constant Arrival Rate. Mgmt. Sci. 18, 560-566.

Gebhard, R.F. (1967). A Queueing Process With Bilevel Hysteretic Service-Rate Control. Nav. Res. Logist. Q. 14, 55-67.

Heyman, D.P. (1968). Optimal Operating Policies for M/G/1 Queueing Systems. Ops. Res. 16, 362-382.

Kendall, D.G. (1953). Stochastic Processes Occurring in the Theory of Queues and Their Analysis by the Method of the Imbedded Markov Chain. Ann. Math. Statist. 24, 338-354.

Kendall, M.G. & Stuart, A. (1963). The Advanced Theory of Statistics, I, 2nd edition. London: Griffin.

Kesten, H. & Runnenburg, J.Th. (1957). Priority in Waiting Line Problems. I and II. Proc. K. ned. Akad. Wet. A, 60, 312-336.

Khintchine, A.J. (1932). Mathematisches über die Erwartung vor einem öffertlichen Schalter. Mat. Sb. 39, 73-84.

Kingman, J.F.C. (1962). The Effect of Queue Discipline on Waiting Time Variance. Proc. Camb. phil. Soc. 58, 163-164.

Lee, A.M. (1966). Applied Queueing Theory. London: MacMillan.

Magazine, M.J. (1971). Optimal Control of Multi-Channel Service Systems. Nav. Res. Logist. Q. 18, 177-183.

Man, F.T. (1973). Optimal Control of Time-Varying Queueing Systems. Mgmt. Sci. 19, 1249-1256.

Moder, J.J. & Phillips, C.R. Jr. (1962). Queueing With Fixed and Variable Channels. Ops. Res. 10, 218-231.

Naor, P. (1969). On the Regulation of Queue Size by Levying Tolls. Econometrica, 37, 15-24.

Oliver, R.M. & Pestalozzi, G. (1965). On a Problem of Optimum Priority Classification. J. Soc. ind. appl. Math. 13, 890-901.

Phipps, T.E. Jr. (1956). Machine Repair as a Priority Waiting-Line Problem. J. Ops. Res. Soc. Am. 4, 76-85.

Saaty, T.L. (1961). Elements of Queueing Theory With Applications. New York: McGraw-Hill.

Schrage, L. (1968). A Proof of the Optimality of the Shortest Remaining Processing Time Discipline. Ops. Res. 16, 687-690.

Schrage, L.E. & Miller, L.W. (1966). The Queue M/G/1 With the Shortest Remaining Processing Time Discipline. Ops. Res. 14, 670-684.

Scott, M. (1971). Queueing With Control on the Arrival of Certain Type of Customers. CORS J. 8, 75-86.

Shapiro, S. (1965). A Technique to Control Waiting Time in a Queue. IBM Syst. J. 4, 53-57.

Takács, L. (1962). Introduction to the Theory of Queues. New York: Oxford University Press.

Takács, L. (1964). Priority Queues. Ops. Res. 12, 63-74.

Yadin, M. & Naor, P. (1963). Queueing Systems With a Removable Service Station. Opl. Res. Q. 14, 393-405.

Yadin, M. & Naor, P. (1967). On Queueing Systems With Variable Service Capacities. Nav. Res. Logist. Q. 14, 43-53.

Yadin, M. & Zacks, S. (1971). The Optimal Control of a Queueing Process. In Developments in Operations Research, I, ed. B. Avi-Itzhak, pp. 241-252. New York: Gordon & Breach.

Yechiali, U. (1971). On Optimal Balking Rules and Toll Charges in the GI/M/1 Queueing Process. Ops. Res. 19, 349-370.