1981

Operational Analysis of Queues with General Service Times

Jeffrey A. Brumfield

Peter J. Denning

Report Number:
80-357

OPERATIONAL ANALYSIS OF QUEUES

WITH GENERAL SERVICE TIMES

Jeffrey A. Brumfield

Peter J. Denning

Department of Computer Sciences
Purdue University
West Lafayette, IN 47907

CSD-TR-357

January 1981

Abstract. Formulas relating the mean queue length, utilization, and coefficient of variation of service time of a queue during a given observation period are derived using operational analysis. The main formula, a counterpart of the Pollaczek-Khintchin formula for M/G/1 queues, relies on four homogeneity assumptions. The other formulas require fewer assumptions. Simulation experiments compare the robustness of these estimators against the stochastic Pollaczek-Khintchin formula.

Key Words and Phrases: operational analysis, M/G/1 queue, mean queue length, Pollaczek-Khintchin formula.

This work was supported in part by NSF grant MCS78-01729 at Purdue University.

Introduction

Suppose that we wish to estimate $\bar{n}$, the mean number of jobs present at a single-server queue during a given observation period. Let N denote the maximum number of jobs observed in the system. If we estimate Y(n), the rate of arrivals that find n jobs already in the system, and S(n), the mean time between completions when n jobs are in the system, we can estimate p(n), the proportion of time n jobs are present, by calculating the normalized solution of the birth-death iteration:

$$p(n) = p(n-1)\ Y(n-1)\ S(n)\ , \qquad n = 1,\ldots,N\ . \tag{1}$$

Then we can calculate $\bar{n}$ from $\sum_{n=0}^{N} np(n)$. The quantity p(N) denotes the proportion of time the queue is full.

This analysis can be simplified greatly if we approximate the true arrival function by a constant (i.e., assume Y(n) = Y, the overall arrival rate), if we approximate the true service function by a constant (i.e., assume S(n) = $\bar{s}$, the overall mean service time), and if p(N) is negligible. In this case $Y\bar{s} = U$, the utilization, and

$$p(n) = p(n-1)\ U\ , \qquad n = 1,\ldots,N. \tag{2}$$

Moreover, the estimator for mean queue length now has the familiar closed form:

$$\bar{n} = U\ /\ (1-U)\ . \tag{3}$$

The two paragraphs above summarize operational birth-death

analysis, which is analogous to stochastic M/M/1 analysis but applies to a different set of behavior sequences [BUZE76, BUZE80]. It has been an open question whether there is an operational analog for the Pollaczek-Khintchin formula for the mean length of an M/G/1 queue [e.g., COFF73, KLEI75]. One result of this paper is just such a formula:

$$\bar{n} = U + \frac{U^2(CV^2+1)}{2(1-U)} \qquad (4)$$

where U is the observed utilization, $CV^2 = \sigma^2 / \bar{s}^2$ is the squared coefficient of service times of jobs observed in the queueing system, and p(N) is neglected. This estimate of $\bar{n}$ is exact for any flow-balanced behavior sequence satisfying four homogeneity assumptions to be discussed below.

We conducted an experimental study to test the robustness of this estimator and compare it with other estimators for $\bar{n}$ derived under different operational assumptions. For open queueing systems, we found that the estimator is generally more accurate if p(N) is known. We also found that an estimator based on the mean and variance of interarrival times is consistently more accurate than estimators based on the mean and variance of service times. For closed queueing systems, we found that open-queue estimators generally were not robust except at bottlenecks.

Notation

Table 1 lists the operational measures of a queue for an observation period of length T. (See [BUZE80].) Two important identities that follow immediately from these definitions are the utilization law, $U = X\bar{s}$, and Little's law, $\bar{n} = XR$. Two identities pertaining to arrivers are

$$p_A(n) = p(n)\ Y(n)\ /\ Y_0 \tag{5}$$

and

$$Y/Y_0 = 1\ /\ (1-p(N)) \qquad \text{if } T(N)<T. \tag{6}$$

We will consider only behavior sequences that are both single-step and flow balanced. Single step behavior means that arrivals and completions occur one at a time and no arrival coincides with a completion. Flow balanced behavior means that $A = C$; this is equivalent to $n(0) = n(T)$, to $X = Y_0$, and, for single-step behaviors, to $A(n-1) = C(n)$. For such behavior sequences $p_A(n) = p_C(n)$. These restrictions do not constitute important sources of error.

Since we are interested in relations among mean queue length ($\bar{n}$), throughput (X), mean service time ($\bar{s}$), and the coefficient of variation of service time (CV), we need operational notation for these quantities. A service period is an interval during which a job occupies the server; it is not the same as an intercompletion interval, in which the server can be idle. The service periods are indexed $i = 1,\ldots,C$ and s_i is the length of

TABLE 1. Standard operational quantities of a queue.

Symbol	Definition	Description
N		Maximum observed queue length
n(t)		Queue length at time t ($0 \leq t \leq T$)
A(n)		Number of arrivals who find n(t)=n
C(n)		Number of completers who leave when n(t)=n
T(n)		Total time during which n(t)=n
A	$\sum_{n=0}^{N-1} A(n)$	Total number of arrivals
B	$\sum_{n=1}^{N} T(n)$	Total busy time (excludes idle time, T(0))
C	$\sum_{n=1}^{N} C(n)$	Total number of completions
T	$\sum_{n=0}^{N} T(n)$	Total observation time
W	$\sum_{n=1}^{N} nT(n)$	Accumulated waiting time

TABLE 1 (continued). Standard operational quantities of a queue.

Symbol	Definition	Description
$\overline{n}$	W/T	Mean queue length
R	W/C	Mean response time per completed job
S(n)	T(n)/C(n)	Mean time between completions given n(t)=n
$\overline{s}$	B/C	Mean service time
U	B/T	Utilization
X	C/T	Output rate
Y(n)	A(n)/T(n)	Arrival rate given n(t)=n (n = 0, 1, ..., N-1)
Y_0	A/T	Overall arrival rate
Y	A/(T-T(N))	Restricted arrival rate
p(n)	T(n)/T	Overall queue distribution (n = 0, 1, ..., N)
$p_A(n)$	A(n)/A	Arriver's queue distribution (n = 0, 1, ..., N-1)
$p_C(n)$	C(n+1)/C	Completer's queue distribution (n = 0, 1, ..., N-1)
$\overline{n}_A$	$\sum_{n=1}^{N-1} np_A(n)$	Mean queue seen by arrivers
$\overline{n}_C$	$\sum_{n=1}^{N-1} np_C(n)$	Mean queue left behind by completers

the i-th period. Since $B = \sum_{i=1}^{C} s_i$, the mean time between completions is also the mean service time, $\bar{s}$. The second moment of service time is

$$\overline{s^2} = \frac{1}{C} \sum_{i=1}^{C} s_i^2 \tag{7}$$

and the squared coefficient of variation is

$$CV^2 = \sigma^2 / \bar{s}^2 = (\overline{s^2} - \bar{s}^2) / \bar{s}^2 \ . \tag{8}$$

For convenience we will assume that the observation period is aligned with service completions; i.e., there is a completion just before times 0 and T. This strengthens the flow balance assumption slightly.

Arrivals can be grouped by service period. We let a_i denote the number of arrivals during the i-th service period. We let the binary variable b_i be 1 if the i-th service period begins with an arrival (because the arrival ended an idle period) and 0 otherwise. These quantities are related to the total number of completions by

$$C = \sum_{i=1}^{C} (a_i + b_i) \tag{9}$$

from which it follows that

$$\bar{a} + \bar{b} = 1 \ . \tag{10}$$

We let n_i denote the state n(t) just after the i-th service period; n_0 denotes the state just after t=0. The average value

of n_i is the mean queue left behind by completers (and seen by arrivers):

$$\bar{n}_C = \bar{n}_A = \frac{1}{C} \sum_{i=1}^{C} n_i \ . \tag{11}$$

Note that $n_{i-1}b_i = 0$ and that

$$n_i = n_{i-1} + a_i + b_i - 1 \ . \tag{12}$$

Let K(n,s) denote the number of occurrences of the pattern $(n_{i-1},s_i) = (n,s)$ and let $p(n,s) = K(n,s)/C$ denote the distribution of such patterns. Similarly, let K(n) denote the number of occurrences of $n_{i-1} = n$; because flow is balanced, $K(n) = C(n+1)$ and $p_C(n) = K(n)/C$.

Corresponding to A(n), the number of arrivals that find $n(t) = n$, we define A(s) to be the number of arrivals that come during service periods of length exactly s. We also define K(s), the number of service periods of length exactly s; $p(s) = K(s)/C$ is the proportion of services of length s. The total time spanned by service periods of length exactly s is $T(s) = sK(s)$; the arrival rate measured during such periods is $A(s)/T(s)$; and the proportion of busy time covered by such periods is $T(s)/B$.

Let $j = 1,\ldots,C$ index the arrivals. Associate with arrival j the <u>forward residual</u>, r_j, which either is the time remaining in the service period in which j arrives, or is 0 if arrival j begins a service period. The mean is

$$\bar{r} = \frac{1}{C} \sum_{j=1}^{C} r_j \ .$$

Similarly define the backward residual, r_j', which either is the time since the beginning of the service period in which j arrives, or is 0 if arrival j begins a service period. The mean backward residual is $\bar{r}'$. If arrival j occurs within service period i, $r_j' + r_j = s_i$; if arrival j begins a service period, $r_j' + r_j = 0$. It follows that

$$\sum_{j=1}^{C} (r_j' + r_j) = \sum_{s} s\, A(s). \qquad (13)$$

The above notations are summarized in Table 2. Figure 1 illustrates the major quantities for a single-step, flow balanced behavior sequence.

The Mean Queue Length

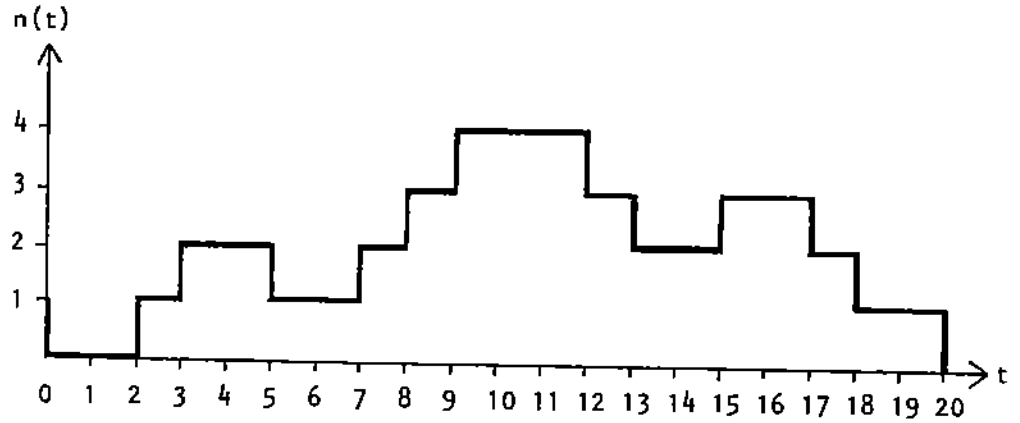
The mean queue length is defined to be

$$\bar{n} = \frac{\text{Area under } n(t) \text{ for } 0<t<T}{T} .$$

As shown in Figure 2, the area under n(t) has two parts: a component depending on queue lengths at the starts of service, weighted by the lengths of service (shaded areas); and a component depending on arrivals during service periods, weighted by forward residuals of service. Since the queue length at the start of the i-th service is $n_{i-1}+b_i$, the first component is the sum of products $(n_{i-1}+b_i)s_i$. Since the j-th arrival is present

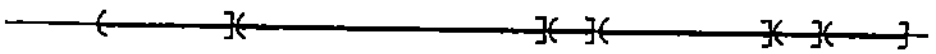
TABLE 2. Service-period oriented operational quantities.

Symbol	Definition	Description
s_i		Length of i-th service period (i=1,...,C)
a_i		Number of arrivals during i-th service
n_i		Queue length just after i-th service
b_i	$\begin{cases} 1 \text{ if } n_{i-1}=0 \\ 0 \text{ otherwise} \end{cases}$	Binary indicator of service periods begun by an arrival
$\bar{s}$, $\overline{s^2}$		First and second moments of s_i
CV^2	$(\overline{s^2}-\bar{s}^2)/\bar{s}^2$	Squared coefficient of variation of s_i
$\bar{a}$, $\overline{a^2}$		First and second moments of a_i
$\bar{b}$, $\overline{b^2}$		First and second moments of b_i
K(n,s)		Number of occurrences of pattern $(n_{i-1},s_i) = (n,s)$ for i=1,...,C
K(n)		Number of occurrences of $n_{i-1} = n$
K(s)		Number of occurrences of $s_i = s$
p(n,s)	K(n,s)/C	Proportion of occurrences of (n_{i-1},s_i) having value (n,s)
$p_C(n)$	K(n)/C	Proportion of occurrences of n_i having value n (completer's distribution)
p(s)	K(s)/C	Proportion of services of length s
A(s)		Number of arrivals within a service period of length s
r_j		Forward residual of service seen by j-th arrival
r_j'		Backward residual of service seen by j-th arrival
$\bar{r}$, $\bar{r}'$		Mean forward and backward residuals



$T = 20$

service periods:



$C = 6$

s_i		3	7	1	4	1	2
n_i	(0)	1	3	2	2	1	0
b_i		1	0	0	0	0	0
a_i		1	3	0	1	0	0

$\bar{s} = 3$

$CV^2 = \frac{13}{27}$

$\bar{n}_C = \frac{3}{2}$

$\bar{b} = \frac{1}{6}$

$\bar{a} = \frac{5}{6}$

n	A(n)	T(n)
0	1	2
1	2	5
2	2	6
3	1	4
4	0	3
	6	20

s	A(s)	K(s)
1	0	2
2	0	1
3	1	1
4	1	1
7	3	1
	5	6

Arrival #	r_j'	r_j
1	0	0
2	1	2
3	2	5
4	3	4
5	4	3
6	2	2
	12	16

$\bar{n} = \sum nT(n)/T = 41/20$

$$\sum_j (r_j' + r_j) = \sum_s sA(s) = 28$$

FIGURE 1. A behavior sequence and its measures.

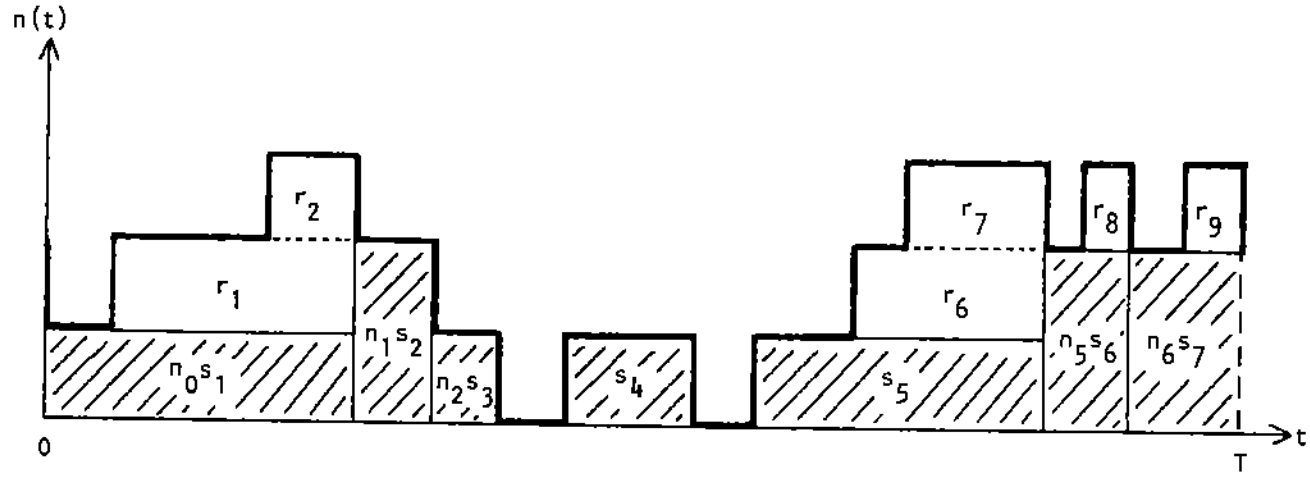


FIGURE 2: Components of the area under n(t)

for time r_j before the next service completion, the second component is the sum of terms r_j. Therefore,

$$\bar{n} = \frac{1}{T} \left(\sum_{i=1}^{C} (n_{i-1}+b_i)s_i + \sum_{j=1}^{C} r_j \right) . \qquad (14)$$

It is of course possible to take measurements of a queueing system (only at arrival and completion times) and determine $\bar{n}$ exactly. It is often desirable, however, to estimate $\bar{n}$ from the quantities $\bar{s}$, CV^2, and U, which can be estimated easily without measuring the queueing system. Further assumptions about behavior sequences are therefore needed.

Homogeneity Assumptions

The definitions impose three constraints on behavior sequences: one-step transitions, flow balance, and alignment of observation period with service completions. None of these assumptions is a significant source of error for behavior sequences of most queueing systems. This section introduces four new assumptions that are sufficient to permit expressing $\bar{n}$ in terms of U, CV^2, N, and p(N). A homogeneity assumption replaces a function by a constant whose value is the mean of the function over its domain.

An important principle underlying these homogeneity assumptions is that the measurements required to determine if

they are true need be taken only at arrival and departure times. They do not depend on any assumptions of finer granularity (e.g., Poisson arrivals).

The first assumption, Homogeneity of Queueing and Service (HQS), permits expressing the first component of mean queue length in Equation 14 in terms of the means $\bar{n}_C$, $\bar{a}$, and $\bar{s}$. The other three assumptions permit expressing the second component of Equation 14 in terms of these means and CV^2 without knowing the details of arrivals. Homogeneity of Arrivals (HA) and Homogeneity of Arrivals and Service (HAS) ignore the possible dependence of arrival rate on queue and service period lengths, respectively. Homogeneity of Residuals (HR) states that there is no overall bias toward forward or backward residuals -- i.e., no overall tendency for arrivals to be bunched at the beginnings or ends of service periods.

> Assumption HQS: Homogeneity of Queueing and Service. The queue length at the start of a service period, $n_{i-1}+b_i$, is independent of the length of that service period, s_i.

Let $p_C(n|s)$ denote the proportion of completions leaving behind $n_{i-1} = n$ given that $s_i = s$. The definitions imply that $p_C(n|s) = K(n,s)/K(s)$. Independence means that $p_C(n) = p_C(n|s)$ -- i.e., that $K(n,s) = K(n)K(s)/C$. Then,

$$\frac{1}{C}\sum_{i=1}^{C} n_{i-1}s_i = \frac{1}{C}\sum_{n,s} n\ s\ K(n,s)$$

$$= \frac{1}{C^2}\sum_{n,s} n\ s\ K(n)\ K(s) \qquad \text{[by HQS]}$$

$$= \sum_{n} n \frac{K(n)}{C} \sum_{s} s \frac{K(s)}{C}$$

$$= \bar{n}_C \bar{s} .$$

A similar argument shows that

$$\frac{1}{C} \sum_{i=1}^{C} b_i s_i = \bar{b} \bar{s} .$$

Hence HQS implies that

$$\frac{1}{C} \sum_{i=1}^{C} (n_{i-1} + b_i) s_i = (\bar{n}_C + \bar{b}) \bar{s} . \qquad (15)$$

Assumption HA: Homogeneity of Arrivals. The conditional arrival rate $Y(n) = A(n)/T(n)$ is a constant independent of n, namely $Y(n) = Y = C/(T-T(N))$ for $n = 0,1,\ldots,N-1$.

This is the homogeneous arrival assumption used by Buzen and Denning [BUZE80]. The given value of constant Y follows from the operational law $\sum_{n=0}^{N-1} Y(n)p(n) = Y_0$. This assumption implies two useful relations among $\bar{n}_C$, $\bar{n}$, $\bar{a}$, $p(0)$, and $p(N)$. By flow balance, $C(n+1) = A(n)$, whence $\bar{n}_C = \sum_{n=1}^{N-1} n \frac{A(n)}{C}$. Assumption HA implies $A(n)/C = T(n)/(T-T(N)) = p(n)/(1-p(N))$, where $p(N)$ is the proportion of time $n(t) = N$; thus

$$\bar{n}_C = \sum_{n=1}^{N-1} \frac{n\, p(n)}{1-p(N)} = \sum_{n=1}^{N} \frac{n\, p(n)}{1-p(N)} - \frac{N\, p(N)}{1-p(N)}$$

so that

$$\bar{n}_C = \frac{\bar{n} - Np(N)}{1-p(N)} \quad . \tag{16}$$

If $p(N) = 0$ then $\bar{n}_C = \bar{n}$; Equation 16 is analogous to the stochastic theorem that arrivers, completers, and outside observers see the same mean queue in an M/G/1 system [BUZE80, COOP72, KLEI75].

Now, the number of occurrences of $b_i = 1$ is the number of arrivals which found the queue empty; therefore, $p_A(0) = \bar{b}$. Equation 10 implies that $\bar{a} = 1-\bar{b}$. Applying the operational laws given by Equations 5 and 6,

$$\begin{aligned} \bar{a} &= 1 - \bar{b} \\ &= 1 - p_A(0) \\ &= 1 - p(0)\ \frac{Y(0)}{Y_0} \\ &= 1 - p(0)\ \frac{Y}{Y_0} \qquad \text{[by HA]} \end{aligned}$$

so that

$$\bar{a} = 1 - \frac{p(0)}{1-p(N)} \quad . \tag{17}$$

If $p(N) = 0$ then $\bar{a} = U$; Equation 17 is analogous to the stochastic theorem that the mean number of arrivals in service periods of an M/G/1 queue is the same as the utilization [COFF73, KLEI75].

Assumption HAS: Homogeneity of Arrivals and Service. The arrival rate in service periods of size s, $A(s)/T(s)$, is a constant independent of s, namely $\bar{a} / \bar{s}$.

This assumption extends the concept of homogeneous arrivals to service periods. The constant $\bar{a} / \bar{s}$ follows from the operational law,

$$\sum_s \frac{A(s)}{T(s)} \frac{T(s)}{B} = \bar{a} / \bar{s} ,$$

which states that the arrival rate for service periods of size s, averaged over the proportion of busy time occupied by such services, is the ratio $\bar{a} / \bar{s}$.

Assumption HR: Homogeneity of Residuals. The total of forward residuals ($\sum_{j=1}^{C} r_j$) equals the total of backward residuals ($\sum_{j=1}^{C} r_j'$).

This assumption assumes that arrivals have no tendency to bunch either toward the start or finish of service periods. It implies in particular that $\bar{r}' = \bar{r}$.

Assumptions HAS and HR imply a useful relation among $\bar{r}$, CV^2, $\bar{s}$, and $\bar{a}$. Applying assumption HR to Equation 13 gives

$$2\bar{r} = \frac{1}{C} \sum_{j=1}^{C} (r_j' + r_j) = \sum_s s \frac{A(s)}{C} .$$

Assumption HAS implies that $A(s) = T(s) \bar{a} / \bar{s}$. The definitions imply that $T(s)/C = sK(s)/C = sp(s)$. Therefore

$$\sum_s s \frac{A(s)}{C} = \sum_s s^2 p(s) \bar{a} / \bar{s}$$
$$= \overline{s^2} \bar{a} / \bar{s} .$$

By definition, $CV^2 = \overline{s^2} / \bar{s}^2 - 1$. Therefore,

$$\bar{r} = \bar{a}\,\bar{s}\,\left(\frac{CV^2+1}{2}\right) . \qquad (18)$$

This is analogous to the expression for mean forward recurrence time in a stochastic renewal process [COFF73, KLEI75].

The four assumptions (HQS, HA, HAS, and HR) are independent of each other. It is possible to contrive behavior sequences having any three of these properties, but not the fourth. Figure 3, which will be discussed shortly, is an example of a behavior sequence satisfying all the assumptions of this paper.

Equations 15-18, the consequences of the four homogeneity assumptions, will be used next to derive an expression for $\bar{n}$.

The Formula for $\bar{n}$

Noting that throughput $X = Y_0 = C/T$, Equation 14 can be rewritten in the form

$$\bar{n} = X \frac{1}{C} \sum_{i=1}^{C} (n_{i-1}+b_i)s_i + X\,\bar{r} .$$

Applying Equations 15 and 18, this reduces to

$$\bar{n} = X\,(\bar{n}_C + \bar{b})\,\bar{s} + X\,\bar{a}\,\bar{s}\,\frac{CV^2+1}{2} .$$

Recalling that $U = X\bar{s}$, this reduces to

$$\bar{n} = U\ (\bar{n}_C + \bar{b} + \bar{a}\ \frac{CV^2+1}{2})\ .$$

On substituting the expressions for $\bar{n}_C$ (Equation 16) and $\bar{a}$ (Equation 17), and solving for $\bar{n}$,

$$\bar{n} = U\ \frac{1-U-Np(N)}{1-U-p(N)} + \frac{U(U-p(N))(CV^2+1)}{2(1-U-p(N))}\ . \qquad (19)$$

This estimator of $\bar{n}$ is exact if the behavior sequence is flow balanced, one-step, service aligned, and satisfies the four homogeneity assumptions.

If the proportion of time the queue attains its maximum observed value is negligible, we can assume $p(N) = 0$ and simplify the estimator. If $p(N) = 0$,

$$\bar{n} = U + \frac{U^2(CV^2+1)}{2(1-U)}\ . \qquad (20)$$

This equation is an operational counterpart of the Pollaczek-Khintchin (PK) formula [COFF73, KLEI75].

If $CV^2 = 1$, Equation 19 reduces to

$$\bar{n} = \frac{U}{1-U-p(N)}\left[1 - (N+1)p(N)\right]\ , \qquad (21)$$

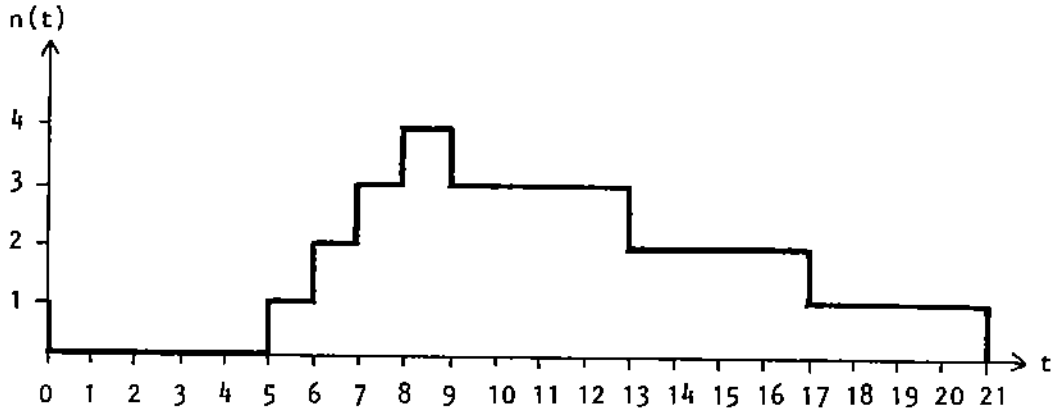
which is identical to the formula for a behavior sequence with homogeneous arrivals (HA) and homogeneous service times (HST -- i.e., $S(n) = \bar{s}$) [BUZE80]. However, any HA/HST behavior sequence will satisfy this formula even if $CV^2 \neq 1$. It is unknown whether a behavior sequence having $CV^2 = 1$ and satisfying all the assumptions of this paper must also have HST.

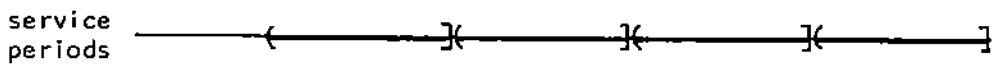
Figure 3 shows a behavior sequence satisfying all the

assumptions of this paper. All the service times are the same, so that $CV^2 = 0$ and assumption HAS is automatically satisfied. Since $A(n)/T(n) = 1/5$ for $n = 0,\ldots,N-1$, assumption HA is satisfied. The sums of forward and backward residuals are 5 each, so HR is satisfied. The sum $\sum_{i=1}^{4} (n_{i-1}+b_i)s_i$ is the same as $(\bar{n}_c+\bar{b})\bar{s}$, so HQS is satisfied. In this case Equation 19 for $\bar{n}$ evaluates to the true value of the mean queue length ($\bar{n} = 34/21$).

The behavior sequence of Figure 3 can be extended by repeating the pattern indefinitely. The resulting behavior sequence, being periodic, has no steady-state limit; and, being deterministic, it will fail any statistical goodness-of-fit test for exponential interarrival times at any given level of confidence. In other words, the extended replication of Figure 3 is a non-steady-state, non-Poisson-arrival behavior sequence for which the operational estimator of $\bar{n}$ is exact. There would be no reason to believe that the stochastic PK formula would apply to this case.

The extended replication of Figure 3 is contrived to make the point that the operational assumptions for the $\bar{n}$ formula are weaker that their stochastic counterparts. It is easy to imagine a deterministic system with the prescribed behavior; given the knowledge that the system is deterministic, no reasonable observer could explain the extended replication by postulating an M/G/1 stochastic process. Nevertheless, the formulae derived in stochastic queueing theory can be applied in this case because they are valid under operational assumptions that are satisfied.





s_i	4	4	4	4	$\bar{s} = 4$, $CV^2 = 0$
n_i	3	2	1	0	$\bar{n}_C = \frac{3}{2}$
b_i	1	0	0	0	$\bar{b} = \frac{1}{4}$
a_i	3	0	0	0	$\bar{a} = \frac{3}{4}$

s	A(s)	T(s)
4	3	16

HAS satisfied

n	A(n)	T(n)
0	1	5
1	1	5
2	1	5
3	1	5
4	0	1

HA satisfied

Arrival #	r_j'	r_j
1	0	0
2	1	3
3	2	2
4	3	1

HR satisfied

$$\frac{1}{4}\sum_{i=1}^{4}(n_{i-1}+b_i)s_i = \frac{4+12+8+4}{4} = \frac{28}{4} = 7$$

$$(\bar{n}_C+\bar{b})\bar{s} = \left(\frac{3}{2}+\frac{1}{4}\right)\cdot 4 = \frac{7}{4}\,4 = 7$$

HQS satisfied

FIGURE 3. Behavior sequence satisfying the four assumptions.

We note that sufficiently long behavior sequences from a steady-state Poisson-arrival queue will satisfy the assumptions of this paper. In such cases both the operational and stochastic formulae for $\overline{n}$ will give the same (correct) answer for mean queue length.

We note also that the foregoing analysis works for multi-server queues if the "service period" is interpreted as the busy period preceding a departure. In this case, $\overline{s}$ is the mean time between departures, not the mean job service time, and may be difficult to estimate.

Granularity of Analysis

The operational estimator for $\overline{n}$ depends altogether on seven assumptions: single-step, flow balance, service alignment, and the four homogeneity assumptions (HQS, HA, HAS, and HR). In contrast, the stochastic estimator for $\overline{n}$, the PK formula for an M/G/1 queue, is conceptually simpler because it depends on three assumptions: steady-state, Poisson arrivals, and independence of number of arrivals and queue length [SAAT61].

There are two essential differences between the operational and stochastic approaches. The first difference is that operational formulae relate parameters of individual behavior sequences whereas stochastic formulae relate parameters of

ensembles of behavior sequences. Operational analysis shows that many properties associated with ensembles are in fact properties of behavior sequences separately.

The second difference is in the granularity of time. The four operational homogeneity assumptions are constraints on aggregate measures conditioned on either queue length or service period: measurements to verify (or refute) them can be completely specified as actions to be taken at arrival or completion events. These assumptions do not constrain the behavior of the system between arrival or completion events. In contrast, the postulates of the Poisson process constrain the behavior of the system in every infinitesimal interval of time.

In fact, the Poisson postulates at the infinitesimal level imply the operational homogeneity assumptions at the arrival/completion level (for sufficiently long behavior sequences). The converse is not true.

Robustness

The formula for $\overline{n}$ may perform poorly when applied to a server embedded in a closed system. The reason is that the feedback inherent in a closed system can destroy one or more of the homogeneity assumptions. The formula for $\overline{n}$ works best in open systems.

Consider the case of a high-CV server in a closed network where other servers have low CV. When a very long job occupies the given server, the queue will build up to its maximum of N; after the long job leaves, the queue will empty as the backlog of short jobs completes. This behavior will tend to associate the large values of n_{i-1} with small s_i, and the small values of n_{i-1} with the large s_i -- violating HQS. Almost all arrivals will be observed while the long job holds the server -- violating HAS. If the delay through the rest of the network is short, arrivals will tend to be bunched at the beginning of the long job's service period -- violating HR.

For these reasons, we looked for alternate estimators of $\bar{n}$ that depend on less restrictive homogeneity assumptions. We also undertook an experimental study of the robustness of various estimators for $\bar{n}$.

Alternate Formulae for Mean Queue Length

Alternate derivations for mean queue length formulae are based on the recursion $n_i = n_{i-1} + a_i + b_i - 1$ (Equation 12); they are analogous to Saaty's argument [SAAT61]. They lead to mean queue length estimators relying on fewer homogeneity assumptions than the previous estimator.

Squaring both sides of Equation 12 and then applying the

identities $n_{i-1}b_i = 0$ and $b_i^2 = b_i$, we have

$$n_i^2 = (n_{i-1} + a_i + b_i - 1)^2 ,$$
$$= n_{i-1}^2 + a_i^2 - b_i + 1 + 2a_i(n_{i-1}+b_i) - 2n_{i-1} - 2a_i .$$

Summing both sides over i = 1,...,C, dividing by C, invoking flow balance, and applying $\bar{a} = 1-\bar{b}$ (Equation 10) gives

$$\bar{n}_C = \frac{\overline{a^2} - \bar{a}}{2} + \frac{1}{C} \sum_{i=1}^{C} a_i(n_{i-1}+b_i) . \qquad (22)$$

Equation 22 is exact for one-step, service aligned, flow balanced behavior sequences. To convert this to a more convenient estimator, we introduce a new assumption:

> Assumption HAQ: Homogeneity of Arrivals and Queueing. The number of arrivals during a service period, a_i, is independent of the queue length at the beginning of that service period, $n_{i-1}+b_i$.

An argument analogous to that used at Equation 15 shows that HAQ implies

$$\frac{1}{C} \sum_{i=1}^{C} a_i(n_{i-1}+b_i) = \bar{a} \ (\bar{n}_C + \bar{b}) .$$

Applying this result in Equation 22 and solving for $\bar{n}_C$,

$$\bar{n}_C = \bar{a} + \frac{\overline{a^2} - \bar{a}}{2(1-\bar{a})} . \qquad (23)$$

This estimator of $\bar{n}_C$, which depends on the first two moments of

a_i, requires just one homogeneity assumption. (Our previous estimator for $\bar{n}_C$ depends on four homogeneity assumptions.)

Under the assumptions of homogeneous arrivals (HA), Equations 16 and 23 yield a solution for $\bar{n}$. If $p(N) = 0$, then $\bar{n} = \bar{n}_C$ (Equation 16) and $\bar{a} = U$ (Equation 17); Equation 23 becomes

$$\bar{n} = U + \frac{\overline{a^2} - U}{2(1-U)} . \tag{24}$$

This agrees with the previous estimate only if $\overline{a^2} = U^2 (CV^2+1) + U$. We know of no assumptions that force this to be true. We can, however, force this to be approximately true under this assumption:

Assumption LA: Linear Arrivals. The number of arrivals within a service period is directly proportional to that period's length; that is, there is a constant H such that $a_i = H s_i$ for all i. (Obviously, $H = \bar{a} / \bar{s}$.)

It is easy to see that $\overline{a^2} = H^2 \overline{s^2} = \bar{a}^2 (CV^2+1)$ according to assumption LA. In this case Equation 23 reduces to

$$\bar{n}_C = \bar{a} + \frac{\bar{a}^2(CV^2+1) - \bar{a}}{2(1-\bar{a})} . \tag{25}$$

The homogeneous arrivals (HA) assumption, with $p(N) = 0$, reduces Equation 25 to

$$\bar{n} = U + \frac{U^2(CV^2+1) - U}{2(1-U)} , \tag{26}$$

which is similar to Equation 20 but relies on three homogeneity assumptions (HAQ, LA, and HA).

Another estimator for $\bar{n}$ arises from a modified linear arrivals assumption:

> Assumption MLA: Modified Linear Arrivals. The number of arrivals associated with a service period is directly proportional to that period's length; that is, there is a constant H' such that $a_i+b_i = H's_i$ for all i. (Obviously, $H' = 1 / \bar{s}$.)

If we write $a_i = s_i / \bar{s} - b_i$, square both sides, and take the mean, we obtain

$$\overline{a^2} = CV^2 + 1 - \frac{2\overline{sb}}{\bar{s}} + \bar{b} .$$

Assumption HQS implies $\overline{sb} = \bar{s}\,\bar{b}$, whereupon this expression simplifies to

$$\overline{a^2} = CV^2 + \bar{a} .$$

With this, Equation 23 simplifies to

$$\bar{n}_C = \bar{a} + \frac{CV^2}{2(1-\bar{a})} . \qquad (27)$$

The homogeneous arrivals assumption (HA), with p(N) = 0, simplifies this to

$$\bar{n} = U + \frac{CV^2}{2(1-U)} . \qquad (28)$$

Table 3 summarizes the estimators discussed above. All these formulae rest on the basic assumptions that behavior sequences are one-step, service aligned, and flow balanced.

Empirical Results

We constructed a program to simulate a single queue in isolation. Inputs to the simulator were the number of service periods to be generated, the desired mean and coefficient of variation of the interarrival times, and the desired mean and coefficient of variation of service times. Interarrival times and service period lengths were drawn from an Erlang, exponential, or hyperexponential distribution, depending on the specified coefficient of variation. The output of the simulator was a one-step, service aligned, and flow balanced behavior sequence.

Another program measured each behavior sequence. It calculated the actual value of $\overline{n}_c$ as well as the parameters required to estimate this quantity using five of the formulas listed in Table 3: the Pollaczek-Khintchin formula (PK), its operational counterpart (OP-PK), and the three alternate estimators (ALT 1, ALT 2, and ALT 3). The relative errors between the actual value of $\overline{n}_c$ and each estimate were calculated.

Two sets of experiments were performed using these tools. Table 4 summarizes the results of the first experimental set, in

Assumptions	Estimator		Estimator assuming HA and $p(N)=0$	
HQS, HA, HAS, HR	$\bar{n} = U\,\frac{1-U-Np(N)}{1-U-p(N)} + \frac{U(U-p(N))(CV^2+1)}{2(1-U-p(N))}$	[OP-PK]	$\bar{n} = U + \frac{U^2(CV^2+1)}{2(1-U)}$	[PK]
HAQ	$\bar{n}_C = \bar{a} + \frac{\overline{a^2} - \bar{a}}{2(1-\bar{a})}$	[ALT 1]	$\bar{n} = U + \frac{\overline{a^2} - U}{2(1-U)}$	[ALT 1']
HAQ, LA	$\bar{n}_C = \bar{a} + \frac{\bar{a}^2(CV^2+1) - \bar{a}}{2(1-\bar{a})}$	[ALT 2]	$\bar{n} = U + \frac{U^2(CV^2+1) - U}{2(1-U)}$	[ALT 2']
HAQ, MLA, HQS	$\bar{n}_C = \bar{a} + \frac{CV^2}{2(1-\bar{a})}$	[ALT 3]	$\bar{n} = U + \frac{CV^2}{2(1-U)}$	[ALT 3']

TABLE 3: Operationally derived mean queue length estimators.

which the simulator generated behavior sequences corresponding to an M/G/1 queue. Behavior sequences comprising 50, 500, and 2000 service periods were studied. For each number of service periods, 50 behavior sequences were generated, approximately one-third of which had service times drawn from Erlang, another third exponential, and a final third hyperexponential distributions. As the length of a behavior sequence increases, the observed mean queue length will converge to that predicted by the PK formula. Each row of Table 4 shows the results of a group of 50 behavior sequences: the frequency with which the estimator appeared at various ranks, its mean rank, and its mean relative error.

Table 4 shows that none of the estimators is good for short behavior sequences. Behavior sequences of length 1000 were needed to get the errors of the best estimators below 20%; at length 2000, the best errors were still near 18%.

The operational P-K formula gave smaller relative errors than the stochastic analog for 36 of the 50 sequences of length 50, for 35 of the 50 sequences of length 500, and for 32 of the 50 sequences of length 2000. The rate of convergence of the OP-PK estimator to the PK estimator is slow. It depends on the rate at which p(N) approaches 0 (where N increases with sequence length).

The ALT 1 estimator, which relies only on one homogeneity assumption (HAQ), performed as well as or better than the OP-PK estimator. However, its ranking was more variable, a sign of

TABLE 4. Experiment set #1: M/G/1 queue in isolation.

Behavior sequence lengths: 50, 500, and 2000 customers

Interarrival times: Exponential with mean of 1.0

Service times: Equal number of Erlang ($0.2 < CV < 0.8$), exponential ($CV = 1.0$), and hyperexponential ($2.0 < CV < 5.0$), all with mean of 0.8

Behavior Sequence Length	Estimator of $\bar{n}_c$	Frequency with which the estimator ranked 1	2	3	4	5	Mean Rank	Mean Relative Error
50	P-K	3	6	9	16	16	3.72	2.492
	OP P-K	7	10	11	15	7	3.10	2.292
	ALT 1	16	8	12	4	10	2.68	1.261
	ALT 2	17	11	5	5	12	2.68	1.013
	ALT 3	7	15	13	10	5	2.82	1.645
500	P-K	5	12	15	14	4	3.00	0.328
	OP P-K	4	22	21	3	0	2.46	0.296
	ALT 1	16	13	8	9	4	2.44	0.288
	ALT 2	13	1	4	1	31	3.72	0.489
	ALT 3	12	2	2	23	11	3.38	0.575
2000	P-K	7	21	17	5	0	2.40	0.176
	OP P-K	14	19	14	3	0	2.12	0.175
	ALT 1	14	9	19	8	0	2.42	0.176
	ALT 2	6	1	0	7	36	4.32	0.505
	ALT 3	9	0	0	27	14	3.74	0.414

less robustness. All three alternate estimators performed better than the PK estimators on short sequences, probably because they require fewer assumptions.

Table 5 summarizes the results of the second experimental set, in which the simulator generated behavior sequences corresponding to a G/M/1 queue. Each behavior sequence comprised 1000 service periods. Interarrival times had coefficients of variation ranging from 0.2 (Erlang) to 6.0 (hyperexponential). When the coefficient of variation of the interarrival times differed significantly from 1.0, both PK and OP P-K performed poorly.

Formula ALT 3 gave consistently the best performance over all types of arrivals. However, when the coefficient of variation of the service times differs significantly from 1.0 (as in Table 4), ALT 3 may not perform well either.

The overall conclusions from these two sets of experiment are that the OP-PK formula is more accurate than the PK formula (because it takes p(N) into account), that neither of the PK formulas is robust if the arrival coefficient of variation differs much from 1.0, and that ALT 3 works well for any arrival coefficient of variation as long as the service time coefficient of variation is not too different from 1.0. If both arrival and service time coefficients of variation differ significantly from 1.0, none of the formulas works well.

Because the most common cases in which both arrival and

TABLE 5. Experiment set #2: G/M/1 in isolation.

Behavior sequence length: 1000 customers

Interarrival times: Erlang, exponential, and hyperexponential with mean of 1.0

Service times: Exponential with mean of 0.8

Inter-Arrival CV	Estimator of $\bar{n}_c$	Frequency with which the estimator ranked 1	2	3	4	5	Mean Rank	Mean Relative Error
0.2-0.8	P-K	2	2	2	7	37	4.50	0.808
	OP P-K	2	5	5	36	2	3.62	0.790
	ALT 1	18	25	3	4	0	1.86	0.180
	ALT 2	0	1	37	1	11	3.44	0.499
	ALT 3	28	17	3	2	0	1.58	0.164
1.0	P-K	16	7	14	7	6	2.60	0.165
	OP P-K	4	26	17	3	0	2.38	0.160
	ALT 1	6	11	18	13	2	2.88	0.176
	ALT 2	0	3	1	4	42	4.70	0.443
	ALT 3	24	3	0	23	0	2.44	0.161
2.0-6.0	P-K	1	0	31	18	0	3.32	0.737
	OP P-K	0	1	0	31	18	4.32	0.767
	ALT 1	6	4	8	0	32	3.96	1.072
	ALT 2	6	33	11	0	0	2.10	0.329
	ALT 3	37	12	0	1	0	1.30	0.262

service time coefficients of variation differ from 1.0 occur in closed queueing networks, we conducted a third study of our estimators. For this purpose we used Balbo's data [BALB79], which included exact solutions for 24 three-station closed networks with different combinations of service time coefficients of variation at the stations. This data included values for $\bar{n}$ (not $\bar{n}_c$) and contained nothing about $\overline{a^2}$; hence, among the alternate forms, we could compare only ALT 2' and ALT 3' against exact values.

Table 6 summarizes the results. Balbo's 24 networks contain a total of 72 service stations. The number of stations having each of the five values of service time coefficient of variation used in the experiment is shown in Table 6. The mean relative error for each estimator is tabulated according to the service time coefficient of variation of the station at which it is measured. It is clear that only the OP-PK formula gives tolerable approximation, and even then only for service stations whose service time coefficient of variation is less than 2.0.

Table 7 compares the OP-PK estimates with the results of two iterative approximations for closed networks studied by Balbo: Marie's method [MARI79] and the Modified Extended Product Form (MEPF) method [SHUM77]. Both iterative methods produced significantly smaller errors than the open-queue formula OP-PK.

TABLE 6. Experiment set #3: 3-station closed networks [BALB79].

Topologies: 15 central server and 9 fully connected networks

Number of customers: 6

Input parameters: Service time CV, U, N, and p(N)

CV of service time: 0.6, 1.0, 2.0, 5.0, or 10.0

Service Time CV	Number of Stations	Relative error in estimate of $\bar{n}$			
		P-K	OP P-K	ALT 2'	ALT 3'
0.6	8	7.491	0.278	2.270	2.068
1.0	40	2.483	0.205	1.447	1.781
2.0	10	2.323	0.362	1.795	4.648
5.0	4	8.403	3.183	7.916	17.984
10.0	10	33.902	10.453	33.423	80.226
ALL	72	7.710	1.824	6.388	14.006

TABLE 7. Comparison of OP-PK with closed network estimators.

Service Time CV	Relative error in estimate of $\bar{n}$		
	OP-PK	MEPF Approx	Marie's Approx
0.6	0.278	0.046	0.039
1.0	0.205	0.027	0.030
2.0	0.362	0.095	0.069
5.0	3.183	0.142	0.028
10.0	10.453	0.078	0.021

Conclusion

We have derived an operational counterpart of the traditional Pollaczek-Khintchin (PK) formula for the mean queue length at an M/G/1 queue. Our formula (OP-PK) is exact for flow-balanced behavior sequences that satisfy four homogeneity assumptions. It takes into account p(N), the proportion of time the queue is at its maximum observed length. If p(N) can be neglected (as it may for a very long behavior sequence), our formula has the same form as the stochastic Pollaczek-Khintchin formula for an M/G/1 queue (although the symbols have different interpretations). Experimental studies reveal that the operational formula tends to be more accurate than the stochastic formula. The operational formula extends to multiserver queues if the statistics $\overline{s}$ and CV are measured for the busy period preceding each completion.

The operational formulas relate parameters measurable in any given behavior sequence; their homogeneity assumptions constrain the behavior at the same granularity of time as measurements of parameters are taken. In contrast, the stochastic formulas relate parameters of ensembles of behavior sequences; their Markovian assumptions constrain the behavior in every infinitesimal interval of time.

Three alternate estimators for mean queue length were derived. Each is based on different homogeneity assumptions. One of the alternates, which is based on the mean interarrival time and the coefficient of variation of service times, was found

experimentally to be more robust than any of the others over a wide range of arrival coefficients of variation as long as the service coefficient of variation was not too much different from 1.0. The experimental study showed that none of the formulas worked well for very short behavior sequences -- typically 1000 service periods needed to be observed to get the estimation errors below 20%. We are not aware of stochastic counterparts of these alternate formulas.

When applied to data for queues in closed networks, all the estimators produced large errors whereas iterative algorithms intended for closed networks were considerably more accurate. The operational PK formula performed best. Even in its best case, however, its error was larger than 20% while the closed-network approximations yielded errors less than 3%. These experiments confirm that the assumptions on which open queue analysis depend may be seriously violated in closed networks.

Acknowledgements

We are grateful to Subhash Agrawal for suggesting several improvements to this paper.

References

BALB79 Balbo, G., "Approximate Solutions of Queueing Network Models of Computer Systems," Ph.D. Thesis, Department of Computer Sciences, Purdue University, West Lafayette, In., December 1979.

BUZE76 Buzen, J. P., "Fundamental Operational Laws of Computer System Performance," Acta Informatica, Vol. 7, No. 2, 1976, pp. 167-182.

BUZE80 Buzen, J. P. and P. J. Denning, "Measuring and Calculating Queue Length Distributions," Computer, Vol. 13, No. 4, April 1980, pp. 33-44.

COFF73 Coffman Jr., E. G. and P. J. Denning, Operating Systems Theory, Prentice-Hall, Englewood Cliffs, NJ, 1973.

COOP72 Cooper, R. B., Introduction to Queueing Theory, Macmillian, New York, 1972.

KLEI75 Kleinrock, L. Queueing Systems, Volume I: Theory John Wiley & Sons, New York, 1975.

MARI79 Marie, R. A., "An Approximate Analytical Method for General Queueing Networks," IEEE Transactions on Software Engineering, Vol. SE-5, No. 5, Sept. 1979, pp. 530-538.

SAAT61 Saaty, T. L., Elements of Queueing Theory McGraw-Hill, New York, 1961.

SHUM77 Shum, A. W. C. and Buzen, J. P., "The EPF Technique: a Method for Obtaining Approximate Solutions to Closed Queueing Networks with General Service Times," Proc. 3rd International Symposium on Modeling and Performance Evaluation of Computer Systems, North-Holland, October 1977.