OPERATIONAL RESPONSE TIME FORMULAS AND THEIR SENSITIVITY TO ERROR

Jeffrey A. Brumfield

TR-83-11 August 1983

Abstract. Operational assumptions can be used to derive four different response time formulas for an isolated service center. There are two sources of error in using these formulas for performance prediction: parameter estimation error and assumption violation. An analysis of the error expressions gives insight into the expected accuracy of the response time estimates.

1. INTRODUCTION

Performance evaluation of computer systems has been an important application of queueing theory. Performance analysts use queueing models to study the effects of proposed changes to existing systems and to predict the performance of systems being developed. Validations have shown that queueing models reproduce many observed performance quantities with high accuracy.

Operational analysis [DENN78] provides a framework for studying queueing phenomena during finite time periods. Operational assumptions have been used to derive many common queueing formulas for isolated queues and networks of queues. Because the violation of operational assumptions can be easily quantified, the operational framework is convenient for studying the sensitivity of formulas to assumption error. Suri [SURI83] has recently shown that performance measures for product form queueing networks are not sensitive to small violations in one of the operational assumptions.

In this paper, we summarize the more important operational formulas for the response time at an isolated service center and the assumptions needed to derive them. We then define measures for parameter and assumption errors and express the errors in the response time formulas in terms of these errors. Finally, we analyze the structures of the error expressions to gain insight into the expected accuracy of each formula.

The main goal of this paper is to give insight into the applicability of queueing formulas. The amount of notation has been minimized; detailed derivations of the formulas have been omitted. Derivations of all results presented in this paper can be found in BRUM82.

Table 1 summarizes some important performance quantities and the operational variables we will use to represent them. The state of an isolated service center can be characterized by the number of customers at the center. A graph of the number of customers during a time period is called a behavior sequence. If at most one arrival or one departure occurs at a time, all the performance quantities in Table

Table 1. Operational quantities for an isolated service center.

$\overline{n}$	mean queue length
R	mean customer response time
S	mean customer service time
CV	coefficient of variation of service times
U	utilization
X	mean completion rate (throughput)
N	maximum queue length
$p(N)$	proportion of time queue length is N

1 can be determined directly from the behavior sequence.

2. OPERATIONAL ASSUMPTIONS

Operational assumptions place restrictions on system behavior during a time period. These assumptions are said to be testable because we can theoretically determine from measured data whether each assumption is satisfied. In reality, we would seldom test whether operational assumptions hold. The detailed values needed to verify the assumptions would allow the direct computation of most performance quantities. For future time periods the data is not available. Also, we would not expect operational assumptions to hold exactly for a given time period.

There are several types of operational assumptions. *Basic assumptions* are used to give the desired meanings to operational variables. The quantity S is defined as the total server busy time divided by the number of completions. For this to be the mean customer service time, no customers may be "in service" at the beginning or end of the time period. Sufficient conditions for this are that the time period begin and end with a customer completion and that a non-preemptive, non-shared queueing discipline be used. The quantity R is defined as the total customer waiting time divided by the number of completions. A sufficient condition for this to be the mean customer response time is that the time period begin and end with no customers in the queue.

The *flow balance assumption* requires that the customer arrival rate is equal to the completion rate. Equivalently, the initial queue length is equal to the final queue length.

Independence assumptions (also called homogeneity assumptions) assert that two sets of quantities are independent. Given a sequence of pairs (x_1,y_1), ..., (x_k,y_k), the x's and y's are independent if for all x and y,

$$p(x_i{=}x \;\&\; y_i{=}y) \;=\; p(x_i{=}x) \cdot p(y_i{=}y) \;\;,$$

where $p(\cdot)$ denotes the proportion of occurrences of a condition. For example, the homogeneity of queueing and service (HQS) assumption requires that the service period lengths be independent of the queue lengths at the beginnings of the service periods. In the above definition, if the x_i are distinct then the y_i must be identical. An example of this restricted type of independence assumption is the homogeneous arrivals (HA) assumption, that requires customer arrival rates be the same during all observed queue lengths.

Representative arrivals assumptions assert that arriving customers see some of the same performance quantities as a continuous observer of the system. These quantities may include mean queue length, proportion of time the queue is empty, and mean forward service period residual.

All of the independence assumptions and representative arrivals assumptions are summarized in the appendix, along with the additional notation needed for their formal definitions. Behavior sequences have been constructed that satisfy most combinations of these assumptions. No subset of the assumptions seems to imply any other of the assumptions.

3. RESPONSE TIME FORMULAS

Table 2 lists four operational formulas for mean response time. All of these formulas rely on the basic assumptions and on the flow balance assumption; the necessary independence and representative arrival assumptions are shown in the table. Detailed derivations of these formulas appear in BRUM82, BUZE80, and KOWA81. Little's formula, $\bar{n} = RX$, can be used to produce analogous formulas for mean queue length.

Several of the formulas are similar in appearance to well-known results from stochastic queueing theory [KLEI75]. The first two formulas correspond to the formulas for the M/M/1/N and M/M/1/∞ queues; the last formula is similar to the Pollaczek-Khinchine formula for an M/G/1 queue.

The similarities are limited to the structure of the formulas. Variables in stochastic formulas represent steady-state quantities; operational variables represent quantities for a finite time period. Stochastic assumptions place restrictions on underlying processes; operational assumptions place restrictions only on observed data. A detailed comparison of the stochastic and operational frameworks can be found in SEVC79.

Each formula in Table 2 is undefined either when $U = 1$ or when $U = 1\text{-}p(N)$. It has been shown [BUZE80] that $R = SN/2U$ can be used to eliminate the discontinuity in formula (1). Similar results have not been discovered for the other formulas. There appear to be no physical manifestations of these discontinuities. Behavior sequences can be constructed that satisfy the assumptions and have finite

Table 2. Response time formulas.

	Formula	Assumptions
(1)	$R = \frac{S}{1-U-p(N)} [\, 1 - (N+1)p(N) \,]$	HA, HS
(2)	$R = \frac{S}{1-U}$	RA1, HS
(3)	$R = S \frac{1-U-Np(N)}{1-U-p(N)} + \frac{S\,(U-p(N))\,(CV^2+1)}{2\,(1-U-p(N))}$	HA, HAS, HR, HQS
(4)	$R = S + \frac{US\,(CV^2+1)}{2\,(1-U)}$	RA1, RA2, RA3, HQS

response times. Yet, the formulas cannot be applied.

When $CV^2 = 1$, formula (3) simplifies to formula (1). However, the assumptions needed to derive formula (3) together with the fact that $CV^2 = 1$ are not equivalent to the assumptions used in deriving formula (1). The same relationship holds for formulas (4) and (2).

Each response time formula is exact for a set of behavior sequences that satisfy the necessary assumptions. Because $p(N) > 0$, the sets of behavior sequences for formulas (1) and (2) are disjoint and the sets for formulas (3) and (4) are disjoint.

4. SOURCES OF ERROR

Operational formulas are most often used to predict performance measures for future time periods. If estimates of the necessary parameters can be obtained, any of the formulas in Table 2 can be used to compute an estimate of the response time. This estimate will contain error from two sources. First, the assumptions under which the formula holds are unlikely to be satisfied during the time period of interest. Second, the exact values of the parameters will not be known. Understanding how the assumption and parameter errors affect the computed value is important when choosing an estimator.

The first step in analyzing error is to quantify the error in each parameter and assumption. To simplify the analysis, we will not consider errors in the basic assumptions or the flow balance assumption. A detailed analysis of the flow balance assumption can be found in BRUM83.

To measure the error in a parameter, we use the signed relative error in the estimate. For example, the error in the estimate $\hat{U}$ of the utilization is $\epsilon_U = (U-\hat{U}) \; / \; U$. We also use relative error for assumptions that equate scalar quantities. For example, the error in the first representative arrival assumption (RA1) is $\epsilon_{RA1} = (\overline{n}_A - \overline{n}) \; / \; \overline{n}_A$.

For assumptions involving vectors of values, it is easier to measure error using an immediate consequence of the assumption. For example, the consequence of the HA assumption used in deriving the response time formulas is an expression for the mean queue length seen by arriving customers,

$$\overline{n}_A \;=\; \frac{\overline{n} - Np(N)}{1 - p(N)} \quad .$$

The error in assumption HA is then measured by

$$\epsilon_{HA} \;=\; \left(\overline{n}_A - \frac{\overline{n} - Np(n)}{1 - p(N)} \right) \; / \; \overline{n}_A \quad .$$

We have shown [BRUM82] that the magnitudes of the consequence errors are no greater than the maximum component error in the estimate vector.

5. ERROR FORMULAS AND BOUNDS

Once we have quantified the sources of error, we can derive an exact expression for the relative error in each response time estimate in terms of the parameter and assumption errors. In practice, we would not know the exact value of each error measure; we would more likely be able to bound the magnitude of each error and perhaps know its sign. Our purpose in deriving exact formulas is to better understand how the errors interact. Bounds and approximations can be obtained from these exact formulas.

The simplest response time formula, $R = S / (1-U)$, has two parameters and relies on two assumptions. If ϵ_S and ϵ_U are the parameter errors, and ϵ_{RA1} and ϵ_{HS} are the assumption errors, then the relative error in the response time estimate is

$$\epsilon_R = \epsilon_{ASMP} + \epsilon_{PARM} - \epsilon_{ASMP}\,\epsilon_{PARM}$$

where

$$\epsilon_{ASMP} = \left(\frac{1}{1-U} \right) \left\{ \epsilon_{HS} + \left(1-\frac{S}{R}\right) \epsilon_{RA1} - \epsilon_{HS}\,\epsilon_{RA1} \right\}$$

and

$$\epsilon_{PARM} = \epsilon_S + \frac{U}{1-\hat{U}}\,\epsilon_U - \frac{U}{1-\hat{U}}\,\epsilon_U\,\epsilon_S \ .$$

This can be expanded to a sum of 15 terms, each containing from one to four error measures. The factor $1-\frac{S}{R}$ is a weight having value in [0,1].

Error expressions can also be derived for the other response time formulas. The number of terms in the error formulas grows combinatorically with the number of assumptions and parameters. Yet, the error formulas all have a similar structure, which we will analyze in the next section.

To bound the magnitude of the response time error, error measures can be replaced by bounds on their magnitudes and weights can be replaced by 1. Exact parameters can be expressed in terms of estimates and parameter error bounds. If parameter and assumption errors are sufficiently small, terms containing a single error measure may be dominant. If higher order terms are omitted, the bound will only be approximate. For example, in the error formula just considered the magnitude of the error due to assumption violation has the approximate bound

$$|\epsilon_{ASMP}| \lesssim \left(\frac{1}{1-U} \right) \{ |\epsilon_{HS}| + |\epsilon_{RA1}| \} .$$

We have found the bounds produced by this method are often too pessimistic to be useful. Frequently, errors are both positive and negative, and some cancellation occurs. The bounds do not reflect this possibility. Examples can be constructed in which large errors in parameters and assumptions cancel, giving a response time estimate that is surprisingly accurate.

6. ANALYSIS OF ERROR FORMULAS

While the exact formulas for the relative errors in the response time estimates are of limited use in practice, their structures give us important information about assumption and parameter errors. To simplify the analysis of the formulas, we first consider only assumption error and assume exact values of all parameters are known. Then, we study parameter error by assuming all assumptions are satisfied exactly.

If assumption violation is the only source of error, the relative errors in formulas (1) and (3) have the common form

$$\epsilon_R = \left(\frac{1-p(N)}{1-U-p(N)} \right) f(\text{assumption errors}) .$$

The errors in formulas (2) and (4) are of the form

$$\epsilon_R = \left(\frac{1}{1-U}\right) f(\text{assumption errors}) \ .$$

The function f(assumption errors) is a sum of terms, each of which contains one or more assumption errors. Some terms include weights whose values reflect the importance of the error in the assumption. The exact structure of f depends on the response time formula. The equation for ϵ_{ASMP} in section 5 illustrates a simple f function.

Since assumption errors may be positive or negative, the errors may cancel or reinforce. But, in every formula the aggregate assumption error is multiplied by either $\frac{1-p(N)}{1-U-p(N)}$ or $\frac{1}{1-U}$. If the magnitude of this multiplier is greater than 1, the aggregate assumption error will be magnified; if the magnitude of the multiplier is less than 1, the error will be damped.

The multiplier is similar to a condition number in numerical analysis. Time periods for which the multiplier is large are "ill conditioned" in that even small assumption errors may produce large errors in the computed response time. The value of the multiplier is easily computed since it requires the same parameters as the response time formula.

Figure 1 shows a contour plot of the multiplier $\frac{1-p(N)}{1-U-p(N)}$ as a function of $p(0) = 1-U$ and $p(N)$. Since $p(0)$, $p(N)$, and $p(0)+p(N)$ are restricted to values between 0 and 1, the domain of this function is the triangular region having vertices (0,0), (0,1), and (1,0). We observe that the multiplier may assume any positive or negative value except those in the range (0,1). Extreme magnification of the assumption error occurs when $p(N) \approx p(0)$. As $p(0)$ approaches 1 or $p(N)$ approaches 1, magnification diminishes. The magnitude of the multiplier is less than 1 when $p(N) \geq (1+p(0)) \ / \ 2$; in this case the aggregate assumption error is reduced.

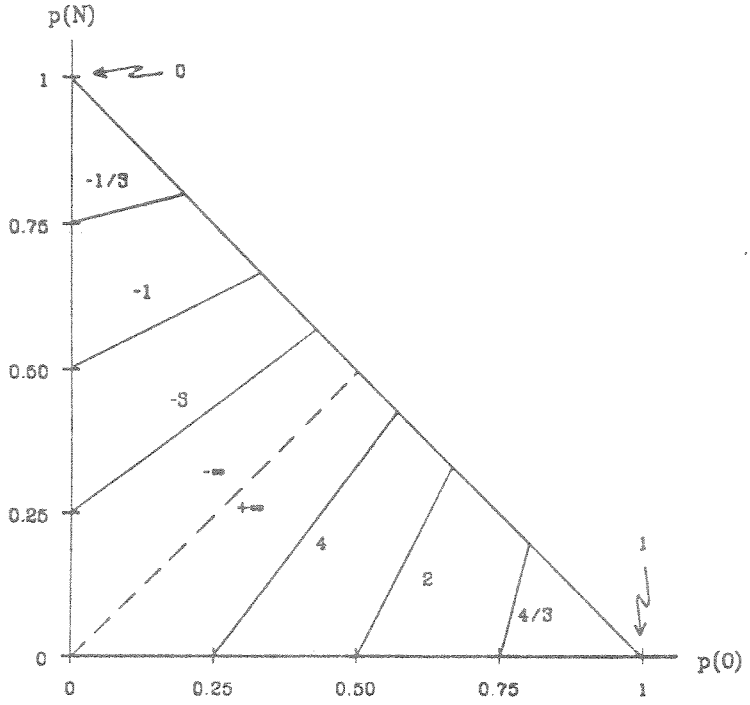

Figure 1. Contour plot of the multiplier $\frac{1-p(N)}{1-U-p(N)}$.

Figure 2 shows a graph of the multiplier $\frac{1}{1-U}$ as a function of U. For comparison with Figure 1, Figure 3 shows a contour plot of the multiplier as function of $p(0)$ and $p(N)$. Because the multiplier is always greater than 1, the aggregate assumption error is always magnified. As U approaches 1, the magnification increases without bounds.

The quantities $p(0)$ and $p(N)$ can usually be measured or estimated more easily than the errors in the assumptions. If assumption errors are not known, it is impossible to determine which formula will produce the most accurate response time estimate. However, if there is a large difference in the magnitudes of the multipliers, it seems reasonable to favor the estimator whose multiplier is smaller.

Comparing the two multipliers shows $\frac{1}{1-U}$ has smaller magnitude when

$$p(N) \leq \frac{2\,p(0)}{1 + p(0)} \ .$$

If $p(0) < 0.25$ and $p(N) \leq 0.35$, both multipliers can still magnify errors by at least a factor of 4.

When all assumptions used in deriving a formula are satisfied, the only source of error is parameter error. The relative error in the response time of formula (4) expressed in terms of the parameter errors ϵ_S, ϵ_U, and ϵ_{CV} is

$$\epsilon_R = \left(\frac{S}{R}\right)\epsilon_S + \left(\frac{1}{R}\,\frac{US\,(CV^2+1)}{2\,(1-U)}\right) \cdot$$

$$\{\ \epsilon_U + \epsilon_S + \frac{CV^2}{CV^2+1}\,\epsilon_{CV} + \frac{\hat{U}}{1-\hat{U}}\,\frac{\epsilon_U}{1-\epsilon_U}$$

$$-\ \text{second order terms} + \text{third order terms}$$

$$-\ \epsilon_U\,\epsilon_S\,\frac{CV^2}{CV^2+1}\,\epsilon_{CV}\,\frac{\hat{U}}{1-\hat{U}}\,\frac{\epsilon_U}{1-\epsilon_U}\ \}$$

If $CV^2 = 1$ and $\epsilon_{CV} = 0$ this reduces to

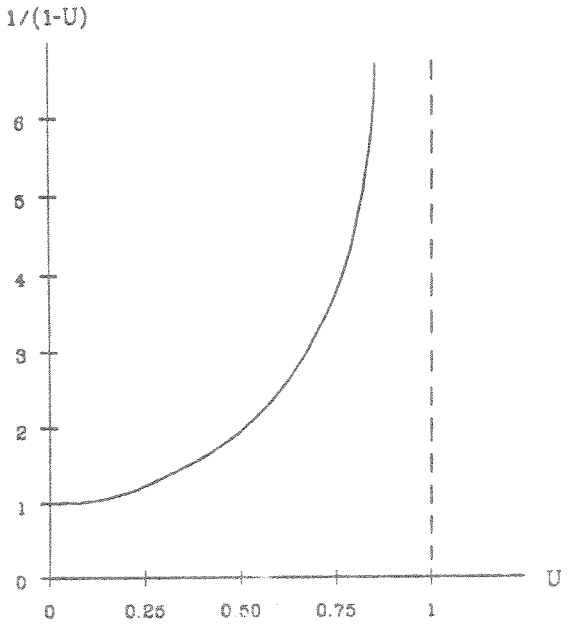

Figure 2. Graph of the multiplier $\frac{1}{1-U}$.

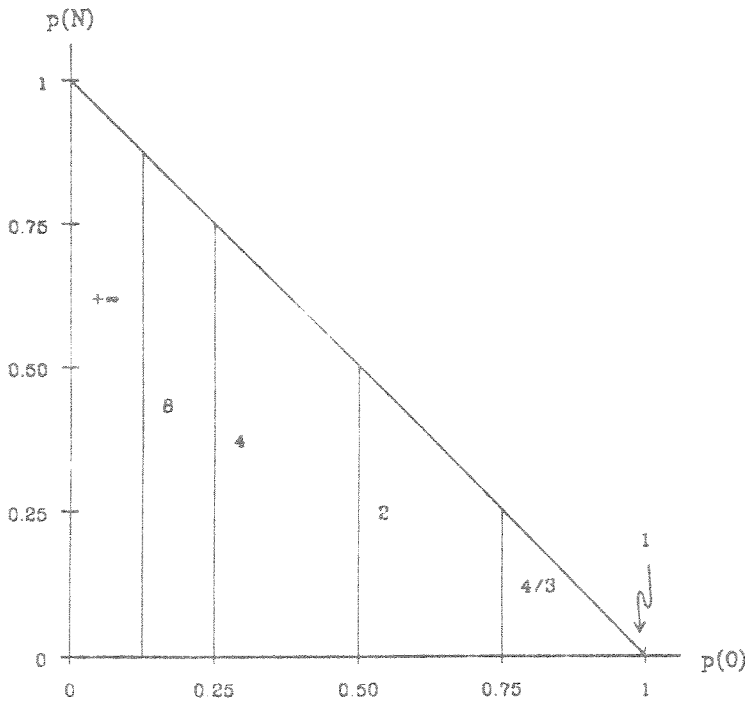

Figure 3. Contour plot of the multiplier $\frac{1}{1-U}$.

$$\epsilon_R = \epsilon_S + \frac{U}{1-\hat{U}}\,\epsilon_U - \frac{U}{1-\hat{U}}\,\epsilon_U\,\epsilon_S ,$$

which is the parameter error for formula (2) presented in section 5.

The quantities $\frac{S}{R}$ and $\frac{1}{R}\,\frac{US\,(CV^2+1)}{2\,(1-U)}$ are weights having values in [0,1]. The quantity $\frac{CV^2}{CV^2+1}$ is always less than 1, but approaches 1 as CV^2 becomes large. This shows that the error in the estimate of CV^2 is less important if CV^2 is close to 0.

While ϵ_S and ϵ_{CV} enter into the error formulas in a simple way, $\hat{U}$ and ϵ_U combine to form a more complicated term. This quantity can be rewritten in terms of U and $\hat{U}$ as follows:

$$\frac{\hat{U}}{1-\hat{U}}\,\frac{\epsilon_U}{1-\epsilon_U} = \frac{U}{1-\hat{U}}\,\epsilon_U = \frac{U-\hat{U}}{1-\hat{U}} .$$

As U approaches 1 this term approaches 1, causing the parameter error in formula (2) to approach 1 regardless of the value of ϵ_S.

Figure 4 shows a contour plot of $\frac{U-\hat{U}}{1-\hat{U}}$. If $\hat{U}$ is an underestimate of U, the term will be bounded by 1; if $\hat{U}$ is an overestimate of U, the magnitude of this term can be arbitrarily large. We see from the graph that if the true utilization is large, it is much better to underestimate U than to overestimate it. As the true utilization approaches 1, small errors in $\hat{U}$ become much more significant.

Parameter errors, like assumption errors, can cancel. Usually we will not be able to anticipate cancellation. To reduce error in the response time estimate, it is important to estimate the parameters as accurately as possible, being careful not to overestimate $\hat{U}$ if utilization is high.

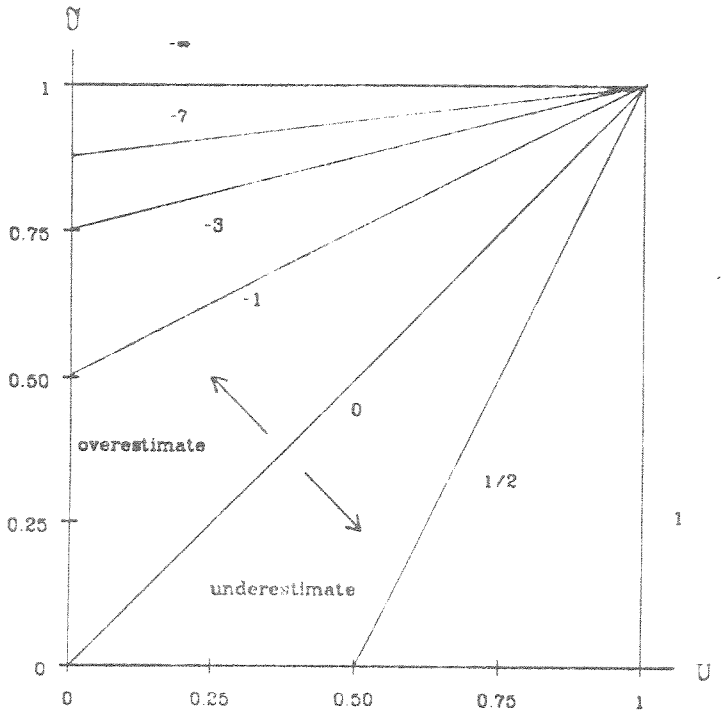

Figure 4. Contour plot of the quantity $\frac{U-\hat{U}}{1-\hat{U}}$.

7. SUMMARY

For time periods that satisfy the necessary assumptions, the four response time formulas presented in this paper express exact relationships among the performance quantities. If the assumptions are not satisfied or the exact values of the parameters are not known, the formulas are only estimators of response time. We have derived exact expressions for the relative error in each response time formula in terms of assumption and parameter errors.

These error expressions reveal a complex interaction among the parameter and assumption errors. Errors may cancel or reinforce, depending upon their signs. Weights on error measures determine their relative importance. Aggregate assumption errors are multiplied by a factor that may magnify or reduce them. This multiplier may be used like a condition number in numerical analysis.

There are several directions for future research. An experimental study could investigate how well operational assumptions are satisfied by real systems and whether assumption errors tend to cancel. These results, together with the formulas in this paper, could help explain the apparent success of queueing models.

In some cases, performance analysts may know the signs of errors and perhaps upper and lower bounds on their values. This additional information should allow us to more tightly bound the relative errors in the response time estimates. Such formulas have not yet been derived. Finally, the techniques in this paper may be applicable to other queueing formulas, including those for queue length distributions and for queues embedded in closed networks.

Acknowledgments

Peter J. Denning provided valuable guidance throughout the course of this research. Discussions with Wolfgang Kowalk and Rajan Suri helped to clarify many of the ideas in this paper.

This research was supported in part by National Science Foundation Grant MCS78-01729 at Purdue University and a University Research Institute award at the University of Texas at Austin.

References

BRUM82 J. A. Brumfield, "Operational Analysis of Queueing Phenomena," Ph.D. Thesis, Department of Computer Science, Purdue University, December 1982.

BRUM83 J. A. Brumfield and P. J. Denning, "Operational State Sequence Analysis," in *Performance '83*, Proceedings of the 9th International Symposium on Computer Performance Modelling, Measurement, and Evaluation, North-Holland Publishing Company, New York, 1983, pp. 269-283.

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

DENN78 P. J. Denning and J. P. Buzen, "The Operational Analysis of Queueing Network Models," *Computing Surveys*, Vol. 10, No. 3, September 1978, pp. 225-261.

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

KOWA81 W. Kowalk, "Extensions of Operational Analysis," in *Messung, Modellierung und Bewertung von Rechensystemen*, Informatik-Fachberichte No. 41, Springer, February 1981.

SEVC79 K. C. Sevcik and M. M. Klawe, "Operational Analysis Versus Stochastic Modelling of Computer Systems," *Proceedings of Computer Science and Statistics: 12 Annual Symposium on the Interface*, University of Waterloo, Ontario, Canada, May 1979.

SURI83 R. Suri, "Robustness of Queueing Network Formulas," *Journal of the ACM*, Vol. 30, No. 3, July 1983.

Appendix 1. Additional operational notation.

Symbol	*Description*
$S(n)$	Mean busy time between completions when queue length is n
$Y(n)$	Mean arrival rate when queue length is n
Y	Overall mean arrival rate
$Y_S(x)$	Mean arrival rate during service periods of length x
Y_S	Mean arrival rate during a service period
$\overline{n}_A$	Mean queue length seen be arrivers
$p_A(0)$	Proportion of arrivals occurring when queue is empty
$p(0)$	Proportion of time queue is empty
$\overline{r}$	Mean forward service period residual
$\overline{r}'$	Mean backward service period residual
$\overline{r}_A$	Mean forward service period residual seen by arrivals
$\overline{q}$	Mean queue length at beginning of service period
$\overline{qs}$	Mean of service period lengths times initial queue lengths
$\overline{s}$	Mean service period length

Appendix 2. Operational assumptions.

Independence Assumptions

HS	Homogeneity of service	$S(n) = S \quad 1 \leq n \leq N$
HA	Homogeneity of arrivals	$Y(n) = Y \quad 0 \leq n \leq N-1$
HQS	Homogeneity of queueing and service	$\overline{q}\,\overline{s} = \overline{qs}$
HAS	Homogeneity of arrivals and service	$Y_S(x) = Y_S \quad \forall x$
HR	Homogeneity of residuals	$\overline{r} = \overline{r}'$

Representative Arrival Assumptions

RA1	Representative mean queue length	$\overline{n}_A = \overline{n}$
RA2	Representative empty queue proportion	$p_A(0) = p(0)$
RA3	Representative service period residual	$\overline{r}_A = \overline{r}$