Operational Treatment of Queue Distribution and Mean Value Analysis

Jevvrey P. Buzen

Peter J. Denning

Report Number:
79-309
OPERATIONAL TREATMENT OF QUEUE DISTRIBUTIONS
AND MEAN VALUE ANALYSIS*

Jeffrey P. Buzen
BGS Systems, Inc.

and

Peter J. Denning
Computer Sciences Department
Purdue University

August 1979

CSD-TR 309

*This work supported in part by NSF Grant MCS78-01729 at Purdue University
Operational Treatment of Queue Distributions
and Mean Value Analysis

ABSTRACT

Relationships among the queue length distributions seen by an arriving job, a completing job, and an outside observer are derived using operational analysis. A simplified derivation of the Sevcik-Mitrani theorem is then presented and used as the basis for discussing Reiser and Lavenberg's mean-value analysis. Two new results are presented: an algorithm for computing queue length distributions from conditional throughputs in closed, product form queueing networks; and an operational bound on the errors that can arise in certain theorems when homogeneity is violated.
Overview

Queueing theory is the mathematical foundation for most analytic models of computer performance. In this paper we use operational analysis to study three basic queueing distributions: the queue length seen by an arriving job, by a completing job, and by an outside observer.

We derive relationships among these distributions for individual queues, and for queues in closed networks with product form solutions. For the latter, we also present a simple derivation for the Sevcik-Mitrani theorem, which states that an arriving job sees the same distribution as an outside observer studying the same network with the arriving job deleted [SEVC78; also LAVE79].

We use the Sevcik-Mitrani theorem to motivate mean value analysis, a recent technique for computing mean response times, throughputs, and queue lengths in closed queueing networks [REIS78, REIS79]. An example of the mean value algorithm is presented, and some comparisons with normalizing-constant analysis are made.

We also present two new results: an algorithm for computing queue length distributions from conditional throughputs in closed networks, and a bound on the errors that can arise in certain theorems when homogeneity is violated. The latter is an operational result with no stochastic counterpart.
Operational Analysis

Operational analysis is a framework for answering questions about the performance of systems during given periods of time. The systems may be real or hypothetical, and the time periods may be past, present, or future.

Operational variables represent quantities that can be measured by observing a system during a time period. Operational analysis is primarily concerned with mathematical relationships among operational variables. These relationships may depend on assumptions that are also expressed in terms of operational quantities. The underlying principle is that variables stand for observable values, and that all assumptions about systems during time periods are experimentally verifiable. (It is not necessary, however, to actually observe values or verify assumptions for an operational analysis to make sense.)

Operational analysis was introduced in 1976 as an alternative to stochastic modeling [BUZZ76]. Many informal, intuitive arguments used to motivate stochastic theorems become rigorous proofs in the formal context of operational analysis. Besides simplifying derivations, operational analysis extends stochastic theorems by demonstrating their validity in cases where conventional stochastic assumptions cannot be justified. Moreover, operational analysis has led to new results about sensitivity factors and error bounds. These results are particularly valuable for prediction because the future validity of operational (or stochastic) assumptions is usually uncertain. DENN78 describes operational analysis in detail and provides many references.
The Three Distributions at a Queue

Associated with every general service system are three distributions of queue length:

1. The overall distribution, \( p(n) \), which gives the proportion of time \( n \) jobs are in the system;

2. The arrivals distribution, \( p_A(n) \), which gives the fraction of arrivals who find \( n \) other jobs in the system; and

3. The completers distribution, \( p_C(n) \), which gives the fraction of completions who left behind \( n \) other jobs in the system.

Each distribution corresponds to a different method of observing the queue.*

Figure 1 illustrates the length \( n(t) \) of a queue during a 10-second observation containing 3 arrivals (marked by "A") and 4 completions (marked by "C"); the table accompanying the figure shows that three distributions may all be different.

Most queueing statistics -- e.g., mean or variance of the queue length -- are expressed relative to the overall distribution, \( p(n) \), because this distribution specifies the state occupancies seen by an outside observer. Yet it is sometimes easier to derive one of the other two distributions during an analysis. The embedded Markov-chain analysis of the \( M/G/1 \) queue, for example, yields the distribution \( p_C(n) \), while for a \( G/M/1 \) queue it yields the distribution \( p_A(n) \). (See COOP72 or KLEI75.) This is why

*Cooper refers to \( p(n) \) as the outside observers' distribution, to \( p_A(n) \) as the arriving customers' distribution, and to \( p_C(n) \) as the departing customers' distribution [COOP72].
Figure 1. Example showing that the three distributions may be different.
queueing analysts are interested in the relationships among these three distributions.

We begin by studying a single-resource queueing system in isolation. Later we will generalize to closed networks of such queues.

A single resource queueing system, which has one queue and a server, is observed for an interval \([0, T]\). The state of the queue at time \(t\), denoted \(n(t)\), gives the number of jobs present (waiting for or receiving service). It varies from a minimum of 0 to a maximum of \(N\) during \([0, T]\). (A nonzero minimum on observed queue length changes the boundary condition but not the nature of the results.) A record of \(n(t)\) for \(0 \leq t \leq T\) is called a behavior sequence, or simply a behavior, of the system.

Define these operational quantities for a given behavior sequence:

\[
A(n) = \text{The number of arrivals who find } n(t) = n, \quad 0 \leq n < N.
\]

\[
C(n) = \text{The number of completions who left when } n(t) = n, \quad 0 \leq n < N.
\]

\[
T(n) = \text{The total time during which } n(t) = n, \quad 0 \leq n < N.
\]

Grand totals are defined as follows:

\[
\sum_{n=0}^{N-1} A(n) \quad \sum_{n=1}^{N} C(n)
\]
\[ T = \sum_{n=0}^{N} T(n) \]

Given these basic quantities, the three queue length distributions are:

\[ p(n) = \frac{T(n)}{T} \quad n = 0, \ldots, N \]

\[ p_A(n) = \frac{A(n)}{A} \quad n = 0, \ldots, N-1 \]

\[ p_C(n) = \frac{C(n+1)}{C} \quad n = 0, \ldots, N-1 \]

Note that \( C(n+1) \) is used to define \( p_C \); this is because \( p_C(n) \) refers to the queue size just after a completion whereas \( C(n) \) refers to the queue size just before a completion. Define also

\[ S(n) = \text{mean time between completions when } n(t) = n \]
\[ = \frac{T(n)}{C(n)} \quad \text{(defined only if } C(n)>0) \]

\[ Y(n) = \text{arrival rate when } n(t) = n \]
\[ = \frac{A(n)}{T(n)} \quad \text{(defined only if } T(n)>0) \]

\[ B = \text{total busy time} \]
\[ = T(1) + T(2) + \ldots + T(N) \]

\[ S = \text{overall mean time between completions} \]
\[ = \frac{B}{C} \]

\[ U = \text{utilization} \]
\[ = \frac{B}{T} \]

\[ Y_0 = \text{overall arrival rate} \]
\[ = \frac{A}{T} \]

\[ Y = \text{restricted arrival rate} \]
\[ = \frac{A}{T-T(N)} \quad \text{(defined only if } T(N)<T) \]
$X = \text{output rate}$

$= C/T$

$W = \text{job-seconds of accumulated waiting time}$

$= \sum_{n=1}^{N} n \cdot T(n)$

$Q = \text{mean queue length}$

$= W/T$

$R = \text{mean response time per completed job}$

$= W/C$

These definitions imply the operational laws:

1. $P_{n}(n) = p(n)(Y(n)/Y_0)$ (if $Y(n)$ defined)

2. $Y/Y_0 = 1/(1-p(N))$ (if $T(N)<T$)

3. $Y_0 = \sum_{n=0}^{N-1} p(n) Y(n)$ (for defined $Y(n)$)

4. $S = \sum_{n=1}^{N} P_n(n-1) S(n)$ (for defined $S(n)$)

5. $X = \sum_{n=1}^{N} p(n)/S(n)$ (for defined $S(n)$)

6. $U = SX$ [Utilization Law]

7. $R = Q/X$ [Little's Law]
Each law can be verified by substituting from the preceding definitions and reducing to an identity. These formulae are called laws because they are valid for every possible behavior sequence \cite{BUZ76, DEN78}; logically, they are tautologies. Little's Law (Eq. 7) plays an important role in mean value analysis.

**Flow Balance and One-Step Behavior**

If the behavior sequences on which \( p(n) \), \( P_A(n) \) and \( P_C(n) \) are observed satisfy certain assumptions (hypotheses), additional relationships (theorems) can be derived. The first two assumptions that we shall consider are *flow balance* and *one-step behavior*.

Flow balance means that the overall arrival rate \( Y_0 \) is equal to the output rate \( X \). This is equivalent to assuming that the total number of arrivals \( A \) is equal to the total number of completions \( C \), or that the initial state \( n(0) \) is the same as the final state \( n(T) \).

One-step behavior means that \( n(t) \) can only change in steps of plus or minus one. There is at most one arrival or one completion at any instant; no arrival coincides with a completion.

If \( n(0) = n(T) \) and if \( n(t) \) can only change in steps of plus or minus one, then \( A = C \) and also the number of transitions from state \( n \) to state \( n+1 \) must equal the number of transitions from state \( n+1 \) to state \( n \):

\[
A(n) = C(n+1) \quad n = 0, \ldots, N-1
\]
Combining this observation with the preceding definitions,

\[ p_A(n) = p_C(n) \quad n = 0, \ldots, N-1. \]  

This theorem is the operational counterpart of a well-known stochastic result [KLEI75, p176]; the derivation provides a "trivial" proof of Kleinrock's Problem 5-6 [KLEI75, p232]. Because of Equation (9) we will not study \( p_C \) further.

Now we will derive recursions for both \( p(n) \) and \( p_A(n) \).

When \( 0 < n < N \):

\[ p(n) = \frac{T(n)}{T} \quad \text{(definition)} \]

\[ = \frac{C(n)}{T(n-1)} \frac{T(n)}{C(n)} \frac{T(n-1)}{T} \]

\[ = \frac{A(n-1)}{T(n-1)} \frac{T(n)}{C(n)} \frac{T(n-1)}{T} \quad \text{(by Equation 8)} \]

\[ = Y(n-1) S(n) p(n-1). \]

This is the equation derived by Buzen for the "generalized birth-death process" [BUZE76]. When \( 0 < n < N \):

\[ p_A(n) = \frac{A(n)}{A} \quad \text{(definition)} \]

\[ = \frac{A(n)}{T(n)} \frac{T(n)}{C(n)} \frac{C(n)}{A} \]

\[ = \frac{A(n)}{T(n)} \frac{T(n)}{C(n)} \frac{A(n-1)}{A} \quad \text{(by Equation 8)} \]

\[ = Y(n) S(n) p_A(n-1). \]
Collecting these results: whenever flow balance and one-step behavior are satisfied,

\begin{align}
(10) \quad p(n) &= Y(n-1) S(n) p(n-1) \quad n = 1, \ldots, N \\
(11) \quad p_A(n) &= Y(n) S(n) p_A(n-1) \quad n = 1, \ldots, N-1
\end{align}

Equation (10) can be used to compute the values of $p(n)$ from measurements or estimates of the $Y(n)$ and $S(n)$: start with a positive value of $p(0)$, say $p(0) = 1$, iteratively compute $p(1)$, $p(2)$, ..., $p(N)$, and then normalize by dividing each $p(n)$ by the sum $p(0) + p(1) + \ldots + p(N)$. The same algorithm can be used with Equation (11) to compute the values of $p_A(n)$.

Equation (10) produces the same formal distribution as the steady state balance equations for a "general birth-death queue" with state dependent Poisson arrivals and state dependent exponential service. Note, however, that no Markovian assumptions were required for the derivation of Equation (10). Equation (10) is valid for any stochastic behavior sequence that satisfies flow balance and one-step behavior. If the behavior sequence is generated by an $M/G/1$ stochastic process with a Coxian service time distribution, Marie's algorithm can be used to calculate the $Y(n)$ and $S(n)$ [MAR78, p45]. See also BUZE78 and LAZ078.

**Homogeneity**

To apply Equations (10) and (11) to an arbitrary system, it is necessary to measure or estimate the values of $Y(n)$ for $n = 0, 1, 2, \ldots, N-1$ and $S(n)$ for $n = 1, 2, \ldots, N$ (a total of $2N$ values).
In some cases, the number of independent variables can be reduced significantly by making one or both of the following assumptions:

\[(12) \quad Y(0) = Y(1) = \ldots = Y(N-1)\]

\[(13) \quad S(1) = S(2) = \ldots = S(N)\]

When combined with Equations (2) and (3), Equation (12) implies that each value of \(Y(n)\) is equal to \(Y\). Likewise, Equations (13) and (14) imply that each value of \(S(n)\) is equal to \(S\).

Equation (12) is called the assumption of homogeneous arrivals; it asserts that the arrival rate is independent of queue size \(n\). Equation (13) is called the assumption of homogeneous services; it asserts that the mean time between completions is independent of \(n\). These equations are examples of the general operational technique of simplifying problems by introducing homogeneity assumptions that allow a set of conditional rates to be replaced by a single, unconditional value.

Homogeneity assumptions reduce the number of independent variables and thereby simplify both the algebraic form and the intuitive implications of our equations. The simplified equations are of interest only when there is reason to believe they will acceptably characterize actual performance. In fact, they often do. Many real systems satisfy Equations (12) and (13) approximately.

Since the aggregate ("macro") values \(Y\) and \(S\) are usually easier to measure or estimate than the individual ("micro")
values of \( Y(n) \) and \( S(n) \), the homogeneity assumptions provide another benefit: they enable the analyst to use independent variables that can be measured or estimated with a higher degree of confidence. Even if Equations (12) or (13) are not satisfied precisely, it is often better to proceed as if they were, because approximate solutions based on variables that are easy to obtain are often more robust than exact solutions based on variables that are difficult to measure or estimate.

Because the assumptions of homogeneity cause the solutions of Equations (10) and (11) to have the same form as well-known stochastic results [BUZE76, BUZE78, DENN78], it is legitimate to inquire whether the operational concepts of flow balance and homogeneity are equivalent to their stochastic counterparts, steady-state and Markovian assumptions (i.e., Poisson arrivals and exponential service times). They are not. Flow balance is a measurable property of finite behavior sequences whereas steady-state is an abstraction for infinite behavior sequences; in fact, steady-state implies flow balance, but not conversely.

Figures 2 and 3 are behaviors of single-step, flow balanced systems with homogeneous arrivals and services. In both cases, the mean service time (\( m \)) is the same as the mean time between completions (\( S \)); the coefficient of variation of service time (\( \rho \)) differs significantly from 1, the value for an exponential distribution. These service times are not well modelled by exponential distributions. A similar statement holds for interarrival times.

We can extend either sequence by repeating the given pattern indefinitely so that the resulting behavior sequence will fail any
FIGURE 2. A homogeneous, flow balanced behavior sequence of a single server queue.
- 11(f2) -

\[ m = k \quad \sigma^2 = \frac{(4k^2 - 2k + 6)}{7} \]

<table>
<thead>
<tr>
<th>( n )</th>
<th>( A(n) )</th>
<th>( C(n) )</th>
<th>( T(n) )</th>
<th>( p(n) )</th>
<th>( p_A(n) )</th>
<th>( Y(n) )</th>
<th>( S(n) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4</td>
<td>-</td>
<td>8k</td>
<td>8/15</td>
<td>4/7</td>
<td>1/2k</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>4</td>
<td>4k</td>
<td>4/15</td>
<td>2/7</td>
<td>1/2k</td>
<td>k</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2k</td>
<td>2/15</td>
<td>1/7</td>
<td>1/7k</td>
<td>k</td>
</tr>
<tr>
<td>( N - 3 )</td>
<td>-</td>
<td>1</td>
<td>k</td>
<td>1/15</td>
<td>-</td>
<td>-</td>
<td>k</td>
</tr>
</tbody>
</table>

**TOTALS**

\[ \begin{align*}
A & = 7 \\
C & = 7 \\
T & = 15k \\
\end{align*} \]

\[ Y = \frac{A}{T-T(N)} = \frac{1}{2k} \]

\[ S = \frac{T-T(0)}{C} = k \]

**FIGURE 3.** Homogeneous, flow balanced behavior sequence of a single server queue.
statistical test for goodness-of-fit to exponential service times (or Poisson arrivals) at any given level of confidence. In other words, the extended replications of Figures 2 and 3 are non-Markovian behavior sequences that satisfy the operational conditions of homogeneity. Whereas a homogeneous behavior sequence need not be Markovian, a long Markovian behavior will be homogeneous.

Implications of Homogeneous Arrivals

Consider any system that satisfies the homogeneous arrivals assumption. Since this implies each value of \( Y(n) \) is equal to \( Y \), Equation (1) reduces to

\[
p_A(n) = p(n) \frac{Y}{Y_0} \quad \text{for } n = 0, 1, 2, \ldots, N-1. \tag{14}
\]

Thus, if arrivals are homogeneous, the arrivers distribution is almost the same as the overall distribution. The only difference is that the arrivers distribution ranges from 0 to \( N-1 \) whereas the overall distribution ranges from 0 to \( N \). The arrivers distribution must therefore be renormalized by multiplying each term by \( Y/Y_0 \). Equation (2) allows replacing the ratio \( (Y/Y_0) \) with \( 1/(1-p(N)) \):

\[
p_A(n) = \frac{p(n)}{1-p(N)} \quad \text{for } n = 0, 1, 2, \ldots, N-1. \tag{15}
\]

In the limit where \( p(N) \to 0 \) as \( N \to \infty \), Equation (15) shows that \( p_A \) and \( p \) are identical. Conversely, if Equation (15) holds, Equations (1) and (2) then imply that \( Y(n) = Y \). The conclusion
is that Equations (14) and (15) are necessary and sufficient for homogeneous arrivals (LAZ078 gives experimental corroboration).

Equations (14) and (15) are theorems that depend only on homogeneous arrivals. Flow balance and one-step behavior are not required. Figure 4 presents a flow imbalanced sequence with \( N = 5, p(N) = 0 \) and \( p_A = p \).

The theorem expressed by Equation (14) can be strengthened to yield bounds on the error introduced when homogeneity is an inexact assumption. The statement, "arrivals are within \( \epsilon \) of being homogeneous" is quantified by writing

\[
| \frac{Y - Y(n)}{Y(n)} | < \epsilon, \quad n = 0, 1, ..., N-1.
\]

Now, let \( p_A(n) \) denote the exact arrivals distribution computed from Equation (1), and let \( \hat{p}_A(n) \) denote the approximate distribution computed by assuming homogeneity. The relative error in estimating the arrivals distribution is

\[
\epsilon_A(n) = | \frac{\hat{p}_A(n) - p_A(n)}{p_A(n)} |.
\]

On using Equation (14) to reduce \( \hat{p}_A(n) \) to \( p(n)Y/Y \), and Equation (1) to reduce \( p_A(n) \) to \( p(n)Y(n)/Y \), this error term simplifies to

\[
(16) \quad \epsilon_A(n) = | \frac{Y - Y(n)}{Y(n)} | < \epsilon .
\]

In words, the relative error in the approximation to \( p_A(n) \)
\[ p(n) = \frac{1}{5} \] for \( n = 0, \ldots, 4 \)

**Figure 4.** A flow imbalanced behavior sequence for which \( p = p_A \).
is the same as the relative error in the homogeneous arrival assumption. If the conditional arrival rate is within $\varepsilon$ of being homogeneous, the homogeneous approximation to $p_A$ will be within $\varepsilon$ of its true value.

The result $p_A = p$ is familiar to those who have studied the unbounded stochastic $M/G/1$ queue in steady-state. Saaty [SAAT61, p186] outlines a lengthy proof given by Khintchine [KHINT60], but Kleinrock presents a shorter proof [KLEI75, p18]. In addition, Cooper presents a derivation of Equation (15) for $M/M/1/N$ queues [COOP72]. The operational analog of Khintchine's result (Equation 15) is simpler still; it is also more general because it does not require flow balance and one-step behavior. Moreover, it leads to an error analysis.

Operational error bounds such as Equation (16) are especially important for predictions [BU2E79]. Suppose that Equation (14) is used to predict the values of $P_A(n)$ from estimates of $p(n)$ for a future time period. Since this equation depends on homogeneous arrivals, the analyst is interested in the errors that can arise if the assumptions are not satisfied exactly.

In the operational case, Equation (16) quantifies the concept of approximate homogeneity; it can be used to bound the error in using Equation (14) when homogeneity is not guaranteed. In contrast, it is difficult to quantify the concept of an arrival process being "approximately Poisson", and there is no stochastic error bound comparable to Equation (16). The new forms of sensitivity analyses that become possible in an operational context are an important advantage.
of operational analysis over stochastic analysis.

Implications of Homogeneous Services

Consider any flow-balanced, one-step behavior sequence in which \( S(n) = S \). The overall mean queue length is

\[
Q = \sum_{n=1}^{N} n \, p(n)
\]

and the mean queue length seen by arriving customers is

\[
Q_A = \sum_{n=0}^{N-1} n \, p_A(n).
\]

Now, by Equation (1) \( p_A(n) = p(n)Y(n)/Y_0 \); since \( S(n) = S \), Equation (10) implies \( p(n)Y(n) = p(n+1)/S \); therefore, \( p_A(n) = p(n+1)/SY_0 \).

Since flow is balanced, \( SY_0 = SX = U \). Accordingly,

\[
Q_A = \sum_{n=0}^{N-1} \frac{n \, p(n+1)}{U}
\]

\[
= \sum_{n=0}^{N-1} \frac{n \, p(n+1)}{U} - \sum_{n=0}^{N-1} \frac{p(n+1)}{U}
\]

So that

\[
Q_A = \frac{Q}{U} - 1
\]

Since \( Q = RX \) by Little's Law, \( \frac{Q}{U} = \frac{RX}{U} = R/S \). Therefore the mean response time for homogeneous services is

\[
R = S(1+Q_A).
\]
Equation (18) shows that average response time $R$ is the same as the time that the set of $1+Q_A$ jobs (in the queue just after the arrival) would take to complete if each required exactly $S$ seconds of service. The intuitive result is valid even though the mean residual service of the job being served at an arrival instant is not necessarily equal to $S$.

Implications of Complete Homogeneity

Consider any one-step flow-balanced behavior in which both arrivals and services are homogeneous. By Equation (14), $p_A(n) = p(n)Y/Y_0$; substituting into the definition of $Q_A$,

$$Q_A = (Q - Np(N))Y/Y_0.$$  

By Equation (17), $Q_A = Q/U - 1$. On setting these two expressions for $Q_A$ equal and using Equation (2), one finds an expression for mean queue length,

$$Q = \frac{U}{1-U-p(N)} (1-(N+1)p(N)).$$  

From Little's Law, $R = Q/X = SQ/SX = SQ/U$, or

$$R = \frac{S}{1-U-p(N)} (1-(N+1)p(N)).$$  

Equations (19) and (20) are the operational counterparts of the mean-queue and response time formulae for the $M/M/1/N$ queue [COOPO72, KLEIT75]. If the queue is unbounded and $Np(N) > 0$ with increasing $N$, Equation (19) reduces to the form $Q = U/(1-U)$ and Equation (20) to the form $R = S/(1-U)$, as for the $M/M/1$

* In the special case that $p(0) = p(N)$, the denominator of this formula is 0. However, in this case, Equation (10) shows that $p(n) = p(n-1)$ because $SY = SY_0/(1-p(N)) = U/(1-p(N)) = (1-p(0))/(1-p(N)) = 1$. For this case $p(n) = 1/(N+1)$ and $Q = N/2$. 

- 16 -
queue. Although these equations have the same form as their stochastic counterparts, their interpretation is different -- e.g., they apply to non-Markovian behaviors such as those in Figures 2 and 3.

The Sevcik-Mitrani Theorem

The preceding discussion deals with single-queue systems. Many real systems comprise several interconnected queues. Models of closed queueing networks are the basis for almost all successful analyses of multiprogrammed computer systems.

Equations (1)-(20) apply to any queue, whether it is part of a network or not. When a queue belongs to a network, its arrival function is determined jointly by the output (service) functions of the other queues that feed it. For this reason it is helpful to derive additional results that relate queueing distributions to the service functions of all the queues in a closed queueing network.

A fundamental theorem for closed queueing networks is the Sevcik-Mitrani theorem [SEVC78]. It states that the arrivals distribution at any queue is the same as the overall distribution at that queue when the network has a load of one job less. In other words, the arriving customer sees the same queue distribution as the outside observer who studies the same network with the arriving customer removed. This theorem was first noted informally by Reiser and Lavenberg [REIS78, REIS79] and later proved by them [LAV79].

The following lemma will be used to prove the theorem. Consider
a flow-balanced, single-step behavior sequence. Because \( A(n) = C(n+1) \) and \( A = C \),

\[
p_A(n) = \frac{A(n)}{A} = \frac{C(n+1)}{C} = \frac{T}{C} \frac{C(n+1)}{T(n+1)} T(n+1),
\]

so that

\[
(21) \quad p_A(n) = \frac{p(n+1)}{X S(n+1)}.
\]

Equation (21) has a simple intuitive interpretation: \( p(n+1)/S(n+1) \) is the departure rate (per unit time from state \( n+1 \)), and \( X \) is the overall departure rate from all states. The ratio of these departure rates is the proportion of departures that occur from state \( n+1 \) or, equivalently, the proportion of departures that leave the queue in state \( n \). Under the conditions of flow balance and one step behavior, this is then the proportion of arrivals that find the system in state \( n \). Note that this lemma does not require the queue to be part of a network.

We now turn our attention to single class closed queueing networks with product form solutions [JACK63, GORD65, DENN78].

The state of a closed queueing network of \( K \) devices (queues) is a vector \( n = (n_1, \ldots, n_K) \) specifying the number of jobs present at each device. The state space for load \( N \), denoted \( S(N,K) \), is the set of all such vectors for which \( n_1 + \cdots + n_K = N \). A behavior sequence of such a network is a record of \( n(t) \) for all \( 0 \leq t \leq T \). Suppose that the given behavior sequence is flow balanced, one-step, and network homogeneous. Network homogeneity means that the
transfer rate between any pair of devices \((i,j)\) depends only on
the queue length at the source, \(n_1\) [Denn77,Denn78]. Network
homogeneity is not a Markovian assumption; Sevcik and Klawe [Sevc79]
have presented examples of network-homogeneous behavior sequences
that do not have Markovian servers. Markovian behavior sequences
are network-homogeneous in the limit.

For a flow-balanced, one-step, network-homogeneous behavior
sequence, the proportion of time each state is occupied follows
the product form solution [Buze73, Klei75, Denn78]:

\[
p(n) = \frac{1}{g(N,K)} \prod_{i=1}^{K} F_i(n_i).
\]

In this formula, \(F_i\) is the device factor

\[
F_i(n) = \begin{cases} 
1, & n = 0 \\
\lambda_i S_i(n) F_i(n-1), & n > 0 
\end{cases}
\]

where \(\lambda_i\) is the mean number of visits per job to device \(i\) and
\(S_i(n)\) is the mean-time-between-completions function for \(n_i(t) = n\).
The function \(g(N,K)\) is a normalizing constant:

\[
g(N,K) = \sum_{n \in S(N,K)} \prod_{i=1}^{K} F_i(n_i).
\]

The normalizing constant can be calculated in time \(O(N^2K)\) from
[Buze73]:

\[
g(n,k) = \sum_{j=1}^{k} F_k(j) g(n-j, k-1).
\]
For any \( n \leq N \) and \( k \leq K \), the "partial constants" \( g(n,k) \) are well-defined in the original behavior sequence. Note, however, that the partials \( g(n,k) \) may also be interpreted as the normalizing constant of a different (flow balanced, homogeneous) behavior sequence for a subsystem comprising only devices \( 1, \ldots, k \) and having load \( n \) jobs; the values of \( V_i \) and \( S_i(j) \) for \( 1 \leq i \leq k \) and \( 1 \leq j \leq n \) in this different sequence are the same as those in the original. By interpreting these partials as normalizing constants of different (but related) behavior sequences, we can develop equations relating measurable quantities of one sequence to those of another.

Let \( p(n_i = n_i|N) \) denote the proportion of time during which \( n_i \) customers are observed at device \( i \), given that \( N \) jobs are in the entire network. Similarly, let \( p_A(n_i = n_i|N) \) denote the proportion of arrivals to device \( i \) that occur when its queue length is \( n_i \). These definitions are identical to those given earlier for \( p(n) \) and \( p_A(n) \), except that the index of the device \( i \) and the total network load \( N \) are explicit.

The overall distribution of queue length at device \( i \) is

\[
p(n_i = n_i|N) = \sum_{n \in S(N,K)} p(n) \cdot \sum_{n_i = n} p\Bigg(n_i = n_i|n_i = n\Bigg).
\]

We can use Equation (21) to obtain \( p_A(n_i = n_i|N) \) once \( p(n_i = n_i|N) \) is known.

To simplify the notation (without losing generality) assume \( i = K \). Substituting the product form solution into the definition
of \( p(n_k=n|N) \), we obtain

\[
p(n_k=n+1|N) = \sum_{n \in S(N,K)} \frac{1}{g(N,K)} \prod_{i=1}^{K} F_i(n_i)
\]

\[
= \frac{F_{K}(n+1)}{g(N,K)} \sum_{n \in S(N-n-1,K-1)} \prod_{i=1}^{K-1} \ F_i(n_i)
\]

so that

\[
(22) \quad p(n_k=n+1|N) = F_{K}(n+1) \frac{g(N-n-1,K-1)}{g(N,K)}.
\]

The definition of device factor allows us to replace \( F_{K}(n+1) \) with \( V_K S_{K}(n+1)F_{K}(n) \). With this replacement and on regrouping terms, we obtain

\[
p(n_k=n+1|N) = \left[ V_K \frac{g(N-1,K)}{g(N,K)} \right] S_{K}(n+1) \left[ F_{K}(n) \frac{g(N-n-1,K-1)}{g(N-1,K)} \right]
\]

The first bracketed term is the definition of the throughput at device \( K \) for network load \( N \), namely \( x_K(N) \). (See DENN78.) The second bracketed term is the definition of \( p(n_k=n|N-1) \); see Equation (22). Therefore,

\[
(23) \quad p(n_k=n+1|N) = x_K(N) S_{K}(n+1) \ p(n_k=n|N-1).
\]

If we divide both sides of Equation (23) by \( x_K(N) S_{K}(n+1) \), the left side will reduce to \( p_A(n_k=n|N) \) according to Equation (21). Therefore,

\[
(24) \quad p_A(n_k=n|N) = p(n_k=n|N-1).
\]
Equation (24) is the operational form of the Sevcik-Mitrani theorem for a single class network.

In Equations (23) and (24), the distributions on the left side are defined for a behavior sequence under load \( N \); the distributions on the right side are defined for a different behavior sequence of the same network under load \( N-1 \). These sequences match in the sense that all the values of \( V \) and \( S_i(n) \) are the same. These equations hold for networks having the product form solution.

Calculating a Queue Distribution

An interesting corollary of the Sevcik-Mitrani theorem is that a knowledge of the system throughput, \( X_0(N) \), and the device parameters suffices to calculate any queue distribution \( p(n_i=n|N) \). By the forced flow law [DENN78], \( X_1(N) = V_1 X_0(N) \). With Equation (23), this implies

\[
p(n_i=n|N) = \begin{cases} 
X_1(N) S_i(n) p_i(n-1|N-1), & \text{for } n = 1, \ldots, N \text{ and } N \geq 1; \\
1, & \text{for } n = 0 \text{ and } N = 0 
\end{cases}
\] (25)

This recursion is also deducible from results in BRUE78 and REIS79.

Equation (25) is a simple procedure for calculating \( p(n_i=n|N) \). Begin with \( p(n_i=0|0) = 1 \). Having calculated \( p(n_i=n-1|N-1) \) for \( n = 0, \ldots, N-1 \), apply Equation (25) to calculate \( p(n_i=n|N) \) for \( n = 1, \ldots, N \); then choose \( p(n_i=0|N) \) to normalize. This procedure requires time \( O(N^2) \). Figure 5 illustrates the
The important aspect of Equation (25) is that it is independent of the method used to calculate the system's throughput, $X_0(N)$. In other words, $X_0(N)$ can be calculated as the ratio of the normalizing constants $g(N-1,K)/g(N,K)$ [DENG73, DEMN78]; or it can be calculated by mean-value analysis, as shown in the next section [BAR079, REIS78, REIS79]. Flow balance and homogeneity are such powerful properties that the system's throughput and parameters implicitly contain the information needed to retrieve the queueing distributions.

Mean Value Analysis

Mean value analysis is a new technique for computing mean response times, throughputs, and queue lengths at devices in closed queueing networks [REIS78, REIS79]. We will present the simplest form of the method -- for closed, product form networks with a single job class and homogeneous (queue independent) service times.

Mean value analysis uses the Sevcik-Mitrani theorem to calculate mean values for successively larger network loads $N$. The load $(N)$ and device indices $(i)$ will be shown explicitly. There are three basic equations. If the behavior sequence of a closed queueing network satisfies the conditions of flow balance, single steps, and homogeneous service, Equation (18) implies

$$R_i(N) = S_i (1 + Q_{A_i}(N))$$
FIGURE 5. Iteration step in calculating a queue length distribution.
at device $i$. Since the Sevcik-Mitrani theorem implies that

$$Q_{A_i}(N) = Q_i(N-1)$$

for any closed product form network,

$$R_{i}(N) = S_{i}(1 + Q_i(N-1)) \quad i = 1, \ldots, K$$

This is the first basic equation of mean value analysis.

The other two equations depend only on flow balance and Little's Law. Since $R_{i}(N)$ is the mean response time per visit at device $i$, and since $V_{i}$ is the mean number of visits per job to this device, $V_{i}R_{i}(N)$ is the mean total time a job spends at device $i$. Then the average system holding time (response time) per job at load $N$ is

$$R_{0}(N) = \frac{X}{\sum_{i=1}^{K} V_{i}R_{i}(N)}.$$

Since $X_{0}(N)$ is the network throughput, Little's Law applied to the entire network implies $N = R_{0}(N)X_{0}(N)$, or

$$X_{0}(N) = \frac{N}{\sum_{i=1}^{K} V_{i}R_{i}(N)}.$$

This is the second basic equation of mean value analysis.

The forced flow law states that $X_{i}(N)$, the throughput at device $i$, is $V_{i}X_{0}(N)$ [DENS78]. Little's Law then implies

$$Q_{i}(N) = R_{i}(N) V_{i} X_{0}(N) \quad i = 1, \ldots, K$$

This is the third basic equation of mean value analysis.
Equations (26)-(28) can be applied iteratively to compute $R_i(N), X_0(N)$, and $Q_i(N)$ for any value of $N$, once the values of $V_i$ and $S_i$ are given. The iteration begins with $N = 1$ and the boundary condition $Q_1(0) = 0$. An example of the mean value iterations with $K = 3$ and

\[
V_1 = 1, \quad V_2 = 2, \quad V_3 = 3
\]

\[
S_1 = 2, \quad S_2 = 1, \quad S_3 = 1
\]

is given in Table 1.

<table>
<thead>
<tr>
<th>$N$</th>
<th>$R_1(N)$</th>
<th>$R_2(N)$</th>
<th>$R_3(N)$</th>
<th>$X_0(N)$</th>
<th>$Q_1(N)$</th>
<th>$Q_2(N)$</th>
<th>$Q_3(N)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.000</td>
<td>1.000</td>
<td>1.000</td>
<td>.143</td>
<td>.286</td>
<td>.286</td>
<td>.429</td>
</tr>
<tr>
<td>2</td>
<td>2.571</td>
<td>1.286</td>
<td>1.429</td>
<td>.212</td>
<td>.545</td>
<td>.545</td>
<td>.909</td>
</tr>
<tr>
<td>3</td>
<td>3.091</td>
<td>1.545</td>
<td>1.909</td>
<td>.252</td>
<td>.779</td>
<td>.779</td>
<td>1.443</td>
</tr>
<tr>
<td>4</td>
<td>3.557</td>
<td>1.779</td>
<td>2.443</td>
<td>.277</td>
<td>.985</td>
<td>.985</td>
<td>2.050</td>
</tr>
</tbody>
</table>

Table 1. Mean Value Iterations

The values of $R_i(N), X_0(N)$, and $Q_i(N)$ in Table 1 could also be computed by the algorithm based on normalizing constants [HUIZU73]. When $R_i(N)$ and $Q_i(N)$ must be computed for every server in the network, the normalizing-constant method and mean-value method require approximately the same number of arithmetic operations and the same amount of storage. There are many applications, however, where $X_0(N)$ is the only
quantity the analyst seeks. In these applications, the
normalizing-constant algorithm requires approximately half as
many arithmetic operations as mean-value analysis.

Mean value analysis can be extended in various ways. Each
is based on altering the method of computing response time in
Equation (26). The simplest extension is the "queueless server",
where device \( i \) comprises at least as many parallel processors as
there are jobs in the network; each processor has mean service
time \( S_i \). In this case, Equation (26) is replaced with

\[
R_i(N) = S_i 
\]  

for all \( N \).

A more complex extension is needed for a general, load-
dependent server. The response time per visit to device \( i \) is

\[
R_i(N) = Q_i(N)/X_i(N) = \sum_{n=1}^{N} p(n_i = n|N)/X_i(N). 
\]

The recursion of Equation (25) reduces this to

\[
R_i(N) = \sum_{n=1}^{N} n S_i(n) p(n_i = n|N-1). 
\]

For the general server, we use Equation (30) instead of (26).
We must, however, also replace the computation of mean queue
distribution at Equation (28) with a computation of the entire queueing
distribution according to Equation (25).

Only the simplest equations applying to a device need be
used at each iterative stage. If device \( i \) has homogeneous
service times. Equations (26), (27), and (28) define the iteration from network load $N-1$ to $N$ at that device. If device $i$ is a queueless server, Equations (29), (27), and (28) define the iteration. If device $i$ is a general server, Equations (30), (27), and (25) define the iteration. Since a homogeneous, flow-balanced network with any of these kinds of devices satisfies the product form solution, the normalizing-constant algorithm [BRUE78, BUZE73] would yield the same results.

Mean value analysis can also be extended to approximate solutions for closed networks that do not have product form solutions. For example, Bard's heuristic extensions to Equation (26) can be used to approximate priority scheduling in multi-class networks [BARD79]. These extensions may ultimately prove to be the most significant practical contribution of mean value analysis.

A primary advantage of mean value analysis (relative to normalizing-constant analysis) is its improved numerical stability. This is because the normalizing-constants can become very large especially if the ratio of the largest to smallest $V_{S_1}$ for all $i$ and $n$ is very large [BRUE78].

Because mean value analysis is new, it is too early to present a comprehensive discussion of its strengths and weaknesses relative to normalizing-constant methods. Yet, its intuitive appeal and the overall simplicity of its iterations make mean value analysis an attractive subject.
Conclusions

The derivations in this paper illustrate the power of operational analysis. Although most results were already known as stochastic theorems, the operational proofs are significantly simpler. These proofs also demonstrate that the theorems are valid in many practical cases where stochastic assumptions cannot be justified. Operational analysis both extends and simplifies stochastic modeling.

This paper has also initiated the study of operational bounds on the errors that can arise if homogeneity is only satisfied approximately. Such bounds are important for predictions applications because the future validity of homogeneity assumptions is never certain. Since it is difficult to quantify the concept of "approximate validity" for stochastic assumptions such as ergodicity or Poisson arrivals, operational analysis has a distinct advantage over stochastic modeling in this regard.

The discussion of the Sevcik-Mitrani theorem, mean value analysis, and the queue length distribution algorithm illustrates that important new results continue to be discovered in the theory of queueing networks. Because of its simplicity and intuitive appeal, the development of heuristic extensions to mean value analysis is a promising area for further investigation.
Acknowledgements

We are grateful to Scott Graham, Len Kleinrock, Andy Langer, Lil Lazuowska, Ken Scvck, and Johnny Wong for helping fix some of these ideas during their formation. The National Science Foundation supported part of this work through its grant MCS78-01729 at Purdue University.
Bibliography


