Operational Analysis Revisited: Error Measure Limits of Assumptions

Neal Bengtson

Abstract[- **The assumptions used to develop operational analysis computer performance measures, such as number of jobs at a device or response times, are stated in terms of the data itself, rather than the underlying system which produces the data. In spite of claims of validity and as an aid in introducing queueing theory in teaching, little has been written about operational analysis in the past ten years. Accuracy of operational analysis performance measures depend on data behavior assumptions which can be validated with data based error measures. Increased soundness of the operational analysis approach may be obtained by determining the limits of assumption errors as the time period of observation increases. Part I of this paper is a review of operational analysis and addresses some of the previous concerns with its approach. Part II develops further understanding of operational analysis assumption errors by examining their limits. Limits are found for the assumption errors of job flow balance, homogeneous arrivals and homogenous services. While the job flow balance assumption error measure is shown to approach zero over time, the homogeneity assumption error measures, in general, do not.**

Index Terms— **assumptions, error measures, limiting values, operational analysis**

A. OPERATIONAL ANALYSIS REVISITED

I. INTRODUCTION OF OPERATIONAL ANALYSIS

Operation analysis (OA) was introduced as an aid in computer system performance analysis [1-4]. OA has been shown to produce relations which are closely related to those in traditional queueing theory and is a complementary approach that has been used in many networks of servers performance analyzes and in computer programs [5-11]. In OA system, outputs are observed and basic performance measures are obtained under two rules:

• All assumptions made in the analysis of real system performance should be subject to direct verification,

• All variables appearing in the performance equations of a real system should be calculable by direct measurement.

These rules mean that no assumptions are made that can't be tested for their validity over a given time period of observation.

In traditional stochastic models for a network of queues, the system is assumed to move through its states in a Markov process. Successive service times are independent, successive transitions between servers are independent, service distributions are exponential, and the system reaches a steady

Neal Bengtson is with the School of Business, Barton College, Wilson, NC (e-mail: Bengtson@barton.edu).

state. Tests since the 1960s have shown a good fit between these models and the throughput and response times of real systems.

Queueing network models can be unattractive because of the high computation overhead needed to calculate performance quantities. Computation has been simplified starting with the equivalence property: "Assume that a service facility with s servers and an infinite queue has a Poisson input with parameter λ and the same exponential service-time distribution with parameter μ for each server (the M/M/S model), where $s\mu > \lambda$. Then the steady-state output of this service facility is also a Poisson process with parameter λ" [12]. Jackson Networks can use product form solutions, $P\{(N_1, N_2, \ldots, N_m) = (n_1, n_2, \ldots, n_m)\} = P_{n1} \cdot P_{n2} \cdot \ldots \cdot P_{nm}$. These have been extended by the BCMP theorem [13]. The product form is simpler to calculate then solving the balance equation, but evaluations are still difficult for all but the smallest systems. Further developments have improved system performance evaluation and capacity planning.

"While performance analysts repeatedly found users interested in their queueing models, they constantly faced skepticism because no one trusted the models' assumptions" [12]. The paradox is that real world computing systems consistently violated all the model assumptions, but the models agreed closely with observed throughput and response times. This had an impact on Jeffery Buzen's company, BGS Systems, which built computer industry performance prediction and capacity planning tools. He found that business executives were reluctant to buy products whose modeling assumptions are violated, even when the products were supported by empirical results. As a consequence, Jeffery Buzen thought of another approach. He developed "Operational Analysis" in 1976 [1, 14], which relies on fundamental laws about utilization, throughput, and response time which are always true because of the way they were defined for collected data. "Analysts substitute measured, operational values of parameters for the model's stochastic parameters" [12]. Buzen and Denning built a queueing theory that starts from the operational laws and avoids making any Markovian or equilibrium assumptions.

For a system of queues, $\boldsymbol{n} = (n_1, \ldots, n_K)$ are vectors of the number of jobs at each server. In stochastic modeling they assign an equilibrium probability $p(\boldsymbol{n})$ to each state. In OA, $p(\boldsymbol{n})$ is interpreted as the proportion of time that the system spends in state $\boldsymbol{n}$. OA re-formulates the balance equations among the equilibrium $p(\boldsymbol{n})$ into balances of state transitions,

where entries equal exits, which are observable and can be measured [12].

Three assumptions, job-flow balance (entries equal exits), one-step behavior (each state change is caused by exactly one job completion), and homogeneity (the rate of transitions between two states is identical to the rate of job-flow between the servers causing the transitions), allowed the conclusion that the state occupancies obey the same product form structure as had been found by Jackson, Gordon, and Newell with Markovian assumptions. The operational characteristics of the three assumptions allow us to calculate errors in these assumptions over finite observation periods [15].

II.   REACTION TO OPERATIONAL ANALYSIS

There was a decided mixed reaction to OA, attracting both praise and criticism. Summarized in 1979 by Ken Sevcik [12, 16]

Praise-

"OA makes Stochastic Modeling (SM) obsolete."

"Stochastic Modeling is a security blanket used to smother intuition by those who lack it."

Criticism-

"OA offers nothing but tautologies."

"OA is a smokescreen for trivially deriving the obvious from the known."

The main criticism was of the homogeneity assumption. However, Sevcik was able to give several examples of deterministic systems that are flow-balanced, one-step, and homogeneous, but obviously not Markovian [16]. These examples satisfied product form solutions. According to Denning that was a turning point in the acceptance of OA.

OA invokes a different level of abstraction from SM: the two systems have the same symbols but interpret them differently. SM refers to probabilistic ensembles of system behaviors; OA refers to one behavior at a time. OA is more obviously relevant to real systems than SM. OA generate confidence in applying models by offering assumptions that are understandable and testable. OA and SM are complementary approaches. OA offers much that is new; SM isn't obsolete ([16] quoted in [12]).

A section of Peter Denning's Wikipedia page states:

In the middle 1970s he [Denning] collaborated with Jeffery Buzen on operational analysis, extending Buzen's basic operational laws to deal with all queueing networks. The operational framework explained why computer performance models work so well, even though violating the traditional stochastic Markovian assumptions. It has become the preferred method for teaching performance prediction in computing courses.

In 1981 Denning was present at a debate between Jeffery Buzen and his critics. Denning wrote an allegorical piece about the debate and concluded that neither side could gather any argument that would change minds and that the two sides may never come to an accord. Denning stated that "the OA and SM believers did have one major point of agreement: most everyone found it much easier to teach queueing networks to beginning students when starting with the operational interpretation" [12]. This is the experience of Denning, Sevcik, and Buzen when teaching queueing theory.

Operational analysis gives understanding when teaching about queueing networks. It is the approach taken in a popular book by Lazowski, Zahojan, Graham, and Sevcik, *Quantitative System Performance* [17].

Including the 2006 paper by Denning, there have been only a few papers since 1983 with OA developments. The 1986 25th *IEEE Conference on Decision and Control* included "Operational analysis of multiclass queueing networks," which extends the previous work on single class networks to multiclass networks [18]. With "Measuring Errors in Operational Analysis Assumptions" [15] in *IEEE Transactions on Software Engineering*, assumption error terms are defined and correction terms developed so that performance measures can be adjusted to give exact values for a set of data no matter how much the assumptions used in deriving the performance measure relations are violated. The use of these correction terms is illustrated in "Using Operational Analysis in Simulation: A Queueing Network Example," in the *Journal of the Operational Research Society* [19]. In 1992, Yves Dallery and Cao published "Operational analysis of stochastic closed queueing networks" in *Performance Evaluation* which extended OA in order to analyze the asymptotic behavior of stochastic closed queueing networks. Also shown in that paper, is "that the aggregate network satisfies the operational assumptions yielding the product-form solution" [10]. In addition to queueing networks, suggested applications for the OA approach include telecommunications [20], E-commerce [19, 20], flexible manufacturing systems [23], and Petri nets [24-26].

One question arises: If Operational Analysis is an aid to teaching queueing theory, has been shown to be a valid approach to evaluating networks, and has seen application in several areas of study, then why has there been little acceptance, additional applications, and almost no development in over ten years? One reason may be the uncertainty over how large the errors in OA assumptions may be. That is, once performance measures are defined using OA principles, how reliable are the results given that the assumptions under which the measures are derived are approximations. For any particular period of time the errors can be measured, but different periods of observed time data will probably lead to different errors quantities.

This study examines the OA assumption error measures previously defined [15] over extended time horizons in order to see if assumption errors decrease or stabilize, making the OA performance measure relationships more applicable.

B. UNDERSTANDING ASSUMPTION ERROR LIMITS

III.   EXPLANATION OF OPERATIONAL ANALYSIS ERROR MEASURES

The types of systems considered in this paper are networks of devices. These devices are normally combinations of a single, FIFO queue with infinite capacity and a single server. The number of jobs (customers, items, etc.) both in queue and in service at a device defines the state of the device. There is only one class of job. Performance measures (PMs) for these

devices may be determined by keeping track of the accumulated operational values of the number of arrivals at each device, the number of completions, and the time spent at each state. These quantities are denoted $A_i(n)$, $C_i(n)$, and $T_i(n)$ respectively, where i indicates the device and n the state of the device. These quantities summed over all visited states, $n=m_i,...,N_i$, are denoted A_i, C_i, and T_i respectively [19]. m_i is the minimum visited state and N_i is the maximum visited state.

PMs may be determined directly or, often, indirectly. For example, the average number of jobs at device i for a period $[0,T]$ is directly calculated by

$$Q_i = \sum_{n=1}^{N_i} \frac{nT_i(n)}{T} \qquad (1)$$

or indirectly by

$$Q_i = \bar{n}_i^{AS} = \frac{U_i}{1 - U_i - p_i(N_i)}(1 + (N_i + 1)p_i(N_i)) \qquad (2)$$

where: N_i is the maximum number of jobs seen at the device at any time,

U_i is utilization, $(1-T_i(0))/T$, and

$p_i(N_i)$ is the proportion of time spent at the maximum state, $T_i(N_i)/T$.

Equation (2) is an Operational Analysis (OA) relationship derived under certain assumptions. These assumptions are indicated by the superscript AS in the equation and are:

One step behavior- only one job may arrive or leave at a time,

Job flow balance- the number of arrivals and number of completions in period $[0,T]$ for each state are the same,

Arrival homogeneity- $A_i(n)$ are equal for $n=m_i,...,N_i-1$, and

Service homogeneity- $S_i(n)$, the average time between completions, are equal for $n=m_i+1,...,N_i$.

If any of these assumptions are violated, and they usually

will be over the finite time period of observation, then (2) gives an approximate value for Q_i. Because of the method of calculating PMs, the devices in the network may be thought of as "black boxes." That is, they are not necessarily single queue, single server combinations, but any device or combination of devices such that the data associated with these devices behaves according to the assumptions given. Notice that these OA assumptions say nothing about underlying system characteristics. They are characteristics of data only. The system which produces the data may be stochastic, but, except as noted below, no assumptions are made about the characteristics of the underlying system.

The one step behavior assumption can be known to hold by the structure of a model, particularly for simulation models. The degree to which the other three assumptions hold can be determined by error measures. These measures are used to find correction terms which, when applied to quantities derived with relations developed under the OA assumptions, will give exact PMs for the period of interest [15].

One problem with the correction terms is that to determine them exactly requires an amount of information which is equal to the amount needed to determine the PMs directly. However, the correction terms may still be useful. If a number of PMs are to be found, it may be more efficient to calculate the error measures for the limited number of assumptions needed (usually two or three) than to calculate each PM directly. This is because the OA relationships are often simple equations which involve knowledge of just a few terms and are, therefore, attractive to use. In some situations, it may not be possible to directly measure a performance characteristic without affecting the system behavior. Examples of OA relationships are given in Table I (based on [27]).

Table I. Operational Analysis Quantities and Relationships	
Quantities	**Relationships**
$p(n) = T(n)/T$ = state n time proportion	
$Y(n) = A(n)/T(n)$ = state n arrival rate	
$Y_0 = A/T$ = overall arrival rate	$Y_0 = \sum_{n=0}^{N-1} p(n)Y(n)$
$p_A(n) = A(n)/A$ = state n arrival proportion	$p_A(n) = p(n)[Y(n)]$
$Y = A/(T - T(N))$ = restricted arrival rate	$Y = Y_0/(1 - p(N))$
$S(n) = T(n)/C(n)$ = state n average inter-completion interval	$S(n) = \dfrac{p(n)/p(n-1)}{Y(n-1)}$ (assuming job flow balance)
$U = 1 - p(0)$ =utilization	
$Q = \dfrac{1}{T}\sum_{n=1}^{N} nT(n)$ = average queue length	$Q = \dfrac{U}{1-U-p(N)}\left[1-(N+1)p(N)\right]$ (assuming arrival and service homogeneity)
$Q_A = \left(\sum_{n=1}^{N-1} nT(n)\right)/(T - T(N))$ = average queue length seen by an arriving job	$Q_A = \dfrac{Q - Np(N)}{1 - P(N)}$
$S = (T - T(0))/\sum_{n=1}^{N} C(n)$ =average inter-completion time	
$R = \left(\sum_{n=1}^{N} nT(n)\right)/\left(\sum_{n=1}^{N} C(n)\right)$ =average response time	$R = S(Q_A + 1)$ (assuming service homogeneity)
$X = \sum_{n=1}^{N} C(n)/T$ =average completion time	$X = \dfrac{Q}{R} = \dfrac{U}{S}$

Correction terms may be estimated by various methods. PMs calculated by OA equations are point estimates determined for a specific period of time. During another time period PM values may be quite different, as would be the assumption error measures and subsequent correction terms. Even though correction terms may give an exact value for a PM, they apply only to data of a particular period of time and are, therefore, estimates of the mean correction term value needed for data which are generated by underlying system behavior characteristics.

Since the correction terms are functions of error measures, a logical question is: what are the expected values of the error measures if the observation period is extended indefinitely? The answer to this question will give us an idea of the importance of the assumptions on which the OA relations are based.

For the derivations which follow it is necessary to distinguish between two types of networks of devices: open and closed. The open network is one to which jobs arrive from outside the system modeled by the network, traverse through the network and eventually leave. In this type of system, there

are not a fixed number of jobs in the network. In a closed network, there are a fixed number of jobs that circulate through the system, never increasing or decreasing the number of jobs that the network is handling.

In order to perform some of the algebraic manipulations which follow, it is necessary to make some assumptions about the behavior of the devices for which the OA assumption error measures are calculated. These assumptions fall into two categories: those that are used by OA to derive the error measures and those that are used to study these error measures over time. We will study behavior at a single representative device in the network; so, for simplicity, the device subscript i used above may be dropped.

OA Assumptions:
• The states of a device are $n=0, 1, 2... N$ where $n =$ the number of jobs at the device at time t, $n(t)=n$, and N is the maximum number of jobs ever seen at the device.
• The total number of arrivals, A, and the total number of completions, C, are both > 0.
• One-step behavior holds.

Time Assumptions:
• The limits exist for all functions used.
• The process is ergodic.

If the network is closed, N is known and fixed. If the network is open, N is not fixed and may increase over time.

All of the error measures can be computed as functions of time, T. Thus, for example, the flow balance assumption error measure of device i for a particular state (defined in the next section) can be denoted as

$$e_{F,i}(n,T).$$

For simplicity, and to avoid extra notation, the T will not be used in the error measures that follow, but it is understood that all error measures are functions of time.

The error measures associated with the three assumptions: job flow balance, arrival homogeneity, and service homogeneity, will be examined in order below. In each case, two error measures are defined, one which is state dependent and one which can be thought of as an overall measure [15]. The overall measure of a particular assumption is one which includes every state error measure of the device. The error measures each have the characteristic that when the assumption holds, its value is 0, and when the assumption does not hold, its value is a measure of the degree to which the assumption is violated. The only exceptions to this are the overall measures for arrival and service homogeneity which, in certain cases, will be 0 even when the individual state values for error are not. In that case, these measures can be thought of as weak error measures. They are used because they have useful properties [19].

Expectations of error measures and other quantities will be taken for specific time periods of observation. These expectations are meaningful in the operational analysis context considered here since the system which produces the data is stable in distribution

IV. FLOW BALANCE ASSUMPTION

We can define state dependent error associated with the flow balance assumption as

$$e_F(n) = (n + 1)(p_A(n) - p_C(n))$$

where: $p_A(n)$ = A(n)/A
= The proportion of arrivers that find n jobs at the device,

$p_C(n)$ = C$(n+1)$/C
= the proportion of completers that leave n jobs at the device,

A = A(0)+A(1)+...+A(N-1),

C = C(1)+C(2)+...+C(N).

The overall flow balance error is

$$e_F = \sum_{n=0}^{N-1} e_F(n).$$

Proposition 1: For each device in an open or closed network
$$\lim_{T \to \infty} E[e_F(n)] = 0.$$

Proof: See Appendix.

The proof of Proposition 1 implies that
$$\lim_{T \to \infty} p_A(n) = \lim_{T \to \infty} p_C(n) \quad n = 0,1,\cdots,N-1$$

by the definition of job flow balance. Given Proposition 1 it is easy to show

Proposition 2: For a device in an open or closed network
$$\lim_{T \to \infty} E[e_F] = 0.$$

Proof: See Appendix.

Using similar procedures we can show that the corresponding variances also go to 0.

Proposition 3: For a device in an open or closed network
$$\lim_{T \to \infty} V[e_F(n)] = 0.$$

Proof: See Appendix.

Proposition 4: For a device in an open or closed network
$$\lim_{T \to \infty} V[e_F] = 0.$$

Proof: See Appendix.

Propositions 1 through 4 mean that for time series of sufficient length, any error in the values of measures of performance calculated using relationships that rely on the assumption of job flow balance are going to be small. This result is not surprising and confirms our intuition that for a system in steady state there cannot be a consistent number of arrivals over completions or completions over arrivals.

Most OA relations rely on one or both of the homogeneity assumptions along with the job flow balance assumption. We need to know if the behavior of these assumption error measures also approaches 0.

V. ARRIVAL HOMOGENEITY

If arrivals are homogeneous, then the number of arrivals for a period of observations will be the same at each state of the device, i.e.,

$$A(m)=A(m+1)=\ldots=A(N-1) ,$$

Where m is the minimum state observed during the period of observation.

For each state we can define an error measure to be

$$e_A(n) = p_A(n)\left(\frac{T - T(n)}{T(n)}\right) - 1. \quad (3)$$

The weak overall error measure is

$$e_A^* = \sum_{n=1}^{N-1} e_A(n)\frac{nT(n)}{T - T(N)} \quad (4)$$

.

The * is to distinguish this weak error measure from a stronger one which is equal to 0 only when the assumption holds for each state n. It has been shown that the stronger assumption is unnecessary [15].

Taking the limit as $T\to\infty$ of the expected value of (3), using the definition $p_A(n)=A(n)/A$, then, by the bounded convergence theorem [28], taking the expectations outside the limits, and manipulating yields

$$\lim_{T\to\infty} E[e_A(n)]$$
$$= E\left[\lim_{T\to\infty}\left(\frac{A(n)}{T(n)}\right)\lim_{T\to\infty}\left(\frac{T}{A}\right)\right] - E\left[\lim_{T\to\infty}\left(\frac{A(n)}{T(n)}\right)\lim_{T\to\infty}\left(\frac{T(N)}{A}\right)\right] - 1.$$

Now, since the process is ergodic

$$\lim_{T\to\infty}\left(\frac{A(n)}{T(n)}\right) = \lambda(n)\ (w.p.1),$$

which is the arrival rate of state n. Also,

$$\lim_{T\to\infty}\left(\frac{T}{A}\right) = \frac{1}{\lambda_0}$$

where λ_0 is the unrestricted overall arrival rate. Then,

$$\lim_{T\to\infty} E[e_A(n)] = \frac{\lambda(n)}{\lambda_0} - \lambda(n)E\left[\lim_{T\to\infty}\left(\frac{T(N)}{A}\right)\right] - 1\ (w.p.1). \quad (5)$$

Let

$$\lim_{T\to\infty} E\left(\frac{T(N)}{A}\right) = \frac{1}{\lambda_N}.$$

Now we can state

Proposition 5: For a device in a closed network

$$\lim_{T\to\infty} E[e_A(n)] = \lambda(n)\left(\frac{1}{\lambda_0} - \frac{1}{\lambda_N}\right) - 1\ (w.p.1). \quad \square$$

In an open network the results simplify. Since the time between occurrences of a new N becomes greater as $T\to\infty$, then the times when $T(N)$ takes on new values are less frequent. In general, since A is increasing at a steady rate to ∞ as $T\to\infty$ then

$$\lim_{T\to\infty}\frac{T(N)}{A} = 0.$$

This result is proved in the Lemma which follows. Reintroducing time, t, in notation for clarity in the proof, $t(n(T))$ will be the time spent in state n during a period of observation T. Also, $N(T)=max\{n(t): 0\leq t\leq T\}$, i.e. $N(T)$ is the maximum state during the time period up to T. $A(T)$ is the number of arrivals to the device during the time period T and λ is the average arrival rate.

Lemma: Given a device in a stable system

$$\lim_{T\to\infty}\left(\frac{t(N(T))}{A(T)}\right) = 0$$

.

Proof: Since $A(T)/T\to\lambda$ it is sufficient to show that

$$\lim_{T\to\infty}\left(\frac{t(N(T))}{T}\right) = 0$$

.

Let $p(j)$ be the proportion of the observation time period spent in state j. Then,

$$\sum_{j=0}^{\infty} p(j) = 1,$$

which means that

$$\lim_{i\to\infty}\sum_{j=i}^{\infty} p(j) = 0.$$

This implies that $\forall \varepsilon > 0\ \exists i$ such that

$$\lim_{T\to\infty}\sum_{j=i}^{\infty}\frac{t(j(T))}{T} = \sum_{j=i}^{\infty}\lim_{T\to\infty}\frac{t(j(T))}{T} = \sum_{j=i}^{\infty} p(j(T)) \leq \varepsilon$$

.

Let $\lim_{T\to\infty} N(T) = \infty$, since this is the open system case. This means that $\forall i\ \exists t(i)$ such that $T > t(i)$ implies $N(T) > i$. Fix i, let $T > t(i)$, then $N(T) > i$. $t(N(T))$ is the time spent when the maximum number at the device occurred. Now, $t(N(T)) \leq \sum_{j=i}^{\infty} T(j)$. Dividing by T, and taking limits yields

$$\lim_{T\to\infty}\frac{t(N(T))}{T} = \lim_{T\to\infty}\sum_{j=i}^{\infty}\frac{T(j)}{T} \quad \forall i$$

,

In particular, this is true for the relation above. Then

$$\lim_{T\to\infty}\frac{t(N(T))}{T} \leq \varepsilon. \quad \square$$

Using this result in Equation (5) we have

Proposition 6: For a device in an open network

$$\lim_{T\to\infty} E[e_A(n)] = \frac{\lambda(n)}{\lambda_0} - 1\ (w.p.1). \quad \square$$

Substituting (3) for $e_A(n)$ into (4) gives

$$e_A^* = \sum_{n=1}^{N-1} n\left(p_A(n) - \frac{T(n)}{T - T(N)}\right). \quad (6)$$

Taking the limit of the expected value

$$\lim_{T\to\infty} E[e_A^*] = \lim_{T\to\infty}\left\{E[\bar{n}_A] - E\left[\frac{1}{1 - T(N)/T}\sum_{n=1}^{N-1} n\frac{T(n)}{T}\right]\right\}.$$

Let $\lim_{T\to\infty} E[\bar{n}_A] = Q_A$ be the mean number of jobs at a device seen by an arriver and

$$\lim_{T\to\infty}\left\{E\left[\frac{1}{1 - T(N)/T}\sum_{n=1}^{N-1} n\frac{T(n)}{T}\right]\right\} = L_N \quad (7)$$

be the mean state of the device excluding the maximum state N. Then we have

Proposition 7: For a device in a closed network

$$\lim_{T \to \infty} E[e_A^*] = Q_A - L_N . \qquad \square$$

Examining (7) for an open network we see that $T(N)/T \to 0$ as $T \to \infty$. Also,

$$E\left[\sum_{n=1}^{N-1} n\frac{T(n)}{T}\right] = E\left[\sum_{n=1}^{N-1} np(n)\right] = E\left[\sum_{n=1}^{N} np(n) - Np(N)\right].$$

Then

$$\lim_{T \to \infty} E[\overline{n} - Np(N)] = Q$$

where Q is the mean number of jobs at the device. Therefore,

Proposition 8: For a device in an open network

$$\lim_{T \to \infty} E[e_A^*] = Q_A - Q . \qquad \square$$

The two quantities, Q_A and L_N, are not necessarily the same. Take, for example, the sequence in Figure 1. In this example,

$$Q_A = (1)\left(\frac{1k}{2k}\right) = \frac{1}{2}$$

$$L_N = \frac{(1)(2k)}{(4-1)k} = \frac{2}{3}$$

where k is an arbitrary number of cycles. In this case,

$$\lim_{T \to \infty} E[e_A^*] = \frac{1}{2} - \frac{2}{3} = -\frac{1}{6} .$$
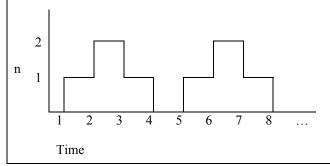

Figure 1. Number, n, at a device over time.

By some algebraic manipulation we can find that the average for any behavior sequence excluding the maximum state, $\overline{n}_N$, is related to average queue length, $\overline{n}$, by

$$\overline{n}_N = \frac{1}{T - T(N)} \sum_{n=1}^{N-1} nT(n)$$

$$= \frac{T}{T - T(N)} \frac{1}{T} \left(\sum_{n=1}^{N-1} nT(n) + NT(N) - NT(N) \right) = \frac{T}{T - T(N)} \left(\overline{n} - \frac{NT(N)}{T} \right)$$

$$= \frac{\overline{n} - Np(N)}{1 - p(N)} . \qquad (8)$$

Equation (6) can be used to give a relationship between the outside observer's average queue length and arriver's average queue length when arrivals are homogeneous. From (6) and homogeneous arrivals,

$$e_A^* = 0 = \overline{n}_A - \frac{1}{T - T(N)} \sum_{n=1}^{N-1} nT(n) .$$

Solving for $\overline{n}_A$ yields

$$\overline{n}_A = \frac{1}{T - T(N)} \sum_{n=1}^{N-1} nT(n)$$

which is the same as

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

Therefore,

$$\overline{n} = \overline{n}_A \left(1 - p(N) \right) + Np(N) \qquad (9)$$

.

Notice that if $p(N)$ is very small, then from (8) $\overline{n}_N \approx \overline{n}$ and from (9) $\overline{n} \approx \overline{n}_A$. Also, note that solving (8) for n and equating with (9) yields $\overline{n}_N = \overline{n}_A$ if the weak homogeneous arrival condition, $e_A^* = 0$, holds.

VI. SERVICE HOMOGENEITY

The homogeneous service error measure [15] is

$$e_S(n) = \frac{1}{p_C(n-1)} \left(\frac{T(n)}{T - T(0)} \right) - 1 . \qquad (10)$$

Taking $\lim_{T \to \infty}$ of the expected value of each side, substituting the definition

$$p_C(n-1) = \frac{C(n)}{C}$$

into the right hand side and manipulating yields

$$\lim_{T \to \infty} E[e_S(n)] = E\left[\lim_{T \to \infty} \left(\frac{T(n)}{C(n)} \right) \lim_{T \to \infty} \left(\frac{C}{T} \right) \lim_{T \to \infty} \left(\frac{1}{1 - T(0)/T} \right) \right] - 1 .$$

$C(n)/T(n)$ is a service rate, therefore the first limit term on the right is the mean time between completions while in state n, $S(n)$. The second limit term is equal to $1/S_o$, where S_o is the unrestricted overall mean time between completions. The last limit term is the reciprocal of the proportion of busy time for the device, i.e., the reciprocal of the utilization. Substituting these descriptions yields

$$\lim_{T \to \infty} E[e_S(n)]$$

$$= \lim_{T \to \infty} E\left[\frac{(\text{time between completions while in state n})}{(\text{overall time between completions})(\text{utilization})} \right] - 1$$

$$= \lim_{T \to \infty} E\left[\frac{S(n)}{S_o U} \right] - 1 .$$

Let

$$G(n, U) = \lim_{T \to \infty} E\left[\frac{S(n)}{S_o U} \right]$$

so that we have

Proposition 9: In an open or closed network

$$\lim_{T \to \infty} E[e_S(n)] = G(n,U) - 1. \quad \square$$

The weak homogeneous service error measure definition [15] is

$$e_S^* = \sum_{n=1}^{N} n p_C (n-1) e_S(n) .$$

Substituting the definition of $e_S(n)$ given in (10) and manipulating gives

$$e_S^* = \sum_{n=1}^{N} \left[n \left(\frac{T(n)}{T - T(0)} - p_C(n-1) \right) \right] \quad (11)$$

or

$$e_S^* = \sum_{n=1}^{N} \left[n \left(\frac{p(n)}{1 - p(0)} - p_C(n-1) \right) \right] .$$

Taking $\lim_{T \to \infty}$ of the expected values of (11) we can derive

$$\lim_{T \to \infty} E[e_S^*] = \lim_{T \to \infty} \left\{ E\left[\frac{1}{T - T(0)} \sum_{n=1}^{N} n T(n) \right] - E\left[\sum_{n=1}^{N} n p_C (n-1) \right] \right\} .$$

The first term on the right is the mean state during the device busy periods. Call it L_0. The second term is the mean queue length seen by a completing job, Q_C. Now we can state *Proposition 10*: In an open or closed network

$$\lim_{T \to \infty} E[e_S^*] = L_0 - Q_C \quad \square$$

Again, the two means, L_0 and Q_C, are not necessarily the same. Using Fig. 1

$$L_0 = \frac{(1)(2k) + (2)(1k)}{(4-1)k} = \frac{4}{3}$$

$$Q_C = (1)\left(\frac{1k}{2k} \right) + (2)\left(\frac{1k}{2k} \right) = \frac{3}{2}$$

where k is the number of cycles. Then,

$$\lim_{T \to \infty} E\left[e_S^* \right] = \frac{4}{3} - \frac{3}{2} = -\frac{1}{6} .$$

VII. CONCLUSION

Table II summarizes the error limit results for the three assumptions studied. Notice that the results for open and closed systems are the same except for the assumption of arrival homogeneity. Even these values will be close if the time spent at the maximum state is small compared to the total time of the observation period. We can conclude that the differences in device performance between open and closed systems are minor when observed for a sufficient length of time.

Of the three main Operational Analysis data assumptions used, only job flow balance is of little concern. The homogeneity assumptions for arrivals and services must be checked for their influence on performance results. Confidence in the use of performance measure relationships such as those in Table I for the types of systems considered in this paper can be enhanced by knowledge of the significance of the assumptions upon which those relationships are based.

Increased knowledge of operational analysis assumption errors can lead to more confident use of performance measures of systems of interest without respect to stochastic assumptions about those systems. As Sevcik and Klawe stated shortly after OA was introduced "Because operational analysis is based on assumptions that can be tested but that are very unlikely to be satisfied exactly in any finite time period, it is very important to develop a means of dealing with 'fuzzy homogeneity' or situations in which the various independence assumptions are satisfied within some tolerance [16].

Further investigation could be made into what types of systems or other circumstances influence the magnitude of long term errors for the homogeneity assumptions. Knowledge of error measures enable correction terms to be formed so that performance measure estimates could be improved. Bounded values of OA performance measure correction terms can also be investigated.

Table II. Limiting Values of Expectations for Various Operational Analysis Assumption Error Measures for Both Open and Closed Networks.

Assumption	Error Measure $\lim_{t \to \infty}$	Network	
		Closed	Open
Flow Balance	$E\left[e_F(n)\right]$	0	0
	$E\left[e_F\right]$	0	0
	$V\left[e_F(n)\right]$	0	0
	$V\left[e_F\right]$	0	0
Arrival Homogeneity	$E\left[e_A(n)\right]$	$\lambda(n)\left(\dfrac{1}{\lambda_0} - p_N(A)\right) - 1$	$\dfrac{\lambda(n)}{\lambda_0} - 1$
	$E\left[e_A^*\right]$	$Q_A - L_N$	$Q_A - Q$
Service Homogeneity	$E\left[e_S(n)\right]$	$G(n,U) - 1$	$G(n,U) - 1$
	$E\left[e_S^*\right]$	$L_0 - Q_C$	$L_0 - Q_C$

REFERENCES

[1] J. P. Buzen, "Fundamental Operational Laws of Computer Systems Performance," *Acta Informatica*, vol. 7, pp. 183-195, 1976.

[2] J. P. Buzen, "Operational Analysis: An Alternative to Stochastic Modeling." *Performance of Computer Installation*," North-Holland Publishing Company, Amsterdam, The Netherlands, 1978.

[3] J. P. Buzen and P. J. Denning, "Operational Treatment of Queue Distributions and Mean Value Analysis," Dept. of Computer Science, Purdue University, CSD-TR 309, 1979.

[4] P. J. Denning and J. P. Buzen, "The Operational Analysis of Queueing Network Models," *Computer Surveys*, vol. 10, no. 3, 1980.

[5] R. Suri, "Robustness of Queueing Network Formulas," *Journal of the Association for Computing Machinery*, vol. 30, no. 3, pp. 564-594, 1983.

[6] E. D. Lazowska, J. Zahorjan, and E. C. Sevcik, "Computer System Performance Evaluation Using Queueing Network Models," *Annual Reviews in Computer Science*, vol. 1, pp. 107-137, 1986.

[7] B. V. Ivanovskii, "Operational Analysis of Communication Networks with Locks," *Automatic Control and Computer Sciences*, vol. 22, issue 3, pp. 51-66, 1988.

[8] P. J. Denning, "Queueing in Networks of Computers," *American Scientist*, vol. 79, May-June, pp. 206-209, 1991a.

[9] P. J. Denning, "In the Queue: Mean Values," *American Scientist*, vol. 79, Sep.-Oct., pp.402-403, 1991b.

[10] Y. Dalley and X. Cao, "Operational Analysis of Stochastic Closed Queueing Networks," *Performance Evaluation*, vol. 14, pp. 43-61, 1992.

[11] M. El-Taha and S. Stidham, "Deterministic Analysis of Queueing Systems with Heterogeneous Servers," *Theoretical Computer Science*, vol. 106, pp. 243-364, 1992.

[12] P. J. Denning, "Operational Analysis," *Computer System Performance Modeling in Perspective,* E. Gelenbe, Ed., Imperial College Press, pp. 21-33, 2006.

[13] F. Baskett, K. Chandy, R. Muntz, and F. Palacios, "Open, closed, and mixed networks of queues with different classes of customers," *Journal of the ACM*, vol. 22, No. 2, pp. 248-260, 1975.

[14] J. P. Buzen, "From Stochastic Modeling to Operational Analysis: The Journey Begins," *Fundamental Concepts in Computer Science*, E. Gelenbe and J-P. Kahane Eds., Imperial College Press, pp.141-149, 2009.

[15] N. M. Bengtson, "Measuring Errors in Operational Analysis Assumptions," *IEEE Transactions on Software Engineering*, vol. SE-13, no. 7, pp. 767-776, 1987.

[16] K. C. Sevcik and M. Klawe, "Operational Analysis versus Stochastic Modeling of Computer Systems," *Proc. Of Computer Science Statistics: 12th Annual Symposium on the Interface*, Waterloo, Canada, pp. 177-184, 1979.

[17] E. D. Lazowska, J. Zahorjan, G. S. Graham, and K. C. Sevcik, *Quantitative System Performance*, Prentice-Hall, Englewood Cliffs, NJ, 1984.

[18] Y. Dallery and R. David, "Operational analysis of multiclass queueing networks," *25th IEEE Conference on Decision and Control*, pp.1728-1732, 1986.

[19] N. M. Bengtson, "Using Operational Analysis in Simulation: A Queuing Network Example," *Journal of the Operational Research Society*, vol. 39, no. 12, Dec. 1988.

[20] J. P. Buzen, "An Overview of Performance Prediction in MVS Systems and SNA Networks," *Proceedings of 1986 ACM Fall Joint Computer Conference*, Dallas, TX, pp. 751-759, 1986.

[21] D. A. Menasce and V. A. Almeida, *Scaling for E-Business: technologies, models, performance, and capacity planning*, Prentice-Hall, Englewood Cliffs, NJ, 2000.

[22] D. K. Buch and V. M. Pentkovski, "Performance Characterization Experience of Multi-Tier E-business Systems using Queuing Operational Analysis," *IEEE International Symposium on Performance Analysis of Systems and Software*, Tucson, AZ, 2001.

[23] M. P. Fanti, B. Maione, G. Piscitelli, and B. Turchiano, "Two Methods for Real-Time Routing Selection in Flexible Manufacturing Systems," *Proc. of the 1992 IEEE International Conference on Robotics and Automation*, pp. 1158-1166, 1992.

[24] W. M. Zuberek, "Throughput Analysis in Timed Petri Nets," *35th Midwest Symposium of Circuits and Systems*, Washington, D. C., pp. 1576-1580, 1992.

[25] W. M. Zuberek, "Throughput Analysis of Simple Closed Time Petri Net Models," *Proceedings of the 36th Midwest Symposium on Circuits and Systems*, Chicago, IL, pp. 2721-2724, 1993.

[26] G. Chiola, C. Anglano, J. Campos, J. M. Colom, and M. Silva, "Operational Analysis of Timed Petri Nets and Application to the Computation of Performance Bounds," *Quantitative Methods in Parallel Systems*, F. Baccelli, A. Jean-Marie, and I. Mitranii Eds., Esprit Basic Research Series, Springer-Verlag, Berlin Heidelberg New York, pp. 161-174, 1995.

[27] J. P. Buzen and P. J. Denning, "Measuring and Calculating Queue Length Distributions," *Computer*, vol. 13, no. 4, pp. 33-44, 1980.

[28] E. Çinlar, *Introduction to Stochastic Processes*, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1975.

[29] J. A. Brumfield, "Operational Analysis of Queuing Phenomena," Ph.D. dissertation, Dept. of Computer Science, Purdue Univ., West Lafayette, IN, 1982.

[30] M. El-Taha and S. Stidham, "Deterministic Analysis of Multi-Channel Queueing Systems with Heterogeneous Servers," Dept. of Operations Research, TR90-7, University of North Carolina, Chapel Hill, NC, 1990.

APPENDIX

The following are proofs of Propositions 1 through 4.

Proposition 1: For each device in an open or closed network

$$\lim_{T \to \infty} E[e_F(n)] = 0.$$

Proof: By definition

$$e_F(n) = (n+1)(p_A(n) - p_C(n)) \; .$$

Substituting the definitions of $p_A(n)$ and $p_C(n)$ and taking limits of expectations gives

$$\lim_{T \to \infty} E\left[e_F(n)\right] = \lim_{T \to \infty} E\left[(n+1)\left(\frac{A(n)}{A} - \frac{C(n+1)}{C}\right)\right]$$
$$= (n+1)\left\{\lim_{T \to \infty} E\left[\frac{A(n)}{A}\right] - \lim_{T \to \infty} E\left[\frac{C(n+1)}{C}\right]\right\} \; . \tag{A}$$

Let $A(n) = C(n+1) + d(n)$ [29]

where

$$d(n) = \begin{cases} -1, \text{ for } n(T) \le n < n(0) \;\; (A < C) \\ +1, \text{ for } n(0) \le n < n(T) \;\; (A > C) \\ 0, \text{ otherwise} \end{cases} \; .$$

Then, $C(n+1) = A(n) - d(n)$ and substituting into (A) gives

$$\lim_{T \to \infty} E\left[e_F(n)\right] = (n+1)\left\{E\left[\lim_{T \to \infty}\left(\frac{A(n)}{A}\right)\right] - E\left[\lim_{T \to \infty}\left(\frac{A(n) - d(n)}{A - \sum_{n=0}^{N-1} d(n)}\right)\right]\right\}$$
$$= (n+1)\left\{E\left[\lim_{T \to \infty}\left(\frac{A(n)}{A}\right)\right] - E\left[\lim_{T \to \infty}\left(\frac{A(n)/A - d(n)/A}{1 - \sum_{n=0}^{N-1} d(n)/A}\right)\right]\right\} \; . \tag{B}$$

The second limit on the right hand side of (B) equals $\lim_{T \to \infty}\left(\frac{A(n)}{A}\right)$ since both $d(n)/A$ and $\sum_{n=0}^{N-1} d(n)/A$ go to 0 in the limit. As $t \to \infty$, $A \to \infty$, but $|d(n)| \le 1$. Therefore, $\lim_{T \to \infty}(d(n)) = 0$. Also, $\lim_{T \to \infty}\left[\left(\sum_{n=0}^{N-1} d(n)\right)/A\right]$ is bounded by

$$\lim_{T \to \infty}(N/A). \tag{C}$$

As is shown in the Lemma which follows, $\lim_{T \to \infty}(N/A) = 0$ and

$$\lim_{T\to\infty}\left(\frac{N}{A}\right) \geq \lim_{T\to\infty}\left[\frac{\sum_{n=0}^{N-1} d(n)}{A}\right] = 0 .$$

Substituting (B) into (C) gives

$$\lim_{T\to\infty} E\big[e_F(n)\big] = (n+1)\left\{\lim_{T\to\infty} E\left[\frac{A(n)}{A}\right] - \lim_{T\to\infty} E\left[\frac{A(n)}{A}\right]\right\} = 0. \qquad \square$$

Lemma: Let the device have jobs arriving with average rate λ. Then

$$\lim_{T\to\infty}\left(\frac{N(T)}{A(T)}\right) = 0$$

Where $N(T)$ is the largest number of jobs at the device at any time up to time T and $A(T)$ is the total number of arrivals to the device up to time T.

Proof: Let $\dfrac{N(T)}{A(T)} = \dfrac{N(T)/T}{A(T)/T}$. Since $\dfrac{A(T)}{T} \to \lambda$ as time increases, it is sufficient to show that $\lim_{T\to\infty}\left(\dfrac{N(T)}{T}\right) = 0$. Let $L(t)$ be the number of jobs at the device at time t. Since the system under consideration is stable, $\dfrac{L(t)}{t} \to 0$ as time increases [30]. Suppose $\lim_{T\to\infty}\left(\dfrac{N(T)}{T}\right) \neq 0$, then $\exists \varepsilon > 0$ and a sequence $\{t_m : t_m \to \infty\}$ such that $\dfrac{N(t_m)}{t_m} \geq \varepsilon \quad \forall m$. For each m define $s_m \leq t_m$ such that $N(t_m) = L(s_m)$, i.e. s_m is a point in time when the maximum number of jobs are at the device. Now,

$$\varepsilon \leq \frac{N(t_m)}{t_m} \leq \frac{L(s_m)}{s_m} .$$

Therefore, we have an $\varepsilon > 0$ and a sequence such that $\dfrac{L(s_m)}{s_m} \geq \varepsilon \quad \forall m$. This is a contradiction to $\dfrac{L(t)}{t} \to 0$. Therefore,

$$\lim_{T\to\infty}\left(\frac{N(T)}{T}\right) = 0 . \qquad \square$$

Proposition 2: For a device in an open or closed network

$$\lim_{T\to\infty} E\big[e_F\big] = 0 .$$

Proof: By definition,

$$e_F = \sum_{n=0}^{N-1} e_F(n) .$$

Then, using Proposition 1

$$\lim_{T\to\infty} E\big[e_F\big] = \lim_{T\to\infty} E\left[\sum_{n=0}^{N-1} e_F(n)\right]$$
$$= \sum_{n=0}^{N-1}\left[\lim_{T\to\infty} E\big[e_F(n)\big]\right] = 0 . \qquad \square$$

Proposition 3: For a device in an open or closed network

$$\lim_{T\to\infty} V\big[e_F(n)\big] = 0 .$$

Proof: By the definition of variance,

$$\lim_{T\to\infty} V\big[e_F(n)\big] = \lim_{T\to\infty}\big[(e_F(n) - E\big[e_F(n)\big])^2\big]$$
$$= \lim_{T\to\infty}\big(e_F^2(n)\big) - 2\lim_{T\to\infty}\big(e_F(n)E\big[e_F(n)\big]\big) + \lim_{T\to\infty} E\big[e_F(n)\big]^2 . \qquad (D)$$

The first term on the right hand side can be changed by substituting the definition of $e_F(n)$ to give

$$\lim_{T\to\infty} e_F^2 = \lim_{T\to\infty}\big[(n+1)^2(p_A(n) - p_C(n))^2\big]$$
$$= (n+1)^2 \lim_{T\to\infty}\left[\left(\frac{A(n)}{A}\right)^2 - 2\left(\frac{A(n)}{A}\right)\left(\frac{C(n+1)}{C}\right) + \left(\frac{C(n+1)}{C}\right)^2\right].$$

Substituting and manipulating $C(n+1)=A(n)-d(n)$ as in Proposition 1 gives

$$\lim_{T\to\infty} e_F^2(n) = (n+1)^2 \lim_{T\to\infty}\left[\left(\frac{A(n)}{A}\right)^2 - 2\left(\frac{A(n)}{A}\right)\left(\frac{A(n)-d(n)}{A-\sum d(n)}\right) + \left(\frac{A(n)-d(n)}{A-\sum d(n)}\right)^2\right]$$

$$= (n+1)^2\left\{\lim_{T\to\infty}\left[\left(\frac{A(n)}{A}\right)^2\right] - 2\lim_{T\to\infty}\left[\left(\frac{A(n)}{A}\right)\left(\frac{A(n)/A - d(n)/A}{1 - \sum d(n)/A}\right)\right] + \lim_{T\to\infty}\left[\left(\frac{A(n)/A - d(n)/A}{1 - \sum d(n)/A}\right)^2\right]\right\}.$$

Now, as $t \to \infty$, $d(n)/A \to 0$ and $\sum d(n)/A \to 0$. Let, $\lim_{T\to\infty}\dfrac{A(n)}{A} \to \dot{p}_A(n)$. Substitution yields

$$\lim_{T\to\infty} e_F^2(n) = (n+1)^2\big[\dot{p}_A(n)^2 - 2\dot{p}_A(n)^2 + \dot{p}_A(n)^2\big] = 0 .$$

The second term on the right hand side of (D) can also be shown to go to 0 by first substituting the error measure definition, then substituting and manipulating $C(n+1)=A(n)-d(n)$ as before to yield

$$- 2 \lim_{T \to \infty} \left(e_F(n) E[e_F(n)] \right)$$

$$= - 2 \lim_{T \to \infty} \left[(n+1) \left(\frac{A(n)}{A} - \frac{C(n+1)}{C} \right) E[e_F(n)] \right]$$

$$= - 2 \lim_{T \to \infty} \left[(n+1) \left(\frac{A(n)}{A} - \frac{A(n) - d(n)}{A - \sum d(n)} \right) E[e_F(n)] \right]$$

$$= - 2 \lim_{T \to \infty} \left[(n+1) \left(\dot{p}_A(n) - \dot{p}_A(n) \right) E[e_F(n)] \right] = 0 .$$

The third term on the right hand side of (D) is also equal to 0 by Proposition 2. Substituting these results into (D) yields,

$$\lim_{T \to \infty} V[e_F(n)] = 0 . \qquad \square$$

Proposition 4: For a device in an open or closed network

$$\lim_{T \to \infty} V[e_F] = 0 .$$

Proof: Substituting the definition of e_F and by Proposition 3

$$\lim_{t \to \infty} V[e_F] = \lim_{T \to \infty} V\left[\sum_{n=0}^{N-1} e_F(n) \right] = \sum_{n=0}^{N-1} \lim_{T \to \infty} V[e_F(n)] = 0 \qquad \square$$

Neal M. Bengtson received B.S. degrees in Aerospace Engineering and in Computer Science from North Carolina State University, M.S. degrees in Industrial and Systems Engineering from The University of Alabama in Huntsville and in Management from North Carolina State University, and the Ph.D. degree in Operations Research from the School of Industrial Engineering at Purdue University.

He worked in the aerospace industry for Fairchild-Hiller Corporation and, as a systems analyst and department head, for Data Processing Associates. He was on the faculty of North Carolina State University in the Computer Science Department until leaving to work at IBM. He has also taught courses for Purdue University, The University of North Carolina at Chapel Hill, and Meredith College. From 1996 to the present he has been at Barton College in the School of Business.