1982

Error Analysis of the Mean Busy Period of a Queue

Peter J. Denning

Wolfgang Kowalk

Report Number:
81-398
ERROR ANALYSIS OF THE MEAN BUSY PERIOD OF A QUEUE

Peter J. Denning†
Wolfgang Kowalk†

April 1982
CSD-TR-398

Abstract: The error between the actual value of the mean busy period and the value estimated by a common queueing formula \( D = S/(1-U) \) is evaluated.

†Denning's address: Department of Computer Sciences, Purdue University, W. Lafayette, IN 47907 USA. Kowalk's address: Fachbereich Informatik, Universität Hamburg, Rotherbaumhause 67/69, 2000 Hamburg 13, West Germany.
1. Introduction

The accuracy of common queueing formulae when used with data obtained from real systems has been of great concern to queueing theorists and performance analysts. In operational analysis [1,2] this question can be addressed directly because the error in a formula can be expressed in terms of the errors in each of the assumptions on which the formula relies. Buzen and Denning illustrated this idea in a paper published in 1980 [3]. Kowalk has since suggested a general method of tracing errors from assumptions to results by annotating each step of a derivation with the error present at that step [4,5].

An example already analyzed fully is the formula for the mean response time, \( R \), of a single server with unbounded queue [6,7]:

\[
R = \frac{S}{1-U},
\]

where \( S \) is the mean service time per completed job and \( U \) is the utilization. This formula is exact when arrivals and services are both homogeneous — i.e., when neither the arrival rate nor the mean service time depends on the queue length [3]. We showed that the actual value of response time lies in the interval \([R-\epsilon, R+\epsilon]\), where

\[
\epsilon = \bar{n}E_A(1+E_S)+(\bar{n}+1)E_S.
\]

In this expression, \( E_A \) is the (absolute) relative error in the arrival assumption, i.e., the maximum of the (absolute) differences between actual and assumed arrival rates divided by the assumed arrival rate. Similarly, \( E_S \) is the relative error in the service assumption. The quantity \( \bar{n} \) is the mean queue length dur-
ing the observation period.

Formula (2) allows us to make statements such as the following. Suppose we know that the queue-dependent arrival rate and mean service times are within 10% of being constant, the mean service time is 2 seconds, the mean queue length is 3, and the utilization is 75%. Then the estimated value of mean response time is $R = 8$ seconds and the error tolerance is $\varepsilon = 0.7$ seconds. Thus the actual value of response time is between 7.3 and 8.7 seconds.

In this paper we will illustrate the principles of error analysis for a very simple case, the length of the mean busy period of a queue.

2. Mean Busy Period: Exact Analysis

Table I summarizes the notation. A busy period is a maximal interval of service during which $n(t) > 0$. We will assume "end effects" have been removed from the data: each end of the observation period $[0, T]$ falls either in an idle period or just after a completion. Under this assumption, the number of busy periods is

$$K = \begin{cases} A(0), & n(0) = 0 \\ A(0) + 1, & n(0) > 0 \end{cases}$$

(3)

where $A(0)$ is the number of arrivals witnessing $n(t) = 0$. 
## TABLE I. OPERATIONAL NOTATION

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A )</td>
<td>Number of arrivals</td>
<td></td>
</tr>
<tr>
<td>( C )</td>
<td>Number of completions</td>
<td></td>
</tr>
<tr>
<td>( T )</td>
<td>Length of observation period</td>
<td></td>
</tr>
<tr>
<td>( n(t) )</td>
<td>Number in system at time ( t )</td>
<td></td>
</tr>
<tr>
<td>( T(n) )</td>
<td>Total time during which ( n(t)=n ) ( (T = \sum_{n=1}^{n} T(n)) )</td>
<td></td>
</tr>
<tr>
<td>( A(n) )</td>
<td>Number of arrivals observing ( n(t)=n )</td>
<td></td>
</tr>
<tr>
<td>( \lambda )</td>
<td>( A / T )</td>
<td>Arrival rate</td>
</tr>
<tr>
<td>( \lambda(n) )</td>
<td>( A(n) / T(n) )</td>
<td>Conditional arrival rate</td>
</tr>
<tr>
<td>( X )</td>
<td>( C / T )</td>
<td>Completion rate</td>
</tr>
<tr>
<td>( U )</td>
<td>( (T-T(0)) / T )</td>
<td>Utilization</td>
</tr>
<tr>
<td>( S )</td>
<td>( (T-T(0)) / C )</td>
<td>Mean service time</td>
</tr>
<tr>
<td>( K )</td>
<td>Number of busy periods</td>
<td></td>
</tr>
<tr>
<td>( D_k )</td>
<td>Length of ( k^{th} ) busy period ( (k=1, ..., K) )</td>
<td></td>
</tr>
<tr>
<td>( \overline{D} )</td>
<td>( \frac{1}{K} \sum_{k=1}^{K} D_k )</td>
<td>Mean busy period</td>
</tr>
</tbody>
</table>

The following analysis depends on the three assumptions summarized in Table II. **Idle Start (IS)** asserts that \( n(0)=0 \), in which case the number of busy periods, \( K \), is the same as \( A(0) \). **Homogeneous Arrivals (HA)** asserts that the arrival rate for empty queue is the same as the overall arrival rate; this is a special case of the general homogeneous arrivals assumption [3,6,7] that \( \lambda(n)=\lambda \) for all \( n \). **Flow Balance (FB)** asserts that the number of arrivals and completions is the same or, equivalently, \( \lambda=X \).

The error introduced by each assumption is denoted by lower-case \( e \) with appropriate subscript. The (absolute) relative error bound is denoted by upper-case \( E \) with appropriate subscript. Relative errors are defined with respect to assumed rather than actual value.
We begin by noting that the HA assumption implies

\[ \frac{T(0)}{A(0)} = \frac{1}{\lambda(0)} = \frac{1}{\lambda} = \frac{T}{A}, \tag{4} \]

which is already an interesting result. It states that the mean interarrival time is the same as the mean idle period when arrivals are homogeneous.

The mean busy period is the total busy time divided by the number of busy periods:

\[ \bar{D} = \frac{T-T(0)}{K} \]

\[ = \frac{T(0)}{A(0)} \frac{T-T(0)}{T(0)} \frac{A(0)}{K} \]

\[ = \frac{1}{\lambda(0)} \frac{U}{1-U} \frac{A(0)}{K}. \tag{5} \]

This identity is reduced as follows:

\[ \bar{D} = \frac{1}{\lambda(0)} \frac{U}{1-U} \quad [\text{IS assumption}] \]
\[ \frac{1}{\lambda} \frac{U}{1-U} \quad [\text{HA assumption}] \]

\[ \frac{1}{\lambda} \frac{U}{1-U} \quad [\text{FB assumption}] \]

On applying the utilization law \((U = XS)\), we obtain

**Theorem 1.** Assumptions IS, HA, and FB imply that the mean busy period length is

\[ \overline{D} = \frac{S}{1-U} \quad (6) \]

This is the same as the response time of a flow balanced queue with homogeneous arrivals and services.

3. **Exact Analysis of Error**

Our next goal is to express the mean busy period in the form

\[ \overline{D} = \frac{S}{1-U} + \varepsilon \]

where the error \(\varepsilon\) is a function of the three assumption errors in Table II. We start by rewriting (5) in the form

\[ \overline{D} = \frac{1}{\lambda} \frac{U}{1-U} + \frac{U}{1-U} \left( \frac{A(0)}{K\lambda(0)} - \frac{1}{\lambda(0)} \right) + \frac{1}{\lambda} \left( \frac{1}{\lambda} - \frac{1}{X} \right) \]

The three bracketed terms reduce to the error measures of Table II as follows:

\[ \frac{A(0)}{K\lambda(0)} - \frac{1}{\lambda(0)} = \frac{\varepsilon_I}{K\lambda(0)} \]

\[ \frac{1}{\lambda(0)} - \frac{1}{\lambda} = \frac{\varepsilon_H}{\lambda\lambda(0)} \]
Therefore,

\[
\frac{1}{\lambda} - \frac{1}{\lambda} = \frac{e_B}{\lambda C}
\]

Theorem 2. The mean busy period length is

\[
\bar{D} = \frac{S}{1 - U} + e,
\]

where the error is

\[
e = \frac{U}{1 - U} \left[ \frac{e_I}{K\lambda(0)} + \frac{e_H}{\lambda\lambda(0)} + \frac{e_B}{\lambda C} \right].
\] (7)

4. Error Bound Analysis

Equation (7) shows that an exact characterization of error requires considerable information about the system during the observation period. It is sometimes useful to work with (simpler) bounds on these errors. By taking the magnitude of the error and applying the triangle inequality, we find

\[
|e| \leq \frac{U}{1 - U} \left[ \frac{|e_I|}{A(0)} \frac{A(0)}{K\lambda(0)} + \frac{|e_H|}{\lambda(0)} \frac{1}{\lambda(0)} + \frac{|e_B|}{C} \frac{1}{\lambda} \right].
\]

On employing the definitions in Table II and the fact \(A(0)/K \leq 1\), we find

Theorem 3. An upper bound on the error of the mean busy period formula is

\[
|e| \leq \frac{U}{1 - U} \left[ \frac{E_I + E_H}{\lambda(0)} + \frac{E_B}{\lambda} \right].
\] (8)
In practice, the queue has some maximum possible length, say $N$, which implies that $E_B \leq N/C$. By observing the queue for a sufficient period, both $E_B$ and $E_I$ can be made arbitrarily small. This implies that the error in the busy period formula is dominated by the error in the HA assumption, i.e.,

$$|e| \leq \frac{U}{1-U} \frac{E_H}{\lambda(0)}$$

for long observation periods.

5. Approximation Analysis

We will briefly illustrate how the error bound (8) can be obtained directly by using an algebra based on approximate equality [4,5]. Let the symbol $=_{e}$ denote "approximately equal with absolute error $\varepsilon \geq 0"$. We write $Y =_{e} Z$ to mean

$$Z - \varepsilon \leq Y \leq Z + \varepsilon$$

or, equivalently, $|Y - Z| \leq \varepsilon$. The notation $=_{e}$ is used only symbolically -- i.e., an expression for $\varepsilon$ must be stated explicitly. The three assumptions in Table II can be expressed in this notation:

Assumption IS: $A(0) =_{e_I} K$  \[ e_I = E_I A(0) \]

Assumption HA: $\lambda(0) =_{e_A} \lambda$  \[ e_A = E_A \lambda \]

Assumption FB: $A =_{e_B} C$  \[ e_B = E_B C \]
Note that

\[
\frac{T(0)}{A(0)} = \frac{1}{\lambda(0)} = \frac{1}{\lambda} = \frac{T}{A}
\]

Therefore, Relation (10) shows that the mean interarrival time is approximately the same as the mean idle period if arrivals are approximately homogeneous. Note that

\[
\frac{1}{\lambda} = \frac{1}{\lambda(0)} = \frac{T}{A}
\]

Note finally this general rule: If \( Y = a_1 a Z \) and \( a = b_1 \), then \( Y = a_1 b Z \) with

\[
\varepsilon = \delta_1 + \delta_2.
\]

The derivation of the mean busy period formula in this algebra is as follows:

\[
\overline{D} = \frac{1}{\lambda(0)} \frac{U}{1-U} \frac{A(0)}{K} \quad \text{[Equation (5)]}
\]

\[
= \frac{1}{\lambda(0)} \frac{U}{1-U} \frac{A(0)}{K}
\]

\[
\Rightarrow \varepsilon_5 = \frac{U}{1-U} \frac{A(0)}{K}
\]

\[
\Rightarrow \varepsilon_4 = \varepsilon_5 + \frac{U}{1-U} \varepsilon_1
\]

\[
\Rightarrow \varepsilon_5 = \frac{U}{1-U} \varepsilon_2
\]

Collecting all error terms and applying the utilization law,

\[
\overline{D} = \varepsilon \quad \frac{S}{1-U}
\]

\[
\varepsilon = \varepsilon_5 = \frac{U}{1-U} \left[ \frac{E_f + E_k}{\lambda(0)} + \frac{E_b}{\lambda} \right]
\]

The error \( \varepsilon \) is identical to the bound in Relation (8). This bound loses accuracy as \( U \to 1 \) (equivalently, \( \lambda(0) \to 0 \)) because \( S/(1-U) \to \infty \) even though the actual
value of $D$ is less than $T$. Stated differently, when $U > 1-S/T$ a better estimate of $D$ is $T$.

5. Acknowledgements

We are grateful to J. Brumfield, J. Buzen, and E. Jessen for valuable comments and advice. Part of this research was supported at Purdue University by NSF Grant MCS78-01729.

6. References


7. J. Brumfield and P. Denning, "Error analysis of homogeneous mean queue and response time estimators," Purdue University, Computer Sciences Department, W. Lafayette, IN 47907, Report