Proceedings of the Industrial Engineering Research Conference (IERC), Atlanta, GA, May 2005

Using Fractional Numbers of Customers to Achieve Throughputs in Queueing Networks

Rajan Suri
Rahul Shinde
Department of Industrial Engineering, University of Wisconsin-Madison
1513 University Avenue, Madison, WI 53706
and
Mary Vernon
Department of Computer Sciences, University of Wisconsin-Madison
1210 W Dayton Street, Madison, WI 53706

Abstract

One of the design parameters in closed queueing networks is N_p, the number of customers of class p. It has been assumed that N_p must be an integer. However, integer choices will usually not achieve the target throughput for each class simultaneously. We use Mean Value Analysis with the Schweitzer-Bard approximation and nonlinear programming to determine the value of N_p needed to achieve the production targets exactly, although the values of N_p may be fractional. We interpret these values to represent the average number of customers of each class in the network. We implement a control rule to achieve these averages and verify our approach through simulation.

Key words: closed queueing networks, mean value analysis, fractional customer populations

1 Introduction

Closed queueing networks are used for modelling and predicting performance in many applications. A few examples of such applications are job shops, automated manufacturing facilities such as flexible manufacturing systems, computer systems performance modelling and material control strategies [9, 6]. These models are popular in diverse fields because they work well in many practical situations [5, 7]. Traditionally, the customer population for different classes of customers in closed queueing networks has been assumed to be an integer. We find that such an assumption is not always suitable because for some systems exact throughput targets for all classes cannot be achieved simultaneously using integer customer populations. For such systems, we introduce the concept of “fractional number of customers”. We provide a control rule to achieve the fractional customer populations and provide a mixed-integer nonlinear optimization model based on Mean Value Analysis with the Schweitzer-Bard approximation to estimate the (fractional) number of customers to meet the throughput targets exactly. The paper is structured as follows. We describe the analytical model in section 2, followed by the motivation for fractional customer populations in section 3. In section 4 we present the nonlinear optimization model. Results from experimental validation are presented in section 5.

2 Multiple Class Mean Value Analysis

The original Mean Value Analysis (MVA) algorithm [3] involves an iterative procedure to calculate the performance measures of a queueing network. The computational effort for the MVA technique increases exponentially with the number of customer classes. We describe an approximation proposed by Schweitzer [4] and extensively tested by Bard [1] which significantly reduces the MVA computational effort for multi-class systems. We refer to the Schweitzer-Bard approximation for Mean Value Analysis as SB-MVA. We describe SB-MVA for closed queueing networks using the variables defined in Table 1. (For a complete description of MVA and SB-MVA we refer the reader to Kant [2].) In Table 1,

$\vec{N} \equiv \{N_1, N_2, \ldots, N_P\}$ represents the population vector of the closed queueing network where P is the total number of customer classes.

Table 1: Mean Value Analysis Variables

Variable	Description
M	Number of stations in the closed queueing network
$T_{m,p}$	Mean processing time for customer of class p at station m
$R_{m,p}(\vec{N})$	Mean response time for customer of class p at station m with $\vec{N}$ customers in the network
$Q_{m,p}(\vec{N})$	Mean queue length for customer of class p at station m with $\vec{N}$ customers in the network
$X_p(\vec{N})$	Throughput of customer class p in the network with $\vec{N}$ customers

The SB-MVA algorithm states that the response time for a customer of class p arriving at a station m is given by

$$R_{m,p}(\vec{N}) = T_{m,p} + \left(\frac{N_p - 1}{N_p}\right) T_{m,p} Q_{m,p}(\vec{N}) + \sum_{r \neq p} Q_{m,r}(\vec{N}) T_{m,r} \quad m = 1, 2, \ldots, M, \quad p = 1, 2, \ldots, P. \tag{1}$$

Applying Little's Law for class p customers in the closed queueing network gives

$$X_p(\vec{N}) = \frac{N_p}{\sum_{m=1}^{M} R_{m,p}(\vec{N})} \quad p = 1, 2, \ldots, P. \tag{2}$$

Applying Little's Law at station m for class p gives

$$Q_{m,p}(\vec{N}) = X_p(\vec{N}) R_{m,p}(\vec{N}) \quad m = 1, 2, \ldots, M, \quad p = 1, 2, \ldots, P. \tag{3}$$

Substituting the value of $X_p(\vec{N})$ from equation (2) into equation (3) we get

$$Q_{m,p}(\vec{N}) = \frac{N_p R_{m,p}(\vec{N})}{\sum_{m=1}^{M} R_{m,p}(\vec{N})} \tag{4}$$

Further substituting the value of $R_{m,p}(\vec{N})$ from equation (1) gives

$$Q_{m,p}(\vec{N}) = \frac{N_p \left(T_{m,p} + \left(\frac{N_p - 1}{N_p}\right) T_{m,p} Q_{m,p}(\vec{N}) + \sum_{r \neq p} Q_{m,r}(\vec{N}) T_{m,r} \right)}{\sum_{k=1}^{M} \left(T_{k,p} + \left(\frac{N_p - 1}{N_p}\right) T_{k,p} Q_{k,p}(\vec{N}) + \sum_{r \neq p} Q_{k,r}(\vec{N}) T_{k,r} \right)} \quad \begin{array}{c} m = 1, 2, \ldots, M \\ p = 1, 2, \ldots, P \end{array} \tag{5}$$

Equation (5) states a system of MP nonlinear equations in the unknowns $Q_{m,p}(\vec{N})$. For given $\vec{N}$ and $T_{m,p}$ $(\forall m, p)$ these equations can be solved to get the values of $Q_{m,p}(\vec{N})$. From these values of queue lengths, the values of the response times and the throughput can be calculated using equations (1) and (2). In the next section, we present the motivation for this research.

3 Motivation for Fractional Number of Customers

Consider a three station closed queueing network with two customer classes. We assume that both customer classes visit all three stations. Each customer has three operation steps and after completing an operation at a station, for its next step it can choose either of the three stations with equal probability. Here we assume that the network is balanced and the processing times for both classes at each station are exponentially distributed with a mean of one. The customer population for the two classes is specified by (N_1, N_2). The throughput of this network for different integer values of customer populations can be calculated using SB-MVA, and values of the throughput for a few combinations

Table 2: Throughput Observed using Integer Number of Customers in the Network

(N_1,N_2)	Class 1 Throughput	Class 2 Throughput
1,1	0.250	0.250
1,2	0.200	0.400
2,2	0.333	0.333
2,3	0.286	0.429
3,3	0.375	0.375

of (N_1, N_2) are shown in Table 2.

From Table 2, it can be seen that if the target throughput desired is (0.27, 0.31) none of the settings will be able to achieve it. The reason for this is the use of integer customer counts. This difficulty becomes more pronounced when the number of customer classes in the network increases [10].

However, from the system of equations in (5), we can see that there is no requirement in the equations that N_p be integer. Thus, suppose that by using fractional values of N_p, we can obtain a solution to equations (5) such that exact throughput targets can be obtained, such as the (0.27, 0.31) target above. The question then arises, can we obtain a physical realization for such "fractional customers". We propose here that we can indeed get such a realization by varying the number of customers between two neighboring integer values so that the time average of these numbers equals the fractional value. We then need to verify that such an approach does indeed achieve (reasonably) the target throughputs. Following is the description of the control rule to set the fractional number of customers wherein we vary the number of customers between two neighboring integer values.

Normally in a closed queueing network when a customer completes all operation steps and leaves, a new customer is immediately introduced into the network. Instead of doing this we use the following algorithm. Let $\llcorner k \lrcorner$ denote the greatest integer less than or equal to k and $\ulcorner k \urcorner$ denote the smallest integer greater than or equal to k. To achieve the fractional value N_p, we vary the number of customers for the customer class between N_p^L and N_p^U where $N_p^L = \llcorner N_p \lrcorner$ and $N_p^U = \ulcorner N_p \urcorner$ using Algorithm 1.

Algorithm 1 (Setting fractional number of customers).

```
Step 1: Initialize simulation with actual number of customers in the network $n_p = N_p^L$ customers.
Step 2: When a customer completes all processing steps and leaves calculate
        $\bar{N}_p$ = the (time) average number of customers for that class
        If $\bar{N}_p < N_p$ then
                      if $n_p = N_p^L$ introduce 2 customers so that $n_p = N_p^U$
                      if $n_p = N_p^U$ introduce 1 customer and maintain $n_p = N_p^U$
        If $\bar{N}_p > N_p$ then
                      if $n_p = N_p^U$ then do not introduce the customer so that $n_p = N_p^L$
                      if $n_p = N_p^L$ introduce 1 customer and maintain $n_p = N_p^L$
        If $\bar{N}_p = N_p$ then introduce 1 customer and maintain same $n_p$
Repeat Step 2 until simulation replication terminating condition.
```

To verify this approach we will use the target throughputs from our previous example, (0.27, 0.31). It turns out that the (fractional) number of customers which will achieve this target is (1,28, 1.48). (We will explain later in this paper how we obtained this number.) Next, we utilize a simulation model using *Arena*® (www.arenasimulation.com). The fractional number of customers are modelled using the control rule described in Algorithm 1. We then compare the throughput predictions from SB-MVA with the throughputs obtained from simulation. Here, we use a warm up period of 20,000 orders and run for 80,000 orders during which statistics are recorded. 10 replications are performed. Although the 95% confidence intervals are not shown they were less than 1% of the observed mean. Table 3 compares the SB-MVA predictions with throughputs from simulation for our example and three additional settings.

Table 3: Comparison of SB-MVA Predictions with Throughputs from Simulation

Populations	Class 1			Class 2		
(N_1, N_2)	SB-MVA Throughput	Simulation Throughput	% *Error*	SB-MVA Throughput	Simulation Throughput	% *Error*
(1.28, 1.48)	0.27	0.2643	2.15	0.31	0.3054	1.51
(3.00, 2.69)	0.39	0.3902	-0.05	0.35	0.3471	0.83
(2.63, 5.90)	0.25	0.2484	0.63	0.56	0.5601	-0.02
(17.50, 5.50)	0.70	0.6996	0.05	0.20	0.2096	-4.83

From Table 3 we see that for the balanced system with exponential processing times, SB-MVA provides reasonable estimates of throughputs even with fractional customers. We have performed several other experiments with unbalanced systems of different sizes and configurations with similar results [10].

NOTE: To achieve a target throughput, N_p can potentially be less than one. Equation (1) cannot be used where the number of customers is less than one. Hence, we develop an extension to SB-MVA to accommodate the situation when $N_p < 1$. For further details (proof and validation) see Shinde and Suri [11]. The SB-MVA extension developed there is as follows. In the extension, equation (1) is replaced with

$$R_{m,p}(\vec{N}) = \begin{cases} T_{m,p} + \sum_{r \neq p} Q_{m,r}(\vec{N}) T_{m,r} & 0 < N_p \leq 1 \\ T_{m,p} + (\frac{N_p - 1}{N_p}) Q_{m,p}(\vec{N}) T_{m,p} + \sum_{r \neq p} Q_{m,r}(\vec{N}) T_{m,r} & N_p > 1 \end{cases} \tag{6}$$

4 Optimization Model to Estimate Customer Population $\vec{N}$

In this section, we propose a mixed-integer nonlinear optimization model (NOM) to achieve our primary objective, which is to calculate the number of customers $\vec{N}$ to achieve, as closely as possible, the target throughput $\vec{Y} \equiv \{Y_1, Y_2, \ldots, Y_P\}$, for a multi-class closed queueing network. Here Y_p is the target throughput for customer class p. The nonlinear optimization model is as follows.

Objective
Minimize -

$$\sum_{p=1}^{P} \left(Y_p - X_p(\vec{N}) \right)^2$$

subject to:

$$R_{m,p}(\vec{N}) = T_{m,p} + \left(\frac{N_p - 1}{N_p}\right) T_{m,p} Q_{m,p}(\vec{N}) \phi_p + \sum_{r \neq p} Q_{m,r}(\vec{N}) T_{m,r} \tag{7}$$

$$Z\phi_p \geq N_p - 1 \tag{8}$$

$$N_p \geq \phi_p \tag{9}$$

$$X_p(\vec{N}) = \frac{N_p}{\sum_{m=1}^{M} R_{m,p}(\vec{N})} \tag{10}$$

$$Q_{m,p}(\vec{N}) = X_{m,p}(\vec{N}) R_{m,p}(\vec{N}) \tag{11}$$

$Q_{m,p}(\vec{N}), R_{m,p}(\vec{N}), X_p(\vec{N}), N_p, Z \geq 0$ and $\phi_p = 0$ or 1

Here Z is a large number and the variables Z, ϕ_p and equations (8), (9) are used to model the response time constraint from equation (6). The values of N_p generated will be called "the prescribed number of customers".

5 Validating the Optimization Model Using Simulation

In this section we validate our overall approach using some examples. For each example, for a given set of target throughputs, we use the NOM to generate the prescribed number of customers which are typically fractional. We then simulate the system, using Algorithm 1 to achieve the fractional customer levels. In [10], we have tested the accuracy of our model in small and large systems and in systems with variable processing times. In this paper, we present results from a three-station four-class network. The routing matrix of this system is shown in Table 4. An "X" in a customer-station pair denotes that the customers from that class visit the station. Upon finishing processing at a station, a customer chooses either of the stations along its routing with equal probability.

Table 4: Routing table for three station four class system

	Customer Class			
Stations	1	2	3	4
1	X	X	X	
2	X	X		X
3	X		X	X

We present results from two different balanced system settings. In the first setting, the utilization for each of the stations is 0.8 while in the second setting the utilization is 0.9. The NOM is solved using GAMS (www.gams.com) on a Sun Ultra 10 440 MHz workstation. The solution times were 0.09 to 0.11 seconds respectively. For the simulations, we use a warm up period of 20,000 orders. The simulation is run for another 100,000 orders during which statistics are recorded. 10 replications are performed during each simulation run. The results are shown in Tables 5 and 6. The 95% confidence intervals, in these cases too, are within 1% of the average simulation throughput.

Table 5: Target versus Actual Throughputs for three station four class system (Utilization = 0.8)

Customer Class	Target Throughput	Prescribed Customer Populations	Observed Throughputs	% Error
1	0.4	3.60	0.4014	-0.34
2	0.2	1.13	0.2133	-6.33
3	0.2	1.13	0.2188	-9.40
4	0.2	1.13	0.2142	-7.12

Table 6: Target versus Actual Throughputs for three station four class system (Utilization = 0.9)

Customer Class	Target Throughput	Prescribed Customer Populations	Observed Throughputs	% Error
1	0.450	7.99	0.4487	0.29
2	0.225	2.59	0.2427	-7.88
3	0.225	2.59	0.2447	-8.74
4	0.225	2.59	0.2438	-8.37

From Tables 5 and 6 we can see that the prescribed customer populations yield throughputs which are within about 10% of the target throughputs for all the cases. Thus, for the closed queueing networks tested, we conclude that the NOM provides reasonable estimates of customer populations.

6 Conclusion

In this paper, we introduced the concept of fractional number of customers to achieve exact target throughputs in closed queueing networks. We estimated the throughputs using SB-MVA with fractional number of customers. We introduced a control rule to achieve fractional number of customers in closed queueing networks and then compared the SB-MVA predictions with the throughputs observed from simulation. Then, we proposed a nonlinear optimization model to estimate the number of customers required to meet target throughputs. For the examples shown, the prescribed number of customers are able to meet the target throughput within 10%. The optimization model can easily be extended to non-product form systems [10].

References

[1] Bard, Y., 1979, “Some Extensions to Multiclass Queueing Network Analysis”, Performance of Computer Systems, Arato, M. (ed.), North Holland, Amsterdam, The Netherlands.

[2] Kant, K., 1992, “Introduction to Computer System Performance Evaluation”, McGraw-Hill, New York.

[3] Reiser, M., and Lavenberg, S.S., 1980, “Mean-Value Analysis of Closed Multichain Queueing Networks”, Journal of the Association for Computing Machinery, 27(2), 313-323.

[4] Schweitzer, P., 1979, “Approximate Analysis of Multiclass Closed Networks of Queues”, Presented at the International Conference, Stochastic Control and Optimization, Amsterdam, The Netherlands.

[5] Spragins, J., 1980, “Analytical Queueing Models: Guest Editor’s Introduction”, IEEE Computer, 13(4), 175-194.

[6] Srinivasan, M.M., Ebbing, S.J. and Swearingen, A.T., 2003, “Woodward Aircraft Engine Systems Sets Work-in-Process Levels for High-Variety, Low-Volume Products”, Interfaces, 33(4), 61-69.

[7] Suri, R., 1983, “Robustness of Queueing Network Formulas”, Journal of the Association for Computing Machinery, 30(3), 564-594.

[8] Suri, R., 1985, “A Concept of Monotonicity and its Characterization for Closed Queueing Networks”, Operations Research, 33(3), 606-624.

[9] Suri, R., Sanders, J.L., and Kamath, M., 1993, “Performance Evaluation of Production Networks”, in Logistics of Production and Inventory, Graves, S.C., Rinnooy Kan, A.H.G., and Zipkin, P.(eds.), Handbooks in OR & MS, Vol. 4, Elsevier Science Publishers, North-Holland, Amsterdam, The Netherlands, 199-286.

[10] Suri, R., Shinde, R., and Vernon, M., 2005, “Using Mean Value Analysis (MVA) to Set Production Card Levels for CONWIP in a Multiproduct Manufacturing System”, Working Paper, Center for Quick Response Manufacturing, University of Wisconsin-Madison.

[11] Shinde, R., and Suri, R., 2005, “Calculating Throughput when Number of Customers in Closed Queueing Networks is less than one”, Working Paper, Center for Quick Response Manufacturing, University of Wisconsin-Madison.