Proceedings of the 1984
Winter Simulation Conference
S. Sheppard, U. Pooch, D. Pegden (eds.)

CONTINUOUS SIMULATION APPROXIMATIONS TO QUEUEING NETWORKS

Gordon M. Clark
Department of Industrial and Systems Engineering
The Ohio State University
Columbus, Ohio 43210

ABSTRACT

Continuous simulations of queueing networks avoid the statistical sampling problems caused by Monte Carlo operations. Also, one may want to simulate a queueing network as part of a larger continuous simulation. The network considered in this paper is nonstationary, has time-dependent exponential service time, and finite queues which block upstream service if they are at capacity. One approach to simulating the network is to define the entire set of Kolmogorov equations and numerically integrate them. This is impractical because the number of equations is quite large even for modest networks. This paper presents two continuous simulation approximations to the exact solution that reduce the number of equations to be integrated to a manageable number. The independence approximation assumes that the probability of a particular system state is approximated by the product of the probabilities each queueing station assumes a particular state. The partition approximation groups states for each queueing station into subsets and uses a weighting factor applied to joint probabilities calculated under the independence assumption. Results from an illustrative example show that partition approximation is more accurate than the independence assumption; however, the independence approximation does very well.

INTRODUCTION

Two different reasons motivate the use of continuous simulations of queueing networks. The first is that a continuous simulation avoids Monte Carlo sampling and the resultant statistical analysis problems. The potential requirement for tremendous sample sizes in order to estimate mean values accurately is well known [1]. No matter how detailed the discrete-event Monte Carlo simulation, it is inherently a numerical approximation because of the sampling errors introduced by Monte Carlo sampling. Thus, any errors in a continuous simulation approximation to a queueing network may be offset by the sampling errors inherent in Monte Carlo simulation.

The second reason is that the queueing network may be a subsystem in a larger continuous simulation. Imagine a jobshop that is modeled as a network of queues that is part of a larger manufacturing capacity-marketing system dynamics model [2]. Note that the system dynamics models are actually continuous simulations. Another example would be the use of several servers in tandem to handle incoming requests for stock at a warehouse which is part of a larger physical distribution system model that is a continuous simulation.

We can relate the use of a continuous simulation of a queueing network to the concept of a delay used in system dynamics models. Delays represent processes requiring elapsed time before system quantities or entities change state [3]. DYNAMO [2] uses the n^{th} order exponential delay to depict delays. A first-order exponential delay is equivalent to the solution to a queue with a time-dependent Poisson arrival process, an unlimited number of servers, and independent service times from the same negative exponential distribution [4]. Thus, we can think of a first-order exponential delay as a nonstationary $M/M/\infty$ queue. An n^{th} order exponential delay describes a queue with an unlimited number of identical servers having n^{th} order Erlang-distributed service times, i.e., a nonstationary $M/E_n/\infty$ queue. Clark [5] defined a queueing delay having a time-dependent Poisson arrival process, s servers, negative exponential service times, and an unlimited waiting line, i.e., a nonstationary M/M/s queue. Moreover, Clark showed how this queueing delay could be approximated efficiently in a larger continuous simulation. This paper considers queueing delays made up of a network of queues with finite capacities for their waiting space.

This paper presents two continuous simulation approximations to a queueing network as defined below. The approximations are capable of computing such performance measures as the mean and variance of the number of entities at each station in the network, the expected output rate at each station, and the average waiting time during the entire simulation time period at each station.

System Represented

The system represented consists of I stations, each one of which has s_i servers and a queue permitting at most $m_i - p_i$ entities where i=1,2,...,I. Thus the maximum number of entities permitted at station i is m_i including the entities being served. A station at capacity blocks service for any entity in service destined for the station at capacity. The probability an entity leaving station i goes next to station j is τ_{ij}. Note that this probability is independent of the path taken by the entity to reach station i. The probability that an entity leaving station i departs the system is

τ_{io}, and τ_{io} is equal to $1-\sum_{j=1}^{I} \tau_{ij}$.

External arrivals can occur to any station and their arrival processes are time-dependent Poisson processes with rate functions $\lambda_i(t)$ for i=1,2,...I. If an external arrival occurs to a station at capacity, the arrival leaves the system immediately. Individual servers have time-dependent

negative exponential service time distributions with $\mu_i(t)$ being the individual server service rate at station i.

The above description specifies a Jackson network [6] with the exceptions that the system is nonstationary, the queues have finite capacity, and queues at capacity block service at upstream stations. This means that the probability of a particular system state is a more complex function than simply the product of the probabilities that the individual stations are at specified states.

KOLMOGOROV EQUATIONS

The system as defined above is a Markov process; thus, we can use the Kolmogorov forward equations which are simultaneous differential equations involving the probabilities of being in each individual system state. When integrated numerically, the solution provides us with a means for calculating time trajectories for our desired output performance measures. In fact, Pritsker [7] suggested that we can simulate Markov queueing systems using Kolmogorov equations.

We can formulate these Kolmogorov equations using the procedure described by Kleinrock [8]. His procedure is readily extended to portray time-dependent arrival rates and service rates by substituting $\lambda_i(t)$ and $\mu_i(t)$ for the constant coefficients λ_i and μ_i.

We define the following notation in order to specify the Kolmogorov equations. Let

n_i = number of entities at station i

$\bar{n}$ = $(n_1, n_2, \ldots, n_I)$

n_o = 1

e_i = $(n_1=0, n_2=0, \ldots, n_{i-1}=0, n_i=1, n_{i+1}=0, \ldots, n_I=0)$

$$g_i(n) = \begin{cases} 1 \text{ if } 0<n<m_i \\ 0 \text{ if otherwise} \end{cases}$$

$g_o(n)$ = 1 for all n

$$h_i(n) = \begin{cases} \mu_i s_i \text{ if } m_i>n>s_i \\ \mu_i n \text{ if } p_i>n>1 \\ 0 \text{ if } n>m_i \text{ or } n<0 \end{cases}$$

$p_r(\bar{n},t)$ = probability the system is in the state $\bar{n}$ at time t.

The derivative of the probability of being in the state $\bar{n}$ at time t is the sum of three quantities or

$$p_r'(\bar{n},t) = -r_1(\bar{n},t)+r_2(\bar{n},t)+r_3(\bar{n},t), \qquad (1)$$

where $r_1(\bar{n},t)$ = instantaneous rate at time t the system leaves the state $\bar{n}$

$r_2(\bar{n},t)$ = instantaneous rate at time t the system enters the state $\bar{n}$ due to an external arrival

$r_3(\bar{n},t)$ = instantaneous rate at time t the system enters the state $\bar{n}$ due to a completed service.

We can write equation (1) because the chance of two or more events in a Markov process is negligible [8]. To calculate $r_2(\bar{n},t)$ we simply sum the expected arrival rates to each station that place us in the state $\bar{n}$ by one arrival at time t. That is,

$$r_2(\bar{n},t) = \sum_{i=1}^{I} \lambda_i(t) g_i(n_i-1) p_r(\bar{n}-e_i,t) \qquad (2)$$

We calculate $r_3(\bar{n},t)$ by summing the expected rates placing the system in the state $\bar{n}$ at time t by a completed service. Thus,

$$r_3(\bar{n},t) = \sum_{i=1}^{I} h_i(n_i+1) \sum_{\substack{j=0 \\ j\neq i}}^{I} \tau_{ij} p_r(\bar{n}+e_i-e_j,t) g_j(n_j-1) \qquad (3)$$

Finally, the instantaneous rate the system leaves the state $\bar{n}$ involves the sum of the expected rates of having an arrival or a service given the system is in the state $\bar{n}$ at time t. That is,

$$r_1(\bar{n},t) = p_r(\bar{n},t) \left[\sum_{i=1}^{I} \lambda_i g_i(n_i) + \sum_{i=1}^{I} h_i(n_i) \sum_{\substack{j=0 \\ j\neq i}}^{I} \tau_{ij} g_j(n_j) \right] \qquad (4)$$

By numerically integrating (1) we can simulate the queueing network and the only errors involved would be those due to numerical integration. This is only practical for very small networks since a Kolmogorov equation is required for each possible system state. If E_q is the total number of equations integrated, then E_q is given by

$$E_q = \prod_{i=1}^{I} (m_i+1), \qquad (5)$$

which grows very quickly with both I and m_i. For example, if $m_1=m_2=m_3=50$ and I=3, than E_q is 132,651.

INDEPENDENCE APPROXIMATION

The independence approximation is one way of reducing the number of equations integrated to a manageable number. This approximation assumes that the joint probability of the system being in a particular state is the product of the individual probabilities each station is in a particular state. That is,

$$p_r(\bar{n},t) = \prod_{i=1}^{I} p_r(N_i=n_i,t), \quad (6)$$

where $p_r(N_i=n_i,t)$ = probability station i has N_i entities at time t.

The Jackson network suggests this approximation because (6) would be exact if it were not for the nonstationarity, the finite capacity of the queues, and the blocking.

To reduce the number of equations integrated, we derive differential equations giving the marginal probabilities of the system state at one station by summing over all possible states at the other stations. That is,

$$p_r'(N_i=u,t) = \sum_{\substack{n \\ n_i=u}} p_r'(\bar{n},t). \quad (7)$$

After evaluating (7) we are left with

$$E_q = \sum_{i=1}^{I} (m_i+1) \quad (8)$$

equations that characterize the network. To illustrate the marked reduction, the independence approximation reduces the 132,651 equations for a three station capacity 50 network to 153 equations.

In integrating (7) we use the approximation given by (6) in order to evaluate probabilities involving more than one station. An interesting observaiton about (7) is that the probabilities involving multiple stations involve no more than two stations, and the joint probabilities appearing on the right hand side of (7) are those of the form $p_r(N_j=u,N_k=v,t)$ where u is restricted to $0,1,2,\ldots,s_j-1$ and m_j.

PARTITION APPROXIMATION

We defined an approximation called the partition approximation in order to improve on the independence approximation and yet have substantially less equations to integrate simultaneously than given by (5). The partition approximation approximates joint probabilities of the form $p_r(N_i=u,N_j=v,t)$, where

$$u = 0,1,2,\ldots,m_i$$

$$v = 0,1,2,\ldots,m_j$$

This approximation requires one to partition the state space for the number of entities at a station into mutually exclusive subsets, at station i let $\alpha_{i,k}$ be the k^{th} such partition where

$$\alpha_{i,k} = \left\{ l_{i,k} \leq u \leq U_{i,k} \right\} \quad k=1,2,\ldots,K_i.$$

Then $l_{i,k}$ is the lower limit on state variables for the k^{th} partition and $u_{i,k}$ is upper limit. We also require these partitions to be collectively exhaustive and mutually exclusive. Thus

$$\bigcup_{k=1}^{K_i} \alpha_{i,k} = \left\{0,1,2,\ldots,m_i\right\}$$

$$\alpha_{i,k} \cap \alpha_{i,s} = \phi \text{ for } k \neq s$$

To calculate approximations to joint probabilities we need to define two probabilities. Let Q(i,k,t) be the marginal probability that N_i is in the k^{th} partition at time t and R(i,k,j,o,t) be the joint probability that N_i is in its k^{th} partition and N_j is in its o^{th} partition at time t. That is,

$$Q(i,k,t) = p_r(N_i \varepsilon \alpha_{i,k},t)$$

$$R(i,k,j,o,t) = p_r(N_i \varepsilon \alpha_{i,k}, N_j \varepsilon \alpha_{j,o},t)$$

Also let $K_i(u)$ = the partition containing the state u for N_i. That is,

$$u \varepsilon \alpha_{i,k_i(u)}$$

The partition approximation is

$$p_r(N_i=u,N_j=v,t) =$$

$$\frac{p_r(N_i=u,t)R(i,k_i(u),j,k_j(v),t)p_r(N_j=v,t)}{Q(i,K_i(u),t)Q(j,K_j(v),t)} \quad (9)$$

Note that the partition approximation is equal to the independence approximation modified by the weight

$$\frac{R(i,K_i(u),j,K_j(v),t)}{Q(i,K_i(u),t)Q(j,K_j(v),t)} .$$

One can prove that the partition approximation satisfies the axioms of probability in that the sum of all probabilities for the bivariate distribution equals 1.

Based upon an analysis of several cases involving a two station network, we selected the following partitions. In this case, the partition is an ordered pair for station (i,j) instead of being the same for each station. First station i has a single partition where $N_i \leq s_j-1$. Next, station j where $j \neq i$, has six subsets where

$$\alpha_{j,1} = \{0\}$$

$$\alpha_{j,2} = \{1,2,3\}$$

$$\alpha_{j,3} = \{4 \leq N_j < 8\}$$

$$\alpha_{j,4} = \{9 \leq N_j \leq 15\}$$

$$\alpha_{j,5} = \{16 \leq N_j \leq 24\}$$

$$\alpha_{j,6} = \{n_j \geq 25\}$$

Of course, station j does not require all six subsets if $m_j<25$. The implementation only used the approximation for probabilities of the form $p_r(N_i=u,N_j=v,t)$ where $u \leq s_i-1$. All other probabilities involving multiple stations were approximated using (6). Using six subsets for station j, then the approximation requires five joint probabilities for each (i,j) pair, $i \neq j$, of the form

$$p_r(N_i \leq s_i-1, N_j \varepsilon \alpha_{j,k},t)$$

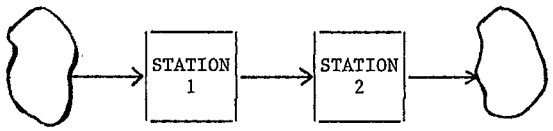
The sixth joint probability can be calculated from $P_r(N_i \leq s_i-1,t)$ and the other five. Thus, this version of the partition approximation requires E_q equations to be integrated where

$$E_q = \sum_{i=1}^{I} (m_i+1)+5 \cdot I(I-1) \qquad (10)$$

Then, if I=3 and each station has a capacity 50, E_q is 183 which is considerably less than the 132,651 equations for all the Kolmogorov equations and only moderately greater than the 153 for the independence approximation.

TEST CASE

To illustrate the errors and computational speed for the two approximations, we simulated a simple two station tandem queue using a continuous simulation.



Each station had a capacity of ten entities, a single server, and the arrival rate to the first station was a sinusoidal function given by

$$\lambda(t) = 1.0 + .5 \sin (2\pi t/60)$$

so the mean arrival rate was one entity per unit time, the amplitude was one-half the mean and the period was 1.0 time units.

We represented two cases to simulate both heavy and light traffic conditions. Both servers had the same service rate which was 2 services per unit time for Case 1 and 1 service per unit time for Case 2. This gave a mean utilization of .5 for Case 1 and 1.0 for Case 2.

First we simulated the tandem network by integrating all 121 Kolmogorov equations. Then we simulated using the two approximations employing 22 equations for the independence approximation and 28 equations for the partition approximation. Each run started in the empty and idle condition, and values of $E(N_i,t)$, $V(N_i,t)$ were sampled and compared every integral value of time until time 240. Using the Kolmogorov equation result as a standard, the performance measure was the error percent range over the 240 comparisons. That is,

error percent range = maximum percent error − minimum percent error

The following table presents the results.

Error % Range-Independence Approximation

Case	$E(N_1,t)$	$E(N_2,t)$	$V(N_1,t)$	$V(N_2,t)$
1	1.78	10.27	3.32	29.3
2	2.90	3.26	7.63	16.235

Error % Range-Partition Approximation

Case	$E(N_1,t)$	$E(N_2,t)$	$V(N_1,t)$	$V(N_2,t)$
1	1.47	5.22	2.93	15.55
2	2.10	1.40	6.85	8.28

CONCLUSIONS

The results show that the independence approximation does surprisingly well. The maximum percent error range in expected entities is 10.3 percent. However, the partition approximation does even better where the maximum percent error range is 5.2%. Note that both approximations have significantly more error when approximating the variance.

The CPU times are approximately proportional to the number of equations integrated. Cases 1 and 2 took 215.0 and 140.8 records, respectively, on an AMDAHL 470.V8 computer for the Kolmogorov equations; however, these two cases only required 44.3 and 25.0 seconds, respectively, for the partition approximation. Of course, larger cases quickly require so many equations that the Kolmogorov equation approach is impractical, and the only alternative is one of the approximations.

REFERENCES

1. Law, A.M., and J.S. Carson, "A Sequential Procedure for Determining the Length of a Steady-State Simulation," Operations Research, Vol. 27, No. 5, pp. 1011-1025, September-October, 1979.

2. Forrester, J.W., Principles of Systems, Wright Allen Press, Cambridge, Mass., 1971.

3. Forrester, J.W., Industrial Dynamics, M.I.T. Press, Cambridge, Mass., 1961.

4. Gross, D., and C. M. Harris, Fundamentals of Queueing Theory, Wiley, New York, 1974, pp. 114-117.

5. Clark, G.M., "Use of Polya Distributions in Approximate Solutions to Nonstationary M/M/s Queues," Communications of the ACM, Vol. 24, No. 4, April 1981, pp. 206-217.

6. Jackson, J.R., "Networks of Waiting Lines," Operations Research, Vol. 15, pp. 518-521, 1957.

7. Pritsker, A. Alan B., "Three Simulation Approaches to Queueing Studies Using GASP IV," Computers and Industrial Engineering, Vol. 1, pp. 57-65, 1976.

8. Kleinrock, L., Queueing Systems Volume 1: Theory, Wiley, New York, 1975.