The stationary state arrival instant probability distribution of closed queueing networks

Elizabeth Varki

Abstract

Closed queueing networks are hard to evaluate since the performance of each queue in the network is dependent on the state of the entire network. The only class of closed networks whose arrival instant distribution has been evaluated is the product form networks. The class of networks analyzed in this paper is more general than the class of product form networks. The queues in the network considered here could be single or parallel fork-join queues. The service time distribution of queues could be dependent on the state of other queues in the network. This paper shows the relationship between the random variable representing a queue's arrival state seen just prior to arrival at the queue and the random variable representing the queue's arrival state seen just prior to arrival at any of the queues within the network. Using this relationship, the paper derives a bound on the mean arrival instant queue length.

Index Terms: arrival instant probability, queue arrival/departure processes, closed queueing networks, embedded Markov chains.

1 Introduction

Queueing networks refer to systems of queues whose input and output processes are related since jobs departing from one queueing server may join the input queue of another server. The layout of a queueing network represents the path taken by jobs as they move from server to server within the network. Queueing networks are used to model real systems that consist of several components that interact with each other. Each of the key components is modeled by a queueing server and the layout of the queueing network models the interaction between the components.

A closed queueing network models systems where there are a fixed number of jobs circulating in the network. There are no job arrivals from outside the network, nor are there job departures from the network. Such networks are used to model computer/communication systems and manufacturing facilities. The stationary state properties of closed queueing networks are difficult to evaluate since the performance of each queue is dependent on the entire state of the network. Further, the arrival/departure process in a closed network is not Poisson [2] nor renewable [3] even when each of the queueing servers has service drawn from the exponential distribution. The only class of closed

queueing networks that is easily solvable is the product form networks [1, 5] since this class of networks has the nice property that the arrival instant distribution for a network with M jobs is identical to the steady state distribution for the network with $M - 1$ jobs [7, 9].

Product form networks are the only type of closed networks that have been extensively studied. As a result, several techniques to quickly compute the mean performance measures of closed product form networks without having to compute the stationary state distribution of the network have been developed. Unfortunately, many current applications cannot be modeled using product form queueing networks. For example, product form networks cannot be used to model parallelism, but most computer/communication systems and manufacturing facilities have components that display parallelism. Thus, solution techniques for non product form networks are required, and in order to develop these techniques one has to study the properties of the underlying queueing network. Understanding properties of the stationary state arrival instant distribution is a logical place to start since the response time distribution can be computed from the arrival instant distribution. Once the response time is known, both the throughput and the queue length distribution of the queueing network can be calculated.

In this paper, we analyze the arrival instant distribution of a class of closed queueing networks that is more general than product form networks. The networks analyzed here could contain parallel (fork-join) queues. Also, the service time at a queue could be dependent on the number of jobs in other queues of the network. There are no prior studies on the arrival instant distribution of this class of networks. In this paper, we show that the random variable representing a queue's arrival instant state observed just prior to arrival at the queue is less than or equal to the random variable representing the queue's arrival instant state observed just prior to arrival at any of the queues in the network. We then use this relationship to relate the mean arrival instant queue length at all the queues in the network to the number of jobs, M, in the closed network.

The remainder of this paper is organized as follows. Section 2 formalizes the class of queueing networks analyzed in this paper. Section 3 then presents the analysis and derives the arrival instant queue length bound. The conclusions are presented in Section 4.

2 Queueing Network Formalization

The queues in the network are arbitrarily assigned a unique number from 1 to K. That is,

$K \doteq$[1] the number of queues in a queueing network.

$M \doteq$ the number of jobs circulating in a closed network.

[1]The symbol $\doteq$ means equal by definition

It is assumed that the M jobs are statistically identical. The number of jobs in the various queues of the network at any point in time is given by the vector $\vec{M} = (m_1, m_2, \cdots, m_K)$ where m_k represents the number of jobs at queue k, and the m_k are such that M= $m_1 + m_2 + m_K$.

The jobs in the network continuously cycle through the network moving from queue to queue. Once a job completes service at queue k, the job moves to queue j ($j = 1, ..$K), depending on the routing probability r_{kj}, where $0 \leq r_{kj} \leq 1$ and $\sum_{j=1}^{j=\mathsf{K}} r_{kj} = 1$ $(k = 1, \cdots, \mathsf{K})$. The routing probabilities are independent of the state of the network and are an input to the modeling study.

The service time at a queue is assumed to be drawn from the exponential distribution. Other than the exponential distribution, no other assumptions are made about a queue's service time. The service time at a queue could be dependent on the number of jobs present at the queue or the number present at other queues in the network. That is, a queue's service time could be network state dependent.

A queue could be a single queue or a parallel fork-join queue where an arriving job divides into tasks. A job at a fork-join queue completes service and departs from the queue only after all its tasks complete service. At a fork-join queue, a job is represented by its tasks. The total number of jobs in the network, M remains unchanged. That is, at a point in time, if a fork-join queue has 10 tasks from 3 jobs, the number of jobs in the fork-join queue is 3 (not 10).

The network dependent service time assumption and the allowing of parallel queues in the queueing network makes this class of network far more general than product-form networks. There are no prior studies on the arrival instant distribution of this class of networks.

3 Arrival State Distribution

Suppose the queueing network is observed from some time instant, T. Let

c_n represent the n^{th} arrival to queue k.

Define random variable

$a_{k,n} \doteq$ the number of jobs seen in queue k by c_n prior to arrival.

Since the service time is assumed to be exponential, $a_{k,n}$ defines the arrival state of queue k just prior to arrival of c_n. (Note that the arriving job does not see itself in the network since this job is in transit.) Since an arriving job can see at most M -1 jobs ahead of it, $0 \leq a_{k,n} \leq$ M $- 1, \forall n$. The sequence of random variables $a_{k,1}, a_{k,2}, a_{k,3}, \cdots$ forms a discrete-time random process.

Now, instead of observing the state of queue k prior to arrivals only at queue k, suppose the state of queue k is observed prior to arrivals at all K queues in the network. Since the network is closed,

there is a departure corresponding to every arrival. So, the state of queue k is observed every time there is a transition from one queue to another. Let

C_n represent the n^{th} arrival (which can occur to any of the K queues) in the network,

Define random variable

$y_{k,n} \doteq$ the number of jobs seen in queue k by C_n.

So, $y_{k,n}$ gives the state of queue k before every arrival (or alternatively, after every departure). The sequence of random variables $y_{k,1}, y_{k,2}, y_{k,3}, \cdots$ forms a discrete-time random process, where $0 \leq y_{k,n} \leq \mathsf{M} - 1 \quad \forall n$.

The random process $\{y_{k,n}\}$ gives the state of queue k prior to all arrivals, while the random process $\{a_{k,n}\}$ gives the state of queue k prior to arrivals only to queue k. Now some of the C_n are also $c_{n'}$, since some of the arrivals occur to queue k. Therefore, the random process $\{a_{k,n}\}$ is embedded in the random process $\{y_{k,n}\}$ as follows:

$$y_{k,1}, y_{k,2}, \cdots, y_{k,n_1} = a_{k,1},\ y_{k,n_1+1}, y_{k,n_1+2}, \cdots, y_{k,n_2} = a_{k,2},\ \cdots$$

The random process $\{y_{k,n}\}$ consists of subsequences of the form

$$y_{k,1'}, y_{k,2'}, \cdots, y_{k,n_j-1}, y_{k,n_j} = a_{k,j} \qquad n_j \geq 1'.$$

The $y_{k,1'}, y_{k,2'}, \cdots, y_{k,n_j-1}$ in each subsequence give the state of queue k prior to when a C_n arrives at queue x, $x \neq k$. Only y_{k,n_j} gives the state of queue k prior to when a c_j arrives at queue k. That is, all arrivals, but for the last, do not increase the number of jobs in queue k. The only arrival that increases the number of jobs in queue k is the last one in the subsequence. But some of the arrivals C_n could be the result of departures from queue k. That is, a job arriving at queue x ($x \neq k$) may have departed from queue k, decreasing the number of jobs in queue k. Thus, the random variables in a subsequence are related to each other in the following manner:

$$y_{k,1'} \geq y_{k,2'} \geq \cdots, \geq y_{k,n_j-1} \geq y_{k,n_j} = a_{k,j}$$

Define random variable

$d_{k,n+1} \doteq$ number of departures (after service) from queue k between the arrivals of C_n and C_{n+1}.

$$d_{k,n+1} = \begin{cases} 0 & : \text{ if } C_{n+1} \text{ did not depart from queue } k \\ 1 & : \text{ if } C_{n+1} \text{ departed from queue } k \end{cases}$$

For every subsequence of the random process $\{y_{k,n}\}$, the random variable $y_{k,n+1}$ is related to the random variable $y_{k,n}$ by:

$$y_{k,n+1} = y_{k,n} - d_{k,n+1}, \qquad \text{when } y_{k,n+1} \text{ and } y_{k,n} \text{ belong to the same subsequence.} \tag{1}$$

When the entire random process $\{y_{k,n}\}$ is considered, some of the $y_{k,n}$ are also $a_{k,n'}$. The random variable $y_{k,n+1}$ is related to the random variable $y_{k,n}$ by:

$$y_{k,n+1} = y_{k,n} - d_{k,n+1} + \triangle_{k,n} \tag{2}$$

where

$$\triangle_{k,n} = \begin{cases} 0 & : \text{ if } C_n \neq c_{n'} \\ 1 & : \text{ if } C_n = c_{n'} \end{cases}$$

That is, $\triangle_{k,n} = 1$ if $y_{k,n}$ =$a_{k,n'}$.

Assume that the random processes $\{y_{k,n}\}$ and $\{a_{k,n}\}$ are ergodic. That is, assume that the random processes $\{y_{k,n}\}$ and $\{a_{k,n}\}$ have limiting distributions that are independent of the initial state. (We will consider the ergodicity of $\{y_{k,n}\}$ and $\{a_{k,n}\}$ later in this section.) In this case, define limiting random variables:

$$\tilde{y}_k = \lim_{n\to\infty} y_{k,n}$$

$$\tilde{a}_k = \lim_{n\to\infty} a_{k,n}$$

Theorem 3.1

$$\tilde{a}_k \leq \tilde{y}_k$$

Proof: From Equation 1, it follows that

$$y_{k,n} \geq a_{k,n'}$$

when $y_{k,n}$ and $a_{k,n'}$ belong to the same subsequence. This relationship is true for all subsequences of the random process $\{y_{k,n}\}$.

If it is assumed that the random processes $\{a_{k,n}\}$ and $\{y_{k,n}\}$ are ergodic, then the theorem holds.

□

Theorem 3.1 can be used to derive a bound on the mean arrival instant queue length at the various queues of a closed network. Let

$$A_k = \mathsf{E}[\tilde{a}_k]$$

$$Y_k = \mathsf{E}[\tilde{y}_k]$$

From Theorem 3.1, the following result follows [8]:

Corollary 3.1

$$A_k \leq Y_k$$

The next theorem uses Corollary 3.1 to compute A_k in terms of M.

Theorem 3.2

$$\sum_{k=1}^{\mathsf{K}} A_k \leq \mathsf{M} - 1$$

Proof: Since there are M jobs in the closed network, C_n sees M -1 jobs at the K queues just prior to arrival. That is,

$$\sum_{k=1}^{\mathsf{K}} y_{k,n} = \mathsf{M} - 1 \quad \forall n$$

If it is assumed that the random processes $\{y_{k,n}\}$, $k = 1, \cdots, \mathsf{K}$, are ergodic, then

$$\sum_{k=1}^{\mathsf{K}} \tilde{y}_k \ = \mathsf{M} \ - 1$$

Hence,

$$\sum_{k=1}^{\mathsf{K}} Y_k \ = \mathsf{M} \ - 1$$

From Corollary 3.1, it follows that

$$\sum_{k=1}^{\mathsf{K}} A_k \ \leq \mathsf{M} \ - 1$$

□

For closed product-form networks, the distribution of $\tilde{a}_k$ equals the steady state distribution of queue k when there is one less job in the network [7, 9]. Thus, for closed product-form networks, $A_k = \mathsf{M} \ - 1$.

We now address the issue of ergodicity of random process $\{y_{k,n}\}$. Since the network is closed, the random process $\{y_{k,n}\}$ is dependent on the random processes $\{y_{j,n}\}$ $\forall j = 1, \cdots, k-1, k+1, \cdots, \mathsf{K}$. So, instead of observing the state of only a single queue k at each transition, the state of the entire network is observed at each transition. Let

$\vec{y_n} = (y_{1,n}, \cdots, y^*_{j,n}, \cdots, y_{\mathsf{K},n}) \doteq$ the arrival state seen by C_n prior to arrival at queue j ($\forall j = 1, \cdots, \mathsf{K}$).

The $y_{k,n}$ are such that $\sum_{k=1}^{\mathsf{K}} y_{k,n} = \mathsf{M} \ - 1$.

Result 3.1 *The random process $\{\vec{y_n}\}$ forms a Markov chain.*

Proof: From Equation 2, $y_{k,n+1}$ is dependent on $y_{k,n}$ and $d_{k,n+1}$ $\forall k = 1, \cdots, \mathsf{K}$.

$d_{k,n+1}$ is dependent on whether C_{n+1} departed from queue k or from one of the other $\mathsf{K} \ - 1$ queues.

Since service times at all the K queues are drawn from the exponential distribution (memory-less), the value of $d_{k,n+1}$ $\forall k = 1, \cdots, \mathsf{K}$, is completely dependent on $\vec{y_n}$.

Thus, the state $\vec{y_{n+1}} = (y_{1,n+1}, \cdots, y^*_{i,n+1}, \cdots, y_{\mathsf{K},n+1})$ is completely dependent on the state $\vec{y_n}$ and is independent of past history.

□

The random process $\{\vec{y_n}\}$ corresponds to the embedded Markov chain formed by observing the network prior to every arrival transition in the network. Now, consider the random process $\{a_{k,n}\}$. Since the network is closed, the value of $a_{k,n} = y_{k,n'}$ is such that $\sum_{k=1}^{\mathsf{K}} y_{k,n'} = \mathsf{M} - 1$. So, the state of $a_{k,n}$ is dependent on the states of $y_{j,n'}$ $(j = 1, \cdots, k-1, k+1, \cdots, \mathsf{K})$. Therefore, the state of the entire network is considered prior to arrivals at queue k. Let

$y'_{j,n} \doteq$ the arrival state of queue j prior to arrival of c_n at queue k.

$\vec{a_{k,n}} = (y'_{1,n}, \cdots, y'_{k-1,n}, y'_{k,n} = a_{k,n}, y'_{k+1,n}, \cdots, y'_{\mathsf{K},n'}) \doteq$ the arrival state of the network prior to arrival of c_n at queue k.

The $y'_{j,n}$ are such that $\sum_{j=1}^{\mathsf{K}} y'_{j,n} = \mathsf{M} - 1$.

The random process $\{\vec{a_{k,n}}\}$ records the state of the network prior to arrivals at queue k while the random process $\{\vec{y_n}\}$ records the state of the network prior to all arrival transitions in the network.

Result 3.2 *The random processes* $\{\vec{a_{k,n}}\}$ $\forall k = 1, \cdots, \mathsf{K}$, *form Markov chains.*

Proof: The random variable $y'_{j,n+1}$ is related to the random variable $y'_{j,n}$ by:

$$y'_{j,n+1} = y'_{j,n} - d'_{j,n+1} + x'_{j,n+1} \quad \forall j = 1, \cdots, \mathsf{K} \tag{3}$$

where

$d'_{j,n+1} \doteq$ the number of departures from queue j between arrivals c_n and c_{n+1} at queue k, and

$x'_{j,n+1} \doteq$ the number of arrivals at queue j between arrivals c_n and c_{n+1} at queue k.

Note that $x'_{j,n+1} = 1$ when $j = k$.

Since the service time at all queues are exponential and the network is closed, the values of $d'_{j,n+1}$ and $x'_{j,n+1}$ $\forall j = 1, \cdots, \mathsf{K}$ are completely dependent on the value of $\vec{a_{k,n}}$.

It follows from Equation 3 that the random process $\{\vec{a_{k,n}}\}$ forms a discrete Markov chain.

□

The random process $\{\vec{a_{k,n}}\}$ corresponds to the embedded Markov chain formed by observing the state of the network prior to every arrival at queue k. The necessary and sufficient conditions for

ergodicity of Markov chains are well known [4, 6]. The networks formalized in this paper satisfy these conditions, so the Markov chains $\{\vec{y_n}\}$ and $\{\vec{a_{k,n}}\}$ have a limiting distribution. The limiting distribution for $\tilde{a}_k$ $(k = 1, \cdots, \mathsf{K})$ can be calculated from the limiting distribution for $\{\vec{a_{k,n}}\}$. Note that the assumption of exponential service times has only been used to prove the existence of the limiting distributions $\tilde{a}_k$ $(k = 1, \cdots, \mathsf{K})$. Theorems 3.1 and 3.2 hold for all closed networks whose queue arrival instant limiting distributions exist.

4 Conclusions

The stationary state arrival instant distribution of a queueing system provides insight into the behavior of the queueing system. The arrival instant distribution is a starting point into computing performance metrics such as queue response time, queue throughput, and queue length. Unfortunately, it is not easy to compute the arrival instant distribution for closed queueing networks since the arrival state at a queue is dependent on the state of the entire network. Product form networks are the only class of closed networks whose arrival instant distribution can be easily computed. The contribution of this paper is that we have presented a bound for the mean arrival instant queue length of a general class of closed networks. The tightest prior bound for general closed queueing networks is much looser ($A_k \leq \mathsf{K} \times \mathsf{M}$) than the bound generated here. Theorem 3.2 has immediate applications for closed fork-join queueing networks since the mean performance measures of this class of networks can now be bounded.

The results derived in this paper hold for closed networks with parallel and non-parallel queues. The queue service times are exponential and can be network state dependent. The assumption of exponential service times ensures that the arrival instant distribution is ergodic. However, the analysis and results derived in the paper are valid for all queueing systems where the arrival instant distribution is ergodic, regardless of the service time distribution.

While deriving the mean arrival queue length bound, this paper has shown the relationship between the random variables representing a queue's arrival instant state just prior to an arrival at the queue and an arrival at any of the queues in the network. More research in the embedded Markov chains $\{\vec{y_n}\}$ and $\{\vec{a_{k,n}}\}$ could provide insights into the nature of input/output processes in closed queueing networks, and this could lead to performance evaluation techniques for this general class of queueing networks.

References

[1] Baskett, F., Chandy, K.M., Muntz, R.R., Palacios, F.G. "Open, closed, and mixed networks of queues with different classes of customers", Journal of the Association of Computing Machinery, Vol. 22, No. 2, April 1975, pp. 248 – 260.

[2] Burke, P.J. "Proof of a conjecture on the interarrival time distribution in an M/M/1 queue with feedback", IEEE Transactions on Computers, Vol. 24, No. 5, May 1976, pp. 575 – 576.

[3] Disney, R. L. "Networks of Queues", In *Encyclopedia of Operations Research & Management Science*, S. I. Gass and C. M. Harris (Eds.). Boston: Kluwer Academic, 1996.

[4] Gross, D., Harris, C.M., *Fundamentals of Queueing Theory*, Wiley-Interscience, 3rd edition, 1998.

[5] Kelly, F. P. "Networks of queues", Advances in Applied Probability, 8, 1976, pp. 416 – 432.

[6] Kleinrock, L., *Queueing Systems, Volume I*, Wiley-Interscience, 1st edition, 1975.

[7] Lavenberg, S.S., Reiser, M. "Stationary state probabilities at arrival instants for closed queueing networks with multiple types of customers", Journal of Applied Probability, 17, 1980, pp. 1048 – 1061.

[8] Ross, S.M., *A first course in probability - 6th ed.*, Prentice Hall, Inc., 2002.

[9] Sevcik, K.C., Mitrani, I. "The distribution of queueing network states at input and output instants", Journal of the ACM, 28, 2, April 1981, pp. 358 – 371.