The A* traffic process in a queue with feedback

Elizabeth Varki

Abstract

This paper establishes the A* traffic process in queueing networks. The A* traffic process links open and closed networks, and explains why a queue with instantaneous Bernoulli feedback has identical limiting queue length distribution as an ordinary M/M/1 queue (with no feedback). We show that the A* embedded queue length process is a Markov chain. The similarities and differences between the feedback queue and the ordinary queue are encapsulated in the A* process. This traffic process can be used to understand open/closed product form and non-product form networks, and this could lead to better performance evaluation tools for non-product form networks.

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

1 Introduction

The M/M/1 queue with instantaneous Bernoulli feedback, shown in Figure 1, embodies the mystique of product-form networks - the limiting queue length process of a queue with feedback has the same distribution as the ordinary M/M/1 queue (without feedback), even though the two queue systems have completely different traffic processes. This queue was first studied by Takàcs [9]. External arrivals are drawn from a Poisson distribution with parameter λ. The service time at the queue is drawn from an exponential distribution with rate μ. A job after completing execution instantaneously departs the network with probability q or is fed back to the queue with probability $p = 1 - q$. Thus, the input process to the queue consists of new arrivals from outside and returning jobs. The Chapman-Kolmogorov equations (and therefore, the steady state flow balance equations) of this network are identical to those of the ordinary M/M/1 queue with arrival rate λ and service rate $q\mu$, as depicted in Figure 2. When $\lambda < q\mu$, the limiting queue length distribution for the network exists. However, the feedback queue and the ordinary M/M/1 queue have very different traffic processes. It has been shown that the input, output, and feedback processes of a feedback queue are not Poisson. Only the external arrivals and the external departures from the queue are Poisson processes [2, 3].

The feedback queue is an example of a product-form network (*i.e.,* Jackson network). Product-form networks have been extensively analyzed, but it is still not entirely understood why some networks have the product-form property. The reason the feedback queue has been extensively studied is that this queue is the simplest of the product-form networks that has non-Poisson traffic flows, but this queue behaves like an ordinary M/M/1 queue with Poisson traffic flows. In this paper, we establish a traffic process that incorporates the similarities and differences between the two queues.

The feedback queue and the ordinary M/M/1 queue are single queue networks. Here, we treat the external Poisson source of arrivals as an "external" queue with infinite number of jobs. The feedback network now consists of two queues - queue 1 is the feedback queue and queue 2 is the external queue as shown in Figure 3. The external queue always has an infinite number of jobs. The state of the network is the number of jobs in the two queues of the network. Since the external queue always has an infinite number of jobs, the Chapman-Kolmogorov equations of the network are unchanged.

Observe the state of the network at every transition; this embedded stochastic process is a Markov chain, and the properties of this chain are well known [1, 4]. Each transition in the network occurs as a result of a job completing service at a queue and departing the queue. This departing job then instantaneously arrives at a queue in the network. Thus, every transition is the result of a job departing a queue and instantaneously arriving at a queue. We observe the state of the stochastic process at every transition, but we view the network from the perspective of the job participating in the transition. The job participating in the transition is not included in the state description. Every transition consists of a queue departure and a queue arrival. For every transition, the state of the network can be viewed just after departure and just prior to arrival. In this paper, we observe the state of the network just prior to every arrival at a queue in the network. The resulting embedded stochastic process is called the Arrival* process, or in short, the A* process.

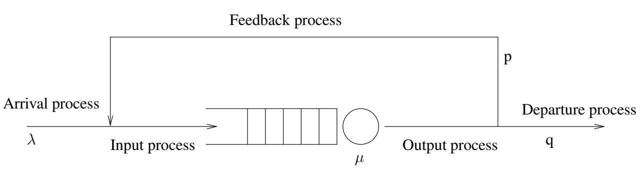

Figure 1: The M/M/1 queue with Bernoulli feedback

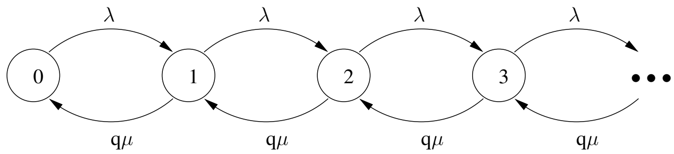

Figure 2: Markov state diagram of the feedback queue and the M/M/1 queue

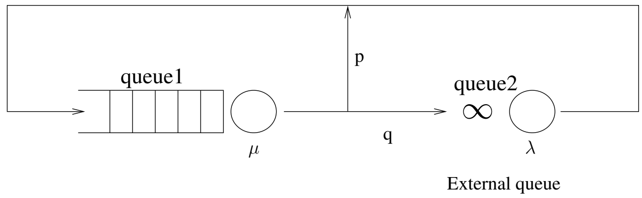

Figure 3: The feedback queue viewed as a network of two queues

The A* traffic processes is the embedded stochastic process resulting from viewing the state of the network just prior to every arrival transition in the network. The arriving job is not included in the state description. In this paper, we show that the A* process is a Markov chain that is recurrent and irreducible. The queueing network configuration is captured in the A* process, so the underlying Markov chain could be periodic or non-periodic depending on the configuration. Prior studies of arrival instant traffic processes have viewed the state of the network just prior to arrival at a single queue [2, 3, 6]. The A* traffic process views the network state prior to arrival at every queue in the network. Typically, such traffic processes are very complex [6]. However, this is not the case with the A* traffic process. It is surprising that this traffic process has not been analyzed since it is simple and captures a lot of information about the queueing network. The stationary distribution of the A* Markov chain can be used to:

1. compute the limiting queue length distribution of the queueing network (if this distribution exists),

2. compute the arrival instant queue length distribution of every queue in the network, where the arrival instant distribution refers to the probability of finding $i \geq 0$ jobs in the feedback queue just prior to a job joining the queue. The arriving job could be either

an external arrival or a feedback arrival.

3. show the relationship between the arrival instant distribution and the limiting queue length distribution.

In this paper, we analyze the A* processes of the feedback queue, an ordinary M/M/1 queue, and a closed feedback network. The A* process explains the product-form behavior of these different networks. In the next section, we analyze the A* traffic process of the feedback network.

2 A* Traffic Process

We first analyze the feedback network shown in Figure 3. The feedback queue is queue 1 while the external queue is queue 2. The jobs in the network move from queue to queue according to the Markov switching rule [2] given below:

$$R = \begin{vmatrix} \mathsf{p} & \mathsf{q} \\ 1 & 0 \end{vmatrix}$$

The Markov switching matrix, R, referred to as the *routing* matrix, is stochastic and finite. Hence, the switching process is irreducible and there exists a solution for the system of stationary linear equations [4]:

$$\mathbf{V}R = \mathbf{V} \tag{1}$$

The solution is unique up to a multiplicative constant. The ratio $\mathsf{V}(i)/\mathsf{V}(j)$ is the mean number of visits to queue i in between two visits to queue j [1]. The solution $\mathbf{V}$ gives the relative *visit counts* to the two queues in the network [7]. The relative visit count of the external queue with respect to the feedback queue is q.

The queue length process for the network is given by:

$$\mathcal{Y} = \{\, Y(\mathsf{T}) : \mathsf{T} \in \mathsf{R}_{+}\}$$

where $Y(\mathsf{T}) = (i, \infty)$ represents the state of the network at time T. There are $i \geq 0$ jobs in the feedback queue and an infinite number of jobs in the external queue. For notational simplicity, the state of the feedback network can be written as:

$Y(\mathsf{T}) = i;\ \ i \geq 0$ is the number of jobs in the feedback queue at time T

$\mathcal{Y}$ is a Markov process that is irreducible. The state space of $\mathcal{Y}$ is the set $E = \{0, 1, 2, 3, \cdots\}$. If $\lambda < \mathsf{q}\mu$, then $\mathcal{Y}$ has a limiting state distribution [4].

Observe the process $\mathcal{Y}$ just prior to every arrival at the feedback queue and the external queue. Consider the embedded stochastic process

$$\mathsf{X} = \{\mathsf{X}_{\mathsf{n}} = Y(\mathsf{T}_{\mathsf{n}}^{-})\}$$

where

T_n^- is time just prior to the n^{th} arrival instant, $n = 1, 2, 3, ..$ and

$$\mathsf{X}_{\mathsf{n}} = \begin{cases} (i^*, \infty) = i^* & \text{if the arrival is to the feedback queue;} \\ (i, \infty^*) = i & \text{if the arrival is to the external queue.} \end{cases}$$

The network state is viewed just prior to an arrival, so the arriving job is not included in the state description. The * symbol in the state description denotes the queue at which the job arrives. The embedded queue length process $\mathsf{X} = \{\mathsf{X}_{\mathsf{n}} : \mathsf{n} = 1, 2, \cdots\}$ is called the Arrival* (or A*) process for the network. The state space of X is the set $E \cup E^{*} * = \{0, 0^*, 1, 1^*, 2, 2^*, 3, 3^*, \cdots\}$.

Theorem 2.1 *The A* process* X *is a Markov chain.*

Proof: We need to show that

$$\mathrm{Prob}\{\mathsf{X}_{\mathsf{n+1}}|\mathsf{X}_0, \cdots, \mathsf{X}_{\mathsf{n}}\} = \mathrm{Prob}\{\mathsf{X}_{\mathsf{n+1}}|\mathsf{X}_{\mathsf{n}}\}.$$

$$\begin{aligned} \text{If } \mathsf{X}_{\mathsf{n}} &= i^*, \text{ then } Y(\mathsf{T}) = i+1, \mathsf{T} \in [T_n, T_{n+1}), \\ & i \geq 0, \qquad (2) \\ \text{if } \mathsf{X}_{\mathsf{n}} &= i, \text{ then } Y(\mathsf{T}) = i, \; \mathsf{T} \in [T_n, T_{n+1}), \\ & i \geq 0. \qquad (3) \end{aligned}$$

If $Y(\mathsf{T}) = (0, \infty) = 0$, then the next transition is an arrival from the external queue at rate λ with probability 1. If $Y(\mathsf{T}) = (i, \infty) = i, i > 0$, then the next transition depends on whether an external arrival occurs or whether the job at the feedback queue completes. The rate of this transition is $\lambda + \mu$.
The external arrival occurs first with probability $\frac{\lambda}{\lambda+\mu}$, the job at the feedback queue completes first with probability $\frac{\mu}{\lambda+\mu}$.
If the job at the feedback queue completes first, then the next arrival is at the feedback queue with probability p or at the external queue with probability q.
Thus, $\mathsf{X}_{\mathsf{n+1}}$ depends only on X_{n}. □

Let Q be the transition matrix of the embedded process X with state space $E \cup E^* = \{0, 0^*, 1, 1^*, 2, 2^*, 3, 3^*, \cdots, \}$. For $i \in E$, $i^* \in E^*$, and $j \in E \cup E^*$, Q is given by [1]:

$$Q(i^*, j) = \begin{cases} \frac{\mathsf{p}\mu}{\lambda+\mu}, & \text{if } j = i^*; \\ \frac{\mathsf{q}\mu}{\lambda+\mu}, & \text{if } j = i; \\ \frac{\lambda}{\lambda+\mu}, & \text{if } j = (i+1)^*; \\ 0, & \text{otherwise.} \end{cases}$$

When the number of jobs in the feedback queue $i > 0$,

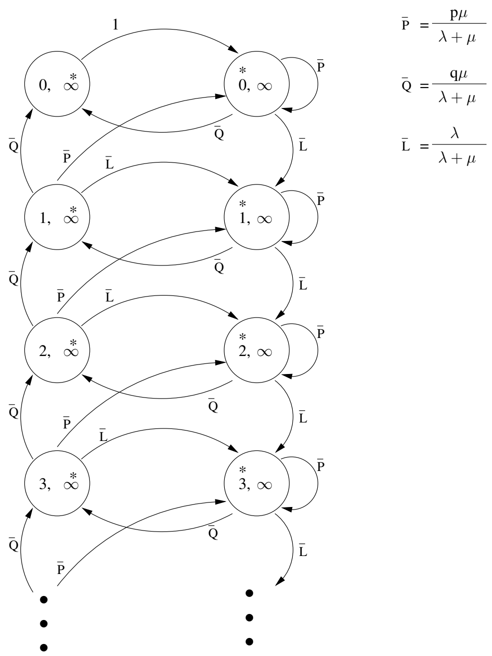

Figure 4: Markov state diagram of the A* process for the feedback network

then

$$Q(i, j) = \begin{cases} \frac{\mathsf{p}\mu}{\lambda+\mu}, & \text{if } j = i - 1^*; \\ \frac{\mathsf{q}\mu}{\lambda+\mu}, & \text{if } j = i - 1; \\ \frac{\lambda}{\lambda+\mu}, & \text{if } j = i^*; \\ 0, & \text{otherwise.} \end{cases}$$

When the number of jobs in the feedback queue $i = 0$, then

$$Q(0, j) = \begin{cases} 1, & \text{if } j = 0^*; \\ 0, & \text{otherwise.} \end{cases}$$

The Markov state diagram of the A* process is given in Figure 4. If $\lambda < \mathsf{q}\mu$, then the system of traffic equations

$$\pi Q = \pi \qquad (4)$$

along with the boundary condition $\sum_{E \cup E^*} \pi(i) = 1$, has a solution called the stationary distribution [4]. Note that it is possible to get a solution for the stationary equations of the Markov chain even when the Markov chain has no limiting distribution [1, 4].

Let p denote the limiting state probability of the queue length process $\mathcal{Y}$:

$$p(i) = \lim_{\mathsf{T}\to\infty} \text{Prob}\{\, Y(\mathsf{T}) = i\}$$

From Equations 2 and 3 it follows that the sojourn times in each state of the Markov chain X has an exponential distribution with parameter γ where:

$$\begin{aligned}\gamma(i^*) &= \lambda + \mu \quad \text{when } i^* \geq 0;\\ \gamma(i) &= \lambda + \mu \quad \text{when } i > 0;\\ &= \lambda \quad \text{when } i = 0.\end{aligned}$$

Therefore, the process $(\mathsf{X}, \mathsf{T}) = \{\mathsf{X_n}, \mathsf{T_n}; \mathsf{n} \in \mathsf{N}\}$ is a Markov renewal process. Equations 2 and 3 state that:

when $\mathsf{X_n} = \mathsf{i}^*$, $\quad Y(\mathsf{T}) = i + 1$ during the sojourn interval $\mathsf{T} \in [T_n, T_{n+1})$;

when $\mathsf{X_n} = \mathsf{i}$, $\quad Y(\mathsf{T}) = i$ during the sojourn interval $\mathsf{T} \in [T_n, T_{n+1})$.

Thus,

$$Y(T_n) = i \quad \Longrightarrow \quad \mathsf{X_n} = \mathsf{i} \ \text{ or } \ \mathsf{X_n} = (\mathsf{i}-1)^* \qquad (5)$$

If $\lambda < \mathsf{q}\mu$, then $\mathcal{Y}$ has a limiting distribution. Since $\mathcal{Y}$ is a recurrent, irreducible Markov process, the limiting distribution of $\mathcal{Y}$ is equal to the stationary distribution of $\mathcal{Y}$. Equation 5 and the fundamental theorem that relates the stationary distribution of $\mathcal{Y}$ at any instant to the stationary distribution of the embedded Markov chain X [1] lead to the following:

Theorem 2.2

$$p(i) = \frac{\pi(i-1^*)/\gamma(i-1^*) \; + \; \pi(i)/\gamma(i)}{\sum_{j\in E\cup E^*} \pi(j)/\gamma(j)} \qquad i \in E \quad (6)$$

The stationary equations for the A* Markov process underlying the feedback queue are easy to solve. Below, we present some key results that follow by solving the stationary equations.

Result 2.1 *The system of traffic equations given by Equation 4 reduce to the following:*

$$\begin{aligned}\lambda\pi(i) &= \mathsf{q}\mu\pi(i+1) \quad i = 0, 1, 2, \cdots\\ \lambda\pi(i^*) &= \mathsf{q}\mu\pi(i+1^*) \quad i^* = 0^*, 1^*, 2^*, \cdots\end{aligned}$$

These are identical in context to the local balance equations for both the feedback queue and the ordinary M/M/1 queue with parameters λ and $\mathsf{q}\mu$:

$$\lambda p(i) \;=\; \mathsf{q}\mu p(i+1) \quad i = 0, 1, 2, \cdots$$

where $p(i)$ is the limiting probability of i jobs in the ordinary M/M/1 queue and the feedback queue.

Result 2.2

$$\mathsf{V}(2)\sum_{i=0}^{\infty} \pi(i^*) = \mathsf{V}(1)\sum_{i=0}^{\infty} \pi(i)$$

where $\mathsf{V}(1)$ and $\mathsf{V}(2)$ are the relative visit counts for the network given by Equation 1.

The result can be independently derived from Equation 1 since $\sum_{i=0}^{\infty} \pi(i^*)$ gives the relative number of visits to the feedback queue when compared to the number of visits to the external queue given by $\sum_{i=0}^{\infty} \pi(i)$.

Result 2.3

$$\mathsf{V}(2)\pi(i^*) = \mathsf{V}(1)\pi(i)$$

This result can be independently derived by using reversibility of product-form queues [5].

We now show how a queue's arrival instant distribution can be derived from the A* process. Let
$\mathsf{X}_\mathsf{n}^\mathsf{a}$ represent the state of the Markov process $\mathcal{Y}$ just prior to job arrivals at the feedback queue. The job arrivals include external arrivals and feedback arrivals.
The A* process incorporates both the arrival instant queue information and the visit count information. The embedded Markov process $\mathsf{X}_\mathsf{n}^\mathsf{a}$ only incorporates the arrival instant queue information. The following result follows:

Result 2.4 *Let $p^a(i)$ represent the limiting probability of finding i jobs in the feedback queue just prior to arrival at the feedback queue.*

$$p^a(i) = C^a \times \pi(i^*), \quad i = 0, 1, 2, \cdots$$

where $C^a = \frac{1}{\sum_{i=0}^{\infty} \pi(i^*)}$.

Theorem 2.2 and Result 2.4 show the relationship between the limiting distribution and the arrival instant distribution of the feedback queue.

3 M/M/1 versus Feedback

The ordinary M/M/1 queue can be viewed as a two queue network as shown in Figure 5, where the M/M/1 queue is queue 1 and the external queue is queue 2. There is no feedback, so the relative visit counts of both queues is 1. The service rate of the M/M/1 queue is $\mathsf{q}\mu$ and the service rate of the external queue with infinite jobs is λ.

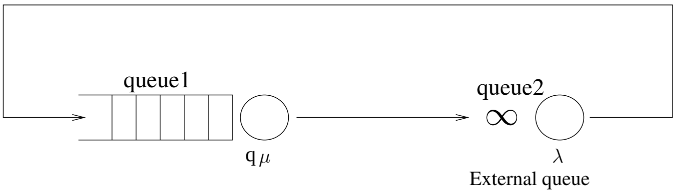

Figure 5: The M/M/1 queue viewed as a network of two queues

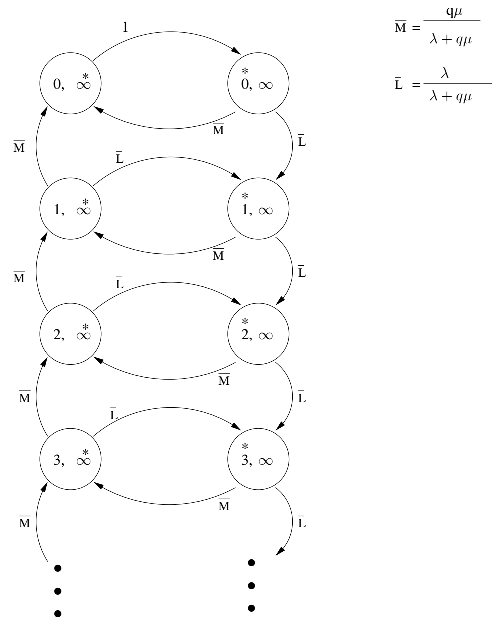

Figure 6: Markov state diagram of the A* process for the M/M/1 queue

Figure 6 shows the Markov state diagram of the embedded A* process for the M/M/1 network. The A* process for the M/M/1 queue is different from the A* process for the feedback queue. The underlying Markov chain of the A* process for the M/M/1 queue is periodic with period 2, while the underlying Markov chain of the A* process for the feedback queue is non-periodic. If $\lambda < \mathsf{q}\mu$, then the Markov chain has a stationary state distribution.

(Note that the stationary state distribution can exist even if the underlying chain is periodic [1, 4].) Let π_M represent the stationary state distribution of the A* process for the M/M/1 queue. The sojourn times in each state of the Markov chain X_M has an exponential distribution with parameter γ_M where:

$$\begin{aligned}\gamma_M(i^*) &= \lambda + q\mu \quad \text{when } i^* \geq 0;\\ \gamma_M(i) &= \lambda + q\mu \quad \text{when } i > 0;\\ &= \lambda \quad \text{when } i = 0.\end{aligned}$$

The queue length process of the feedback queue and the M/M/1 queue are identical. So, we use p to represent the limiting queue length distribution for both the feedback queue and the M/M/1 queue. Using the similar reasoning as in feedback queues:

$$p(i) = \frac{\pi_M(i-1^*)/\gamma_M(i-1^*) \; + \; \pi_M(i)/\gamma_M(i)}{\sum_{j\in E\cup E^*} \pi_M(j)/\gamma_M(j)}, i \in E \quad (7)$$

The stationary equations for the A* Markov process underlying the M/M/1 queue are easy to solve. Below, we present some key results that follow by solving the stationary equations.

Result 3.1 *The system of traffic equations relating to the Markov chain underlying the A* process for the M/M/1 queue reduce to the following:*

$$\begin{aligned}\lambda\pi_M(i) &= \mathsf{q}\mu\pi_M(i+1) \quad i = 0, 1, 2, \cdots\\ \lambda\pi_M(i^*) &= \mathsf{q}\mu\pi_M(i+1^*) \quad i^* = 0^*, 1^*, 2^*, \cdots\end{aligned}$$

Result 3.1 corresponds to Result 2.1 for feedback queues. These two results show the link between the traffic processes of the feedback queue and the M/M/1 queue.

Result 3.2

$$\sum_{i=0}^{\infty} \pi_M(i^*) = \sum_{i=0}^{\infty} \pi_M(i)$$

Result 3.2 corresponds to Result 2.2 for the feedback queue.

Result 3.3

$$\pi_M(i^*) = \pi_M(i)$$

Result 3.3 corresponds to result 2.3 for the feedback queue. The arrival instant distribution for the M/M/1 queue can be computed from the A* process using the approach for deriving Result 2.4 in the feedback queue.

4 Close versus Open

We now analyze the closed queueing counterpart of the open feedback queue. Figure 7 shows the queueing network analyzed in the paper. There are no external arrivals or external departures, and there are a fixed number of jobs constantly circulating in the network. The relative visit count of queue 2 is q compared to queue 1.

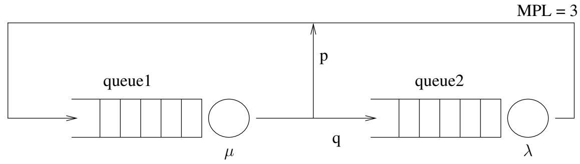

Figure 7: Closed queueing counterpart of the open feedback queue

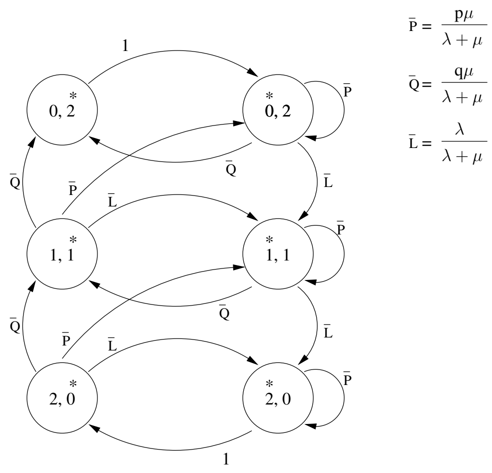

Figure 8: Markov state diagram of the A* process for the closed feedback network with 3 jobs

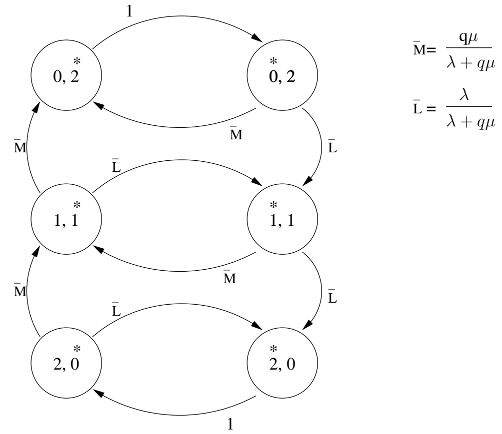

Figure 9: Markov state diagram of the A* process for a closed M/M/1 (no feedback) network with 3 jobs

Figure 8 shows the Markov state diagram of the embedded A* process for the closed network. The Markov chain underlying the A* process for the closed network is similar to the Markov chain for the open feedback queue. The only difference between the two chains reflects the impact of an infinite job population in the open network and a finite job population in the closed network. The stationary state distribution of the A* process, the limiting distribution of the network's queue length, and the arrival instant queue length for each queue can be computed using the approaches presented earlier for the open feedback queue.

Next, we analyze the closed network counterpart of the ordinary M/M/1 queue with no feedback. In Figure 5, replace the external queue with an ordinary queue. Figure 9 shows the Markov state diagram of the embedded A* process for this closed M/M/1 network with 3 jobs. Just as in the case of the feedback queue, the only difference between the open and closed chains reflects the impact of an infinite job population in the open network and a finite job population in the closed network.

Finally, compare the underlying A* Markov chain of the closed feedback network to the A* Markov chain of the closed M/M/1 network. The two networks have identical limiting queue length distributions [7]. Just as in the case of the open networks, it can be shown that the stationary state traffic equations corresponding to the two chains are equivalent. The traffic equations are identical to the local balance equations of the steady-state queue length process.

5 Conclusion

This paper establishes the A* traffic process in queueing networks. In addition to a queue's arrival instant distribution, this traffic process captures a lot of information about the network, such as visit counts, network configuration, and service rates. The A* process is generated by observing the state of a network just prior to every arrival instant at all the queues in the network. The paper shows that the A* queue length process is a Markov chain, and is simpler than traffic processes generated by observing network states prior to arrival at a specific queue in the network. The A* process is a Markov renewal process, so results pertaining to Markov renewal theory are valid for this process.

In this paper, the A* traffic process is used to explain why the feedback queue has identical queue length distribution as an ordinary M/M/1 queue. The A* embedded Markov chain captures both the similarities and differences in the two queues. The A* process is also used to show the similarities between the open feedback queue and a closed feedback network. This traffic process goes a long way in removing some of the mystery of why these dissimilar networks have similar product-form behaviors.

This paper is preliminary; the A* process needs to be evaluated in more detail for a class of networks, such as the BCMP networks. The traffic process could be used to understand the impact of network configuration, visit counts, and service rates on the performance of the network. The A* process could be used to independently prove some well known results. For example, the embedded Markov chain captures network wide information at every arrival instant. For closed networks, this information could be used to prove the equality of arrival instant state distribution and the steady state state distribution in a network with one less job [6, 8]. This paper does not look at the queue departure instant distribution. It would be interesting to see if a queue's departure instant network state distribution can be computed from the A* process (without using the knowledge that arrival and departure instant network state distributions are identical for product-form).

It would also be interesting to look at the A* processes of non-product form networks and compare them to those of product form networks. In fact, our initial motivation for looking at the A* traffic process was to help understand non-product form networks. If one can understand how the flows of product form networks differ from non-product-form networks, then this could lead to better tools to evaluate the performance of these complex networks.

References

[1] Erhan Çinlar. *Introduction to Stochastic Processes*. Prentice Hall College Div, 2007.

[2] Ralph Disney and Peter Kiessler. *Traffic Processes in Queueing Networks (Johns Hopkins Studies in the Mathematical Sciences)*. The Johns Hopkins University Press, 1987.

[3] Erol Gelenbe and Guy Pujolle. *Introduction to Queueing Networks, 2nd edition*. John Wiley & Sons Ltd., 1998.

[4] Donald Gross and Carl M. Harris. *Fundamentals of Queueing Theory*. Wiley-Interscience, 1998.

[5] F. P. Kelly. *Reversibility and Stochastic Networks*. Wiley series in probability and mathematical statistics, 1979.

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

[7] Daniel A. Menasce, Lawrence W. Dowdy, and Virgilio A.F. Almeida. *Performance by Design: Computer Capacity Planning By Example*. Prentice Hall PTR, 2004.

[8] K. C. Sevcik and I. Mitrani. The distribution of queuing network states at input and output instants. *J. ACM*, 28(2):358–371, 1981.

[9] L. Takàcs. A single server queue with feedback. *Bell Systems Technical Journal*, 42:505 – 519, 1963.