Networks of Queues

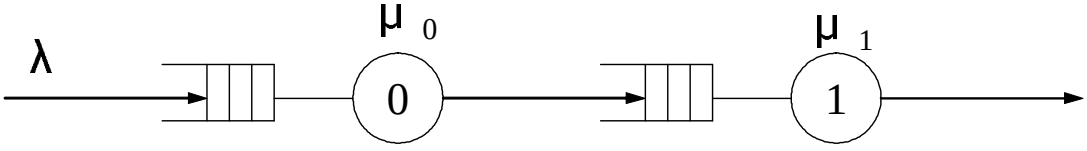
Ting Yan and Malathi Veeraraghavan, April 19, 2004

1. Introduction

Networks of Queues are used to model potential contention and queuing when a set of resources is shared. Such a network can be modeled by a set of service centers. Each service center may contain one or more servers. After a job is served by a service center, it may reenter the same service center, move to another one, or leave the system.

In an open queuing network, jobs enter and depart from the network. In a closed queuing network, jobs neither enter nor depart from the network. Open queuing networks can be further divided into two categories: open feed forward queuing networks and open feedback queuing networks. In an open feed forward queuing network, a job cannot appear in the same queue for more than one time. In an open feedback queuing network, after a job is served by a queue, it may reenter the same queue.

2. Two-Stage Tandem Network with Independent Service Times



The above figure shows a two-stage tandem network composed of two nodes with service rates μ_0 and μ_1, respectively. The external arrival rate for node 0 is λ and the arrival process is Poisson. Assume that the service times at each node are exponentially distributed and mutually independent, and independent of the arrival processes.

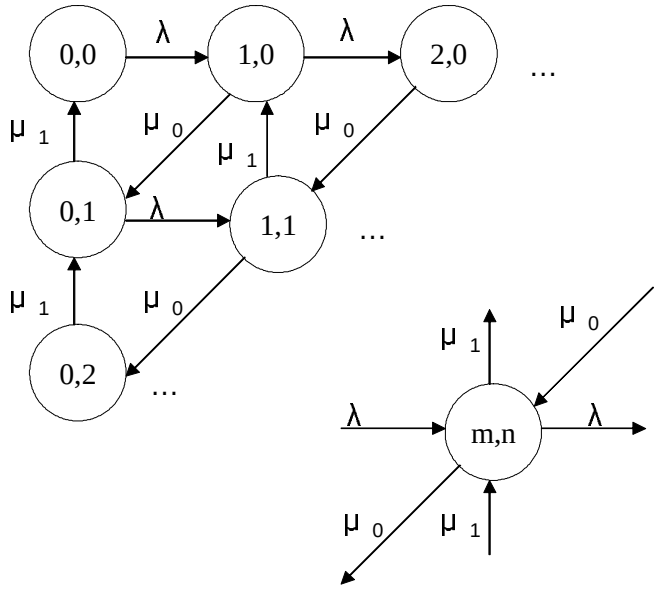
The first interesting question is what the arrival process at node 1. According to Burke's Theorem, it is also Poisson. Then what is the rate of the Poisson process? Intuitively, it is λ. For stability we also need to assume that $\mu_0 < \lambda$ and $\mu_1 < \lambda$. Then both nodes are M/M/1 queues. In order to get the joint distribution of the numbers in both nodes, can we simply apply M/M/1 results and have the following equation? (N_0 and N_1 denote the numbers in node 0

and node 1, respectively. $\rho_0 = \mu_0 / \lambda$, $\rho_1 = \mu_1 / \lambda$.)

$$p(N_0, N_1) = (1-\rho_0)\rho_0^{N_0}(1-\rho_1)\rho_1^{N_1} \tag{2.1}$$

In order to get the above result, we need the assumption that N_0 and N_1 are mutually independent. Does it hold? Burke's Theorem says for M/M/1 queue at any time t, the number in the system is independent of the sequence of departure times prior to t. Therefore, N_0 is independent of N_1. The proof of Burke's Theorem is given in the appendix.

We can also analyze the state diagram for the stochastic process (N_0, N_1) and derive $p(N_0, N_1)$ in a similar way to the derivation of state probabilities for M/M/1 queue.



For the internal states, we have the following balance equations:

$$(\mu_0 + \mu_1 + \lambda)\, p(N_0, N_1) = \mu_0 p(N_0 + 1, N_1 - 1) + \mu_1 p(N_0, N_1 + 1) + \lambda p(N_0 - 1, N_1),$$
$$N_0 > 0, N_1 > 0. \tag{2.2}$$

For the boundary states, we have:

$$(\mu_0 + \lambda)\, p(N_0, 0) = \mu_0 p(N_0 - 1, 0) + \mu_1 p(N_0, 1), N_0 > 0, \tag{2.3}$$

$$(\mu_1 + \lambda)\, p(0, N_1) = \mu_0 p(1, N_1 - 1) + \mu_1 p(0, N_1 + 1), N_1 > 0, \tag{2.4}$$

$$\lambda p(0,0) = \mu_1 p(0,1). \tag{2.5}$$

After all, for normalization we have:

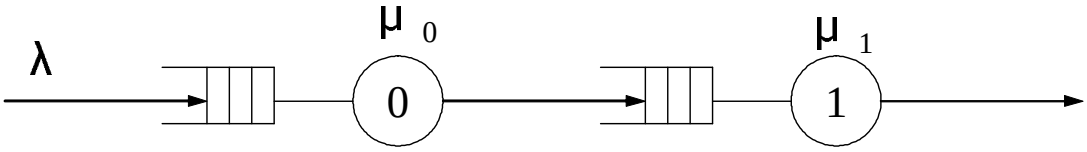
$$\sum_{N_0 \geq 0} \sum_{N_1 \geq 0} p(N_0, N_1) = 1. \tag{2.6}$$

Solve these equations and we obtain the solution,

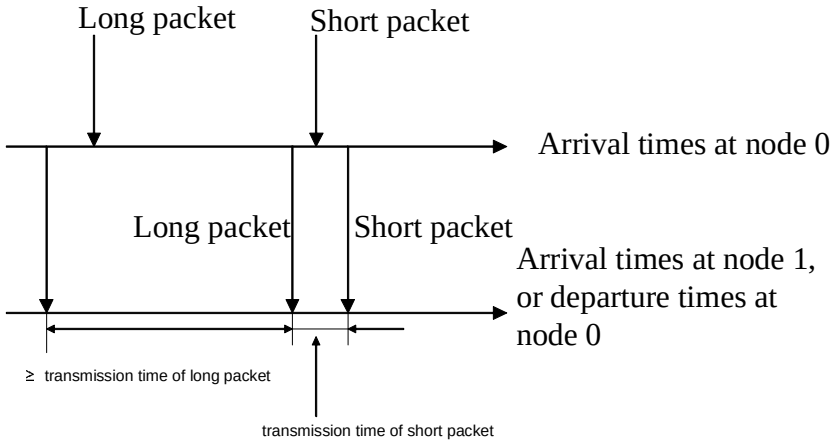
$$p(N_0, N_1) = (1-\rho_0)\rho_0^{N_0}(1-\rho_1)\rho_1^{N_1} \tag{2.7}$$

which is exact the same as (2.1).

3. Two-Stage Tandem Network with Dependent Service Times



Consider that the two nodes are transmission lines, where service time is proportional to the packet length. We also assume that the packet lengths are Poisson and independent of the arrival process. Does the above result still apply? The answer is no. The issue here is that the arrival times for node 2 are strongly correlated with packets lengths, and therefore the service process. The following figure shows why.



From the above figure, we can also see that a long packet suffers less waiting time than a short one does on average. The reason is that it takes longer for a long packet to be transmitted in the first line, and therefore the second line gets more time to empty out.

There exists no analytical results for such networks in which interarrival and service times are dependent. However, *Kleinrock independence approximation* states that "merging several packet streams on a transmission line has an effect akin to restoring the independence of interarrival times and packet lengths" [1] thus an *M/M/1* model can be used to analyze the behavior of each communication link.

When the arrival/service time correlation is eliminated and randomization is used to divide the traffic, Jackson's Theorem provides an analytical approach to derive the average numbers in the system for a broad category of queuing networks.

4. Average Delay

Consider a network composed of nodes and links between nodes. Applying Kleinrock independence approximation, each link can be modeled as an M/M/1 queue. Thus we have the average number of packets in queue or service at link (i, j) is

$$N_{ij} = \frac{\lambda_{ij}}{\mu_{ij} - \lambda_{ij}} \tag{4.1}$$

After summing over all queues, we have

$$N = \sum_{(i,j)} \frac{\lambda_{ij}}{\mu_{ij} - \lambda_{ij}} \tag{4.2}$$

Apply Little's Law and ignore processing and propagation delay, the average delay per packet

$$T = \frac{1}{\gamma} \sum_{(i,j)} \frac{\lambda_{ij}}{\mu_{ij} - \lambda_{ij}} \tag{4.3}$$

where γ is the total arrival rate in the system. If the delay d_{ij} can not be ignored, the formula should be modified to

$$T = \frac{1}{\gamma} \sum_{(i,j)} \left(\frac{\lambda_{ij}}{\mu_{ij} - \lambda_{ij}} + \lambda_{ij} d_{ij} \right) \tag{4.4}$$

And the average delay per packet for a certain traffic stream traversing a path *p* is

$$T_p = \sum_{(i,j) \in p} \left(\frac{\lambda_{ij}}{\mu_{ij}(\mu_{ij} - \lambda_{ij})} + \frac{1}{\mu_{ij}} + d_{ij} \right) \tag{4.5}$$

5. Jackson's Theorem for Open Queuing Networks

Jackson's Theorem provides a general product-form solution for both feed forward and feedback open queuing networks.

The assumptions for Jackson's Theorem are:

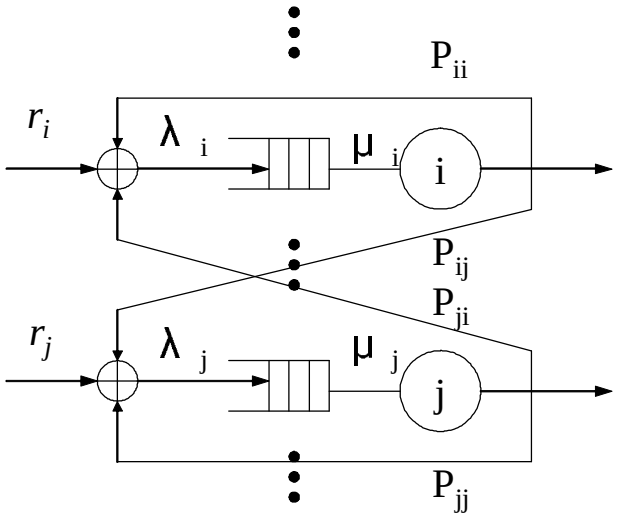
(1) the network is composed of *K* FCFS, single-server queues

(2) the arrival processes for the *K* queues are Poisson at rate $r_1, r_2, \ldots, r_K$;

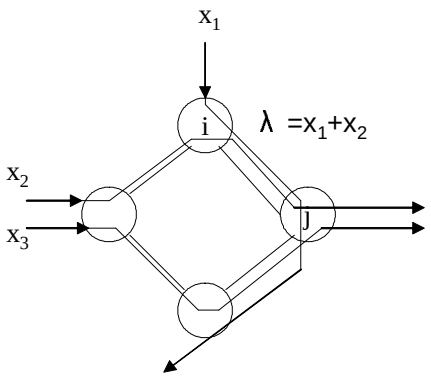
(3) the service times of customers at j^{th} queue are exponentially distributed with mean $1/\mu_j$ and they are mutually independent and independent f the arrival processes;

(4) once a customer is served at queue *i*, it joins each queue *j* with probability P_{ij} or leave the system with probability $1 - \sum_{j=1}^{K} P_{ij}$. P_{ij} is called the routing probability from node *i* to node *j*. For all possible *i* and *j*, P_{ij} compose the routing matrix.

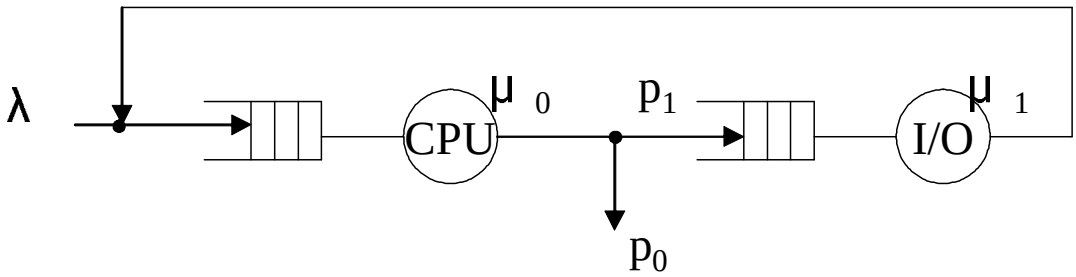
The following figure shows the general structure for open queuing networks.



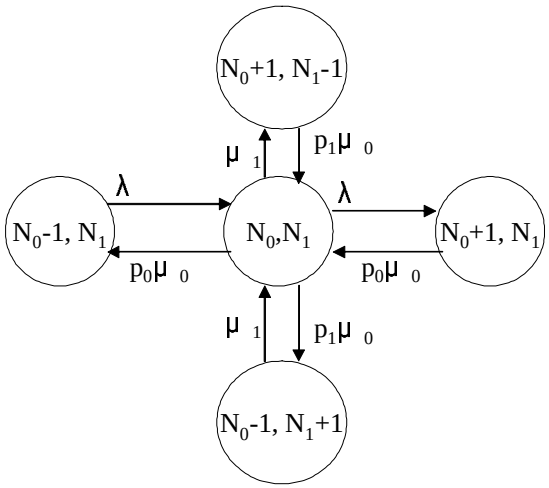
The following figure shows a virtual circuit network example.



Here is a feedback example.



And the following figure gives the internal state transitions.



In order to calculate the arrival rates at each queue, firstly we have the following equations:

$$\lambda_j = r_j + \sum_{i=1}^{K} \lambda_i P_{ij}, \qquad j = 1, ..., K \tag{5.1}$$

Solve the linear equations and λ_i are obtained. Define utilization factor for each queue as $\rho_j = \lambda_j / \mu_j, j = 1, ..., K$.

Then we have:

Jackson's Theorem. Assuming that $\rho_j < 1, j = 1, ..., K$, we have for all $N_1, ..., N_K \geq 0$,

$$P(N_1, ..., N_K) = P_1(N_1) P_2(N_2) ... P_K(N_K) \tag{5.2}$$

where

$$P_j(N_j) = \rho_j^{N_j}(1 - \rho_j), N_j \geq 0 \tag{5.3}$$

Then we have the average number in each queue:

$$E[N_j] = \frac{\rho_j}{1 - \rho_j} \tag{5.4}$$

We may applying Little's Law and get the average response time. For example hen we have only one external arrival with rate λ, we have the average response time formula

$$E[R] = \frac{1}{\lambda} \sum_j E[N_j] \tag{5.5}$$

For feed forward networks, the above result is straightforward. It's trickier for feedback networks. In feedback networks, the arrival process for a queue may not be Poisson. The following is a simple example. Consider a queue in which $r \ll \mu$, and after a customer is served, it is sent back to the same queue with a probability p which is very close to 1. Given there is an arrival, it is very likely that there will be an arrival soon because the customer will be sent back again with a high probability. But when there is no customer in the system, because *r*

is very small, it is very unlikely that there will be an arrival soon. Evidently the arrival process is not memoryless. So the total arrival process may not be Poisson thus the queue is not *M/M/1*. Nevertheless, Jackson's Theorem still holds even when the total arrival process at each queue is not Poisson.

Jackson's Theorem can also be extended to even more general scenarios, for example *M/M/m* queues. We can generalize *M/M/m* or *M/M/*∞ to allow the service rate at each queue to depend on the number of customers at that queue. Suppose the service time at the j^{th} queue is exponentially distributed with rate $\mu_j(m)$, where m is the number in the queue just before the customer's departure. We define

$$\rho_j(m) = \lambda_j / \mu_j(m),\ j = 1,...,K, m = 1,2,... \tag{5.6}$$

and

$$\hat{P}_j(N_j) = \begin{cases} 1, & N_j = 0 \\ \rho_j(1)\rho_j(2)...\rho_j(N_j) & N_j > 0 \end{cases} \tag{5.7}$$

We have:

Jackson's Theorem for State-Dependent Service Rates. We have for all $N_1,...,N_K \geq 0$,

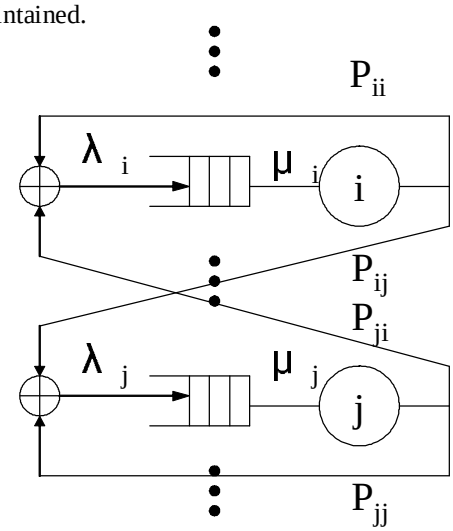
$$P(N_1,...,N) = \frac{\hat{P}_1(N_1)...\hat{P}_K(N_K)}{G} \tag{5.8}$$

assuming $0 < G < \infty$, where G is the normalization factor:

$$G = \sum_{N_1=0}^{\infty} ... \sum_{N_K=0}^{\infty} \hat{P}_1(N_1)...\hat{P}_K(N_K) \tag{5.9}$$

4. Closed Queuing Networks

Closed queuing networks model a system in which multiple resources are shared and no job enters or departs. It can also approximate a system involving multiple resource holding under heavy load. Although jobs enter and depart from these systems, under heavy load, once a job leaves the system, an already waiting job will be put in immediately so that a constant degree of multiprogramming *M* is maintained.



For closed queuing networks, we need to modify the system requirement by

$$\sum_{j=1}^{K} P_{ij} = 1, \qquad i = 1,...,K \tag{6.1}$$

and

$$\lambda_j = \sum_{i=1}^{K} \lambda_i P_{ij} \tag{6.2}$$

Note that there is no external arrival entering the system. Under certain conditions, (6.2) can be solved with the form

$$\lambda_j(M) = \alpha(M)\bar{\lambda}_j, \qquad j = 1,...,K \tag{6.3}$$

Denote

$$\rho_j(m) = \frac{\bar{\lambda}_j}{\mu_j(m)} \tag{6.4}$$

$$\hat{P}_j(N_j) = \begin{cases} 1, & N_j = 0 \\ \rho_j(1)\rho_j(2)...\rho_j(N_j) & N_j > 0 \end{cases} \tag{6.5}$$

and

$$G(M) = \sum_{\{(N_1,...,N_K)|N_1+...+N_K=M\}} \hat{P}_1(N_1)...\hat{P}_K(N_K) \tag{6.6}$$

We have:

Jackson's Theorem for Closed Networks: We have for all $N_1,...,N_K \geq 0$, and $N_1 + ... + N_K = M$,

$$P(N_1,...,N_K) = \frac{\hat{P}_1(N_1)...\hat{P}_K(N_K)}{G(M)} \tag{6.7}$$

How many states does the system have? Or, how many nonnegative integer solutions are there for the equation $N_1 + ... + N_K = M$? A little counting theory gives the result

$$\text{Number of system states} = \binom{M+K-1}{M} \tag{6.8}$$

This number increases exponentially with M and K, it is very difficult to calculate $G(M)$ with (6.6). Simple algorithms have been developed to make this mission possible.

If each node of the system is a single queue, then $\rho_j(m) = \rho_j$ for any m. (6.6) becomes

$$G(M) = \sum_{\{(N_1,...,N_K)|N_1+...+N_K=M\}} \rho_1^{N_1}...\rho_K^{N_K} \tag{6.9}$$

From (6.9), define a polynomial in z

$$\begin{aligned}\Gamma(z) &= \prod_{i=1}^{K} \frac{1}{1-\rho_i z} \\ &= (1+\rho_1 z+\rho_1^2 z^2+...)(1+\rho_2 z+\rho_2^2 z^2+...)... \\ &\quad (1+\rho_K z+\rho_K^2 z^2+...)\end{aligned} \tag{6.10}$$

This is the generating function of $G(1)$, $G(2)$, …

$$\Gamma(z) = \sum_{n=0}^{\infty} G(n) z^n \tag{6.11}$$

where G(0)=1.

Define

$$\Gamma_i(z) = \prod_{j=0}^{i} \frac{1}{1-\rho_j z}, \qquad j = 1,...,K \tag{6.12}$$

and

$$\Gamma_i(z) = \sum_{j=0}^{\infty} G_i(j) z^j, \qquad i = 1,...,K \tag{6.13}$$

where $G_K(j) = G(j)$.

We will be able to get the recursive formula to compute $G_K(j)$:

$$G_i(j) = G_{i-1}(j) + \rho_i G_i, \qquad \begin{matrix} i = 2,3,...,K \\ j = 1,2,...,M \end{matrix} \tag{6.14}$$

with the initial values $G_1(j) = \rho_1^j$, $j = 1,2,...,M$ and $G_i(0) = 1$, $i = 1,2,...,K$.

This algorithm is efficient both in time and space.

Another approach is called Mean Value Analysis, in which the average number of customers and average customer time spent per visit in each queue are directly calculated. Assume the service rate does not depend on states. First, when $M = 0$, we have trivially

$$T_j(0) = N_j(0) = 0, \qquad j = 1,...,K \tag{6.15}$$

. Then the Arrival Theorem is applied to get the recursive formula for $T_j(s)$:

$$T_j(s) = \frac{1}{\mu_j}(1+N_j(s-1)), \qquad \begin{matrix} j = 1,...,K \\ s = 1,...,M \end{matrix} \tag{6.16}$$

Finally Little's Law is applied to get $N_j(s)$:

$$N_j(s) = s\frac{\bar{\lambda}_j T_j(s)}{\sum_{i=1}^{K} \bar{\lambda}_j T_j(s)}, \qquad \begin{matrix} j = 1,...,K \\ s = 1,...,M \end{matrix} \tag{6.17}$$

The Arrival Theorem states that in a closed product-form queuing network, the probability mass function of the number of jobs seen at the time of arrival to node i when there are n jobs in the network is equal to that of the number of jobs at the node with one less job in the network.

Appendix – Burke's Theorem and Reversibility

Burke's Theorem: In steady-state of an *M/M/1*, *M/M/m*, or *M/M/∞* queue, the following hold true:

(a) The departure process is Poisson with the arrival rate λ.

(b) At each time t, the number of customers in the system is independent of the sequence of departure times prior to t.

To prove Burke's Theorem, we need to have some idea on reversibility.

Consider an irreducible and aperiodic DTMC X_n, X_{n+1}, … with transition probability P_{ij} and stationary distribution $\{p_j \mid j \geq 0\}$ with $p_j > 0$ for all j that is in steady-state, that is,

$$P\{X_n = j\} = p_j, \text{ for all } n \tag{0.1}$$

Consider the sequence of states going backward in time X_n, X_{n-1}, … . It can be proved that this sequence is also a Markov chain (how?) and

$$\begin{aligned} P_{ij}^* &= P\{X_m = j \mid X_{m+1} = i\} \\ &= \frac{p_j P_{ji}}{p_i} \end{aligned} \tag{0.2}$$

We say that the Markov chain is *time reversible* if $P_{ij}^* = P_{ij}$ for all *i, j*. It can be easily seen that the reversed chain is also irreducible, aperiodic, and has the same stationary distribution as the forward chain.

If we can find positive numbers p_i, $i \geq 0$, $\sum_{i=0}^{\infty} p_i = 1$ and

$$\sum_{j=0}^{\infty} P_{ij}^* = \sum_{j=0}^{\infty} \frac{p_j P_{ji}}{p_i} = 1, \qquad i = 0,1,... \tag{0.3}$$

then $\{p_i \mid i \geq 0\}$ is the stationary distribution and P_{ij}^* are the transition probabilities of the reversed chain. (Prove it with the global balance equation) A variation for CTMC is used to prove Jackson's Theorem by first guessing and proving the transition rates, and then proving the

stationary distribution as the form given in the theorem satisfies

$$\sum_{j=0}^{\infty} q_{ij} = \sum_{j=0}^{\infty} \frac{p_j q_{ji}}{p_i}, \text{ for all } \ i, j \geq 0 \tag{0.4}$$

A chain is time reversible if and only if the detailed balance equation (directly from definition)

$$p_i P_{ij} = p_j P_{ji}, \ \ i, j \geq 0 \tag{0.5}$$

We know that any birth-death processes are time reversible. So the queuing systems such as *M/M/*1, *M/M/m*, *M/M/m/m*, etc. are all time reversible.

For CTMC, the analysis and properties are analogous and the only difference is that transition rates are used instead of transition probabilities.

For a queuing system, which is time reversible, we may represent the reverse process by another queuing system in which departures correspond to arrivals of the original system and arrivals to departures in the original system. In steady-state, the forward and reversed systems are statistically indistinguishable, which gives part (a) of Burke's Theorem. At each time t, the sequence of departure times prior to t correspond to the arrival times after t in the reversed system. Since arrivals are Poisson, these future arrivals do not depend on or affect the number in the system. Therefore we have part (b) of Burke's Theorem.

This lecture note is taken from [1] mostly, sometimes verbatim. Some materials in [2] are also included.

References

[1] D. Bertsekas and R. Gallager, "Data Networks," Prentice Hall, 1986, ISBN 0-13-196825-4.

[2] K. S. Trivedi, "Probability, Statistics with Reliability, Queueing and Computer Science Applications," Second Edition, Wiley, 2002, ISBN 0-471-33341-7.