Tạp chí Tin học và Điều khiển học, T. 18, S. 3 (2002), 182–186

QUEUING NETWORK THEORY AND ITS APPLICATION TO COMMUNICATION SYSTEMS

LUONG HONG KHANH, VU NGOC PHAN

Abstract. The average arrival rates and the average delay at the queues in a queuing network are two significant quantities evaluating the network performance. The present paper will show how these two quantities can be calculated for the open queuing network and the closed queuing network used in communication systems.

Tóm tắt. Số lượng trung bình các nhu cầu phục vụ xuất hiện và thời gian trung bình một nhu cầu phục vụ nằm tại các nút của mạng hàng đợi là các đại lượng tiêu biểu để đánh giá chất lượng của một mạng dịch vụ. Bài báo này đề cập đến cách xác định hai đại lượng nêu trên đối với mạng hàng đợi đóng và mạng hàng đợi mở mô tả hệ thống truyền tin.

1. INTRODUCTION

The queuing theory initiated by Erlang has gained a wide applicability to communication system design and analysis [4]. The major performance measures in a communication network are blocking probability and delay of calls, packets, or cells. Almost investigation approaches of the network performance are usually based on the queuing theory. A communication network can be mathematically considered as a lattice of N nodes and L links. Customers arrive from outside the network at a queue. Each customer stays in the network to be served and the service needs normally a duration of time. A customer will leave the network after being served. The numbers of calls (generally said arrivals) at a queue at a fixed point and the service time have stochastic properties. Hence, a communication network is represented as a stochastic system.

The queuing theory has been used for designing ATM switching networks [2, 7], for designing and analysing broadband networks [5, 6], for designing and analysing the performance of computer communication network [3], for dynamical and optimal routing in telecommunication networks [1], etc. In Section 2 the fundamental of the open queuing network and closed queuing network is repeated. In Section 3 the modified mean value method is described as a tool for the network analysis. A simple example will make the issue easier to understand.

2. OPEN QUEUING AND CLOSED QUEUING NETWORKS

2.1. Open queuing network

As well known, a queue can be generally described by quintuple $\{A/B/m/N/p\}$. Here A is the name of the arrival distribution, B is the name of the service time distribution, m is the number of servers, N is the queue length and p is the maximum number of customers being permitted in the service system. A queuing network is a collection of queues which are connected in parallel or/ and in tandem. Let $\mathcal{N} = \{1, 2, 3, \ldots, N\}$ and $\mathcal{L} = \{1, 2, 3, \ldots, M\}$ denote the set of nodes and the set of links of a queuing network, respectively. Without loss of generality it is supposed here that the queue length is of infinity and the service is arranged by the first-come first-serve discipline. Furthermore, it is assumed that there is only one queue at each node and the arrival of customer at a queue is an independent random process. Let w_{ij} denote the probability that after a customer is served at queue i it continuously joins the queue j. Thus, in an open queuing network the following condition has to be satisfied

$$\sum_{j=1,\, j\neq i}^{N+1} w_{ij} = 1 \ \text{ for } i \in \mathcal{N}. \tag{2.1}$$

Let λ_i denote the arrival rate from outside and Λ_i denote the total arrival rate at queue i, respectively. Then, it is easy to know that:

$$\Lambda_i = \lambda_i + \sum_{j=1, i\neq j}^{N} \Lambda_j w_{ji} \ \text{ for } i \in \mathcal{N}. \tag{2.2}$$

Let $\Lambda = [\Lambda_1, \Lambda_2, \dots, \Lambda_N]^T$ denote the vector of arrivals in an open queuing network, called the total node flow vector. Let $\lambda = [\lambda_1, \lambda_2, \dots, \lambda_N]^T$ denote the vector of arrivals from outside, called the external node flow vector. Then the equation (2.2) can be commonly written as follows:

$$\Lambda = \lambda + W^T \Lambda. \qquad ((2.3))$$

Here W is a square matrix:

$$W = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1N} \\ w_{21} & w_{22} & \cdots & w_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ w_{N1} & w_{N2} & \cdots & w_{NN} \end{bmatrix}. \tag{2.4}$$

It is to remember that $w_{ii} = 0$ for $i = 1, \dots, N$. From Eq. (2.3) Λ can be obtained by the matrix inversion as following:

$$\Lambda = [I - W^T]^{-1} \lambda. \tag{2.5}$$

In Eq. (2.5), I represents an identity matrix.

Let x_i denote the number of customers at queue i. The vector $x = [x_1, x_2, \dots, x_N]^T$ describes the total number of customers in the queuing network and it is called the state vector of the service system. The following conditions are supposed to be hold.

- The arrival at queue i has the Poisson distribution with rate λ_i.
- The service time of customers at queue i has the exponential distribution with mean $\dfrac{1}{\mu_i}$.
- Arrival and departure process among every queues of the system are mutually independent.
- The transition time from a state to the next one is so short that it can be neglected.

Let $P_r(x_1, x_2, \dots, x_N)$ denote the probabitity that there are x_i customers at queue i, etc. According to these assumptions the balance principle can be expressed as follows

$$\begin{aligned} &P_r(x_1, x_2, \dots, x_N)\Big[\sum_{i=1}^{N} \lambda_i + \sum_{i=1}^{N}\sum_{j=1}^{N} \mu_i w_{ij} + \sum_{i=1}^{N} \mu_i w_{i(N+1)}\Big] \\ &= \sum_{i=1}^{N} \lambda_i P_r(x_1, \dots, x_i - 1, \dots, x_N) + \sum_{i=1}^{N}\sum_{j=1}^{N} \mu_i w_{ij} P_r(x_1, ..., x_i + 1, ..., x_j + 1, \dots, x_N) \\ &\qquad + \sum_{i} \mu_i w_{i(N+1)} P_r(x_1, ..., x_i + 1, \dots, x_N), \end{aligned} \tag{2.6}$$

The solution of (2.6) leads to:

$$P_r(x_1, x_2, \dots, x_N) = \prod_{i=1}^{N} P_r(x_i) = \prod_{i=1}^{N} (1 - q_i) q_i^{x_i}, \tag{2.7}$$

$$P_r(x_1, ..., x_i + 1, ..., x_j - 1, \dots, x_N) = \frac{q_i}{q_j} P_r(x_1, ..., x_i \dots, x_N). \tag{2.8}$$

In Eqs. (2.7) and (2.8) $q_i = \dfrac{\Lambda_i}{\mu_i}$ is called the relative untilization of queue i.

2.2. Closed queuing network

The closed queuing network is a network of queues where the total number of customers inside the network is fixed. That means, there is no customer departing from the network and no customer is allowed to enter into the network. Let K denotes the fixed number of customers in the network, x_i be the number of customers at queue i and N be the total number of queues. It is easily to recognise that.

$$\sum_{i=1}^{N} x_i = K. \tag{2.9}$$

Let $S(K, N)$ denote the state of the system, where:

$$S(K,N) = (x_1, x_2, \ldots, x_N \Big| \sum_{i=1}^{N} x_i = K). \tag{2.10}$$

It can be imagined that the number of the network states is very large. For simplicity, the following assumptions are made.

- In the closed queuing network the first-come first-serve discipline is used.
- The service time at queue i follows the exponential distribution with rate λ_i.
- Service rate of each queue is independent of its queue length.

Let w_{ij} denote the probability that after a customer is served at queue i it continuously joins the queue j like in the open queuing network. Here the condition

$$\sum_{j=1, j\neq i}^{N} W_{ij} = 1 \ \text{ for } \ i \in \mathcal{N} \tag{2.11}$$

is hold. An essential difference between the open queuing network and the closed queuing network is that the total arrival rate at queue i can not be determined by Eq. (2.2). Consider the state of the network when there is no customer departing the network and because of that there is no customer can enter into the network. For this situation the total arrival at queue i is determined by

$$\Lambda_i = \sum_{j=1, i\neq j}^{N} \Lambda_j w_{ji} \ \text{ for } \ i \in \mathcal{N} \tag{2.12}$$

or in the matrix from:

$$\Lambda = W^T \Lambda. \tag{2.13}$$

Equation (2.13) is singular and can not be solved. The transition from state $x = (x_1, x_2, \ldots, x_N)$ to state $x = (x_1, \ldots, x_i + 1, \ldots, x_j - 1, \ldots, x_N)$ corresponds to the situation in which a customer finishes his service at queue j and joins queue i. The following relation is hold.

$$\mu_i w_{ij} P_r(x_1, \ldots, x_i + 1, \ldots, x_j - 1, \ldots, x_N) = \mu_j w_{ji} P_r(x_1, \ldots, x_i, \ldots, x_j, \ldots, x_N) \tag{2.14}$$

or

$$P_r(x_1, \ldots, x_i + 1, \ldots, x_j - 1, \ldots, x_N) = \frac{q_i}{q_j} P_r(x_1, \ldots, x_i, \ldots, x_j, \ldots, x_N). \tag{2.15}$$

Here it is supposed that $w_{ij} = w_{ji}$. The solution of Eq. (2.15) gets the from:

$$P_r(x_1, \ldots, x_i, \ldots, x_N) = \frac{1}{G(K,N)} \prod_{i=1}^{N} q_i^{x_i}, \tag{2.16}$$

where $G(K, N)$ is a normalization constant to guarantee that $P_r(x_1, \ldots, x_i, \ldots, x_N)$ is a proper probabitity distribution, that means:

$$\sum_{x\in S(K,N)} P_r(x) = \sum_{x\in S(K,N)} \frac{1}{G(K,N)} \prod_{i=1}^{N} q_i^{x_i} = 1. \tag{2.17}$$

From (2.17) it is easily to identify that

$$G(K,N) = \sum_{x\in S(K,N)} \prod_{i=1}^{N} q_i^{x_i}. \tag{2.18}$$

By substituting $g_n(k) = G(k,n)$ the following recursive equation can be derived from Eq. (2.18)

$$g_n(k) = g_{n-1}(k) + q_n g_n(k-1), \quad k = 0,\dots,K,\ n = 1,\dots,N. \tag{2.19}$$

$G(K,N)$ is equal to $g_n(k)$ when $n = N$ and $k = K$.

3. APPLICATION TO COMMUNICATION NETWORK

Given a closed queuing network with K packets and N nodes each of which represents a queue. The typical interesting values are the expected number of packets in the network and the expected delay time. It can be show that these expected values can be determined without knowing the normalization constant $G(K,N)$. The expected number of packets in a queue can be expressed by the utilization of queue as follows

$$E\{x_i(K)\} = \sum_{\nu=1}^{K} q_i^{\nu} \frac{G(K-\nu,N)}{G(K,N)}. \tag{3.1}$$

From Eq. (3.1) it is easily to get the recursive equation

$$E\{x_i(K)\} = U_i(K)[1 + E\{x_i(K-1)\}], \quad i = 1,2,\dots,N, \tag{3.2}$$

where $U_i(K)$ is determined by

$$U_i(K) = q_i \frac{G(K-1,N)}{G(K,N)}. \tag{3.3}$$

The first and second terms in Eq. (3.2) represent the average service time and waiting time of packets, respectively. Substitute:

$$S_i(K) = \Lambda_i \frac{G(K-1,N)}{G(K,N)}. \tag{3.4}$$

Let

$$E\{d_i(K)\} = \frac{E\{x_i(K)\}}{S_i(K)} \tag{3.5}$$

be the expected delay time at queue i. After substituting Eqs. (3.2) and (3.4) in Eq. (3.5) it leads to

$$E\{d_i(K)\} = \frac{1}{\mu_i}(1 + E\{x_i(K-1)\}) = \frac{1}{\mu_i} + \frac{1}{\mu_i} E\{x_i(K-1)\}. \tag{3.6}$$

Eq. (3.6) indicates that the expected delay time at a queue equals the sum of average service time and average waiting time.

Example. Consider a closed queuing network of $K = 4$ and $N = 6$. The probalitity matrix W is given by

$$W = \begin{bmatrix} 0 & 0.1 & 0.3 & 0 & 0.3 & 0.3 \\ 0.2 & 0 & 0.1 & 0.6 & 0.1 & 0 \\ 0.2 & 0.5 & 0 & 0.1 & 0.1 & 0.1 \\ 0.3 & 0.1 & 0.1 & 0 & 0.4 & 0.1 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0 & 0.2 \\ 0.4 & 0.1 & 0.1 & 0.2 & 0.2 & 0 \end{bmatrix}. \tag{3.6}$$

The service rate of the queues are $\mu_1 = \mu_2 = \mu_5 = 4,\ \mu_3 = \mu_4 = 5.5,\ \mu_6 = 3$. Queue 6 is the source queue mit arrival rate $\Lambda_6 = 1$. The total packet arrival rate (TPAR) and the relative utilization (RUTI) of the queues are given in Table 3.1.

Table 3.1

Queue	1	2	3	4	5	6
TPAR Λ_i	1.2000	0.3000	0.3000	0.6000	0.6000	1
RUTI q_i	0.3000	0.0750	0.0545	0.1091	0.1500	0.3333
$S_i(K)$	2.6859	0.6715	0.6715	1.3430	1.3430	2.2383
$U_i(K)$	0.6715	0.1679	0.1221	0.2442	0.3357	0.7461
$E\{x_i(K)\}$	1.3152	0.1960	0.1365	0.3070	0.4625	1.5829
$E\{d_i(K)\}$	0.4897	0.2918	0.2032	0.2286	0.3444	0.7072

To calculate $S_i(K)$ and $U_i(K)$ it follows from Table 3.2 that $G(4, 6) = 2.2385$.

Table 3.2

	1	2	3	4	5	6
0	1	1	1	1	1	1
1	0.3000	0.3750	0.4295	0.5386	0.6886	1.0220
2	0.0900	0.1181	0.1416	0.2003	0.3036	0.6443
3	0.0270	0.0359	0.0436	0.0654	0.1110	0.3257
4	0.0081	0.0108	0.0132	0.0203	0.0370	0.1455

4. CONCLUSION

The queuing network theory plays an important role on many fields of technology and science. In this paper the fundamental of the queuing network theory and its application to communication systems have been shown. It is to note that the calculation of the expected arrival and the expected delay at the queue, which is indicated in the expressed example, is very difficult for a large network.

REFERENCES

[1] Ash G. R., *Dynamic Routing in Telecommunication Network*, McGraw Hill, 1997.
[2] Guizani M., Rayes A., *Designing ATM Switching Networks*, McGraw Hill, 1979.
[3] Hayes J. F., *Modelling and Analysis of Computer Communication Networks*, Plenum Press, New York, 1984.
[4] Kleinrock L., *Queuing Theory*, Wiley, 1975.
[5] Minoli D., *Broadband Network Analysis and Design. Bell Communications Research*, Inc. New York University, Artech House, Boston, London, 1993.
[6] Ross K. W., *Multiservice Loss for Broadband Telecommunication Network*, Springer Verlag, 1995.
[7] Zbigniew D., *ATM Resource Management*, McGraw Hill, 1997.

Received August 12, 2001

Institute of Information Technology