# Pollaczek–Khinchin formula for the M/G/1 queue in discrete time with vacations

W.C. Chan
T.-C. Lu
R.-J. Chen

*Indexing terms: Pollaczek–Khinchin formula, M/G/1 queue in discrete time*

**Abstract:** The continuous-time M/G/1 queue with vacations has been studied by many researchers. In the paper the authors report on an investigation of the discrete-time M/G/1 queue using Little's formula and conditional expectation. This direct approach can also be adopted to study the continuous-time case.

## 1 Introduction

Consider a discrete-time M/G/1 queue in which the server begins a vacation of random length each time that the system becomes empty. If the server returns from a vacation to find one or more customers waiting, the server works until the system empties, and then begins another vacation. If the server returns from a vacation to find no customers waiting, the server begins another vacation immediately. We assume that the lengths of vacations are independent and identically distributed (i.i.d.) random variables and are independent of the arrival process as well as the service times of customers.

The continuous-time model has been analysed in a number of papers [1–6]. A key result from these analyses is that the number of customers present in the system at a random point in time in equilibrium is the sum of two independent random variables: the number of Poisson arrivals during a time interval of the residual vacation time, and the number of customers present at a random point in time in equilibrium in the corresponding standard M/G/1 queue.

The purpose of this paper is to present a discrete-time model using a simple and direct method. The Pollaczek–Khinchin (P–K) mean-value formula, either in continuous-time or discrete-time models, can be derived using Little's formula and conditional expectation.

## 2 Discrete-time M/G/1 model

A discrete-time queueing system, such as an ATM network, is characterised by time-slotted and synchronous services [7]. The time axis is divided into equal intervals, called slots. Arriving customers that find the server busy wait in a queue. Servicing of customers is synchronised to start only at slot boundaries.

We will use the following definition for the discrete-time models. Departures take place only at slot boundaries. Without loss of generality, we normalise the length of a slot to unit time. Slots are sequentially numbered in nonnegative integers so that the $j$th slot is located in the time interval $[j-1, j)$, where $j = 1, 2, \ldots$ . The two time points immediately before and after time $j$ are denoted $j^-$ and $j^+$. A customer completing service in slot $j$ will leave the system at time $j^-$ and a customer starting service in slot $(j+1)$ will begin the service at time $j$.

The service times of customers are i.i.d. random variables with probability density function

$$p_k = P\{\text{service time equals } k \text{ slots}\} \quad k \geq 1$$

Now we define two discrete-time models for two different memoryless arrival processes.

*Definition 1*: Model A (bulk arrivals at slot boundaries)

Bulk arrivals occur at slot boundaries with a Poisson distributed size of mean $\lambda$ customers, that is,

$$f(k) = P\{k \text{ customers arrive at a slot boundary}\}$$
$$= \frac{\lambda^k e^{-\lambda}}{k!} \qquad k = 0, 1, 2, \ldots$$

*Definition 2*: Model B (exponential interarrival time)

Interarrival times are identical and exponentially distributed with mean $1/\lambda$ slots. The probability density function of the interarrival time is

$$f(x) = \lambda e^{-\lambda x} \qquad x \geq 0$$

Equivalently the number of arrivals in a slot has the same Poisson distribution as that of model A, However, customers in model B have to wait, on average, an additional half-slot more than that of customers in model A.

*Proposition 1*: The mean service time of customers in model B is a half-slot more than that of customers in model A.

*IEE Proceedings* online no. 19971225

Paper first received 13th June 1996 and in revised form 21st February 1997

W.C. Chan is with the Department of Electrical and Computer Engineering, The University of Calgary, Alberta, Canada T2N 1N4

R.J. Chen and T.C. Lu are with the Department of Computer Science and Information Engineering, National Chiao Tung University, Hsinchu, Taiwan

## 3 Residual service time in discrete-time model

Consider the test customer in model A who arrives at a discrete-time M/G/1 queueing system and finds the server busy. The time interval between the arrival epoch of this test customer and the completion of the service of the customer being served is known as the residual service time $R$ as shown in Fig. 1.

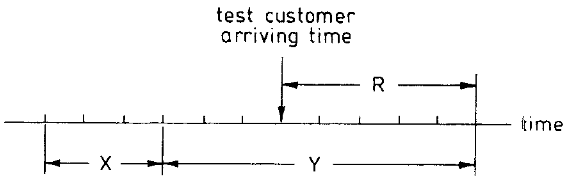

**Fig. 1** *General service time X, the special service time Y and the residual service time R ≤ Y*

From Fig. 1 we see that the test customer must wait at least an amount of time equal to the residual service time $R$. We shall determine the probability distributions of both the special service time $Y$ and the residual service time $R$ in terms of the distribution $p_k$ of the general service time $X$ and of its first and second moments ($m_1$ and $m_2$, respectively).

*Definition 3*: Define the following random variables:

$X$ = the general service time with probability distribution $\{p_k,\ k = 0,\ 1,\ \ldots\}$, the first moment $m_1$ and second moment $m_2$;

$Y$ = the special service time within which the test customer arrives; and

$R$ = the residual service time.

By definition, $p_k = P\{X = k\}$ is the probability that $X$ equals exactly $k$ slots in length. Now let

$$q_k = P\{Y = k\} \qquad k = 0, 1, \ldots$$

be the probability that $Y$ equals exactly $k$ slots in length.

*Proposition 2*: Given $\{p_k\}$, the probability distribution of $Y$ can be expressed as

$$q_k = \frac{kp_k}{m_1} \qquad k = 0, 1, \ldots$$

where $m_1$ is the first moment or mean of $X$.

*Proof*: Since the arrival process is Poisson, which is independent of the service process, the probability of the event $\{Y = k\}$ is proportional to $k$ and $p_k$. Thus $q_k$ can be expressed as the product of $kp_k$ multiplied by a proportional constant $c$,

$$q_k = ckp_k \qquad k = 0, 1, \ldots$$

where the right-hand side expresses the linear weighting with respect to the length of $Y$ and includes a constant $c$ which must be evaluated so as to properly normalise this probability distribution (see p.171 of [8]). It follows that

$$\sum_{k=0}^{\infty} q_k = c\sum_{k=0}^{\infty} kp_k = 1$$

Thus

$$c = \frac{1}{m_1}$$

where $m_1 = \Sigma_{k=0}^{\infty}\ kp_k$, is the mean of $X$.

*Proposition 3*: Let $r_j$ denote the probability that the residual service time $R$ equals exactly $k$ slots in length. Then

$$\begin{aligned} r_j &= P\{R = j\} \\ &= \frac{P\{X > j\}}{m_1} \qquad j = 0, 1, \ldots \end{aligned}$$

*Proof*: Suppose that the special service time $Y$ is exactly $k$ slots in length and that the arrival process is Poisson. The arrival epoch of the test customer occurs at any one of the $k$ time points of $Y$ with equal probability 1/$k$. In other words, given that $Y = k$ the conditional probability of $R$ being exactly $j$ slots in length is

$$P\{R = j | Y = k\} = \frac{1}{k} \qquad j = 0, 1, \ldots, k - 1$$

It follows that the joint probability

$$\begin{aligned} P\{R = j, Y = k\} &= P\{R = j | Y = k\}P\{Y = k\} \\ &= \frac{q_k}{k} \\ &= \frac{p_k}{m_1} \end{aligned}$$

Therefore

$$\begin{aligned} r_j &= P\{R = j\} \\ &= \sum_{k=j+1}^{\infty} P\{R = j, Y = k\} \\ &= \sum_{k=j+1}^{\infty} \frac{p_k}{m_1} \\ &= \frac{P\{X > j\}}{m_1} \qquad j = 0, 1, \ldots \end{aligned}$$

Applying the results of proposition 2 and proposition 3 we obtain the mean value of the residual time as follows:

*Theorem 1*: The mean residual service time $\overline{R}_A$ for model A is given by

$$\overline{R}_{\text{A}} = \frac{m_2 - m_1}{2m_1}$$

where $m_1$, $m_2$ are the first and second moments of $X$, respectively.

*Proof*: By definition, we write

$$\begin{aligned} \overline{R}_{\text{A}} &= \sum_{j=1}^{\infty} jr_j \\ &= \frac{1}{m_1}\sum_{j=0}^{\infty} jP\{X > j\} \qquad \text{proposition 2} \\ &= \frac{1}{m_1}\sum_{j=0}^{\infty} j \sum_{k=j+1}^{\infty} p_k \\ &= \frac{1}{m_1}\sum_{k=1}^{\infty} p_k \sum_{j=0}^{k-1} j \\ &= \frac{1}{m_1}\sum_{k=1}^{\infty} p_k \frac{k(k-1)}{2} \\ &= \frac{1}{2m_1}\left[\sum_{k=1}^{\infty} k^2 p_k - \sum_{k=1}^{\infty} kp_k\right] \\ &= \frac{m_2 - m_1}{2m_1} \end{aligned}$$

Applying the results of proposition 1 and theorem 1, we obtain the mean residual service time for model B as follows:

*Theorem 2*: The mean residual service time $\overline{R}_B$ for the discrete-time model B is given by

$$\overline{R}_{\text{B}} = \frac{m_2}{2m_1}$$

where $m_1$, $m_2$ are the first and second moments of $X$, respectively.

## 4 Pollaczek–Khinchin mean-value formula for the M/G/1 queue in discrete time

We use the conditional expectation and Little's formula to derive the P–K formula. Model B is adopted here because its mean residual service time has the same form as the continuous-time model. The method developed in this Section is also useful for deriving the P–K formula in continuous time.

Depending on the state of the server, the waiting time of an arriving customer can have two different values. Let us define the following events:

$B$ = the arriving customer finds the server busy and $N_q$ customers waiting in the queue;

$I$ = the arriving customer finds the server idle.

Let $W$ denote the waiting time. The average waiting time of a customer can be expressed as

$$E[W] = \rho E[W|B] + (1-\rho)E[W|I] \qquad (1)$$

where $\rho = \lambda E[X]$ is the probability that the server is busy.

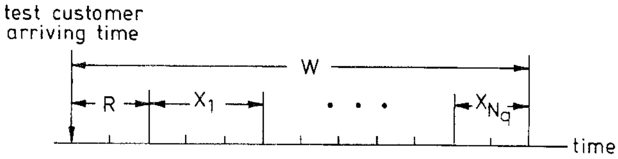

**Fig. 2** *Waiting time for the test customer*

For the event $B$, we have the waiting time as shown in Fig. 2 and

$$\begin{aligned} E[W|B] &= E[R + X_1 + X_2 + \cdots + X_{N_q}|B] \\ &= E[R|B] + E[X_1 + X_2 + \cdots + X_{N_q}|B] \\ &= E[R|B] + E[X|B]E[N_q|B] \\ &= E[R|B] + E[X]E[N_q]/\rho \\ &= E[R|B] + E[W] \end{aligned}$$

Thus we find

$$E[W|B] = E[R|B] + E[W]$$

where $X_{N_q} = 0$ if $N_q = 0$ and

$$E[X_{N_q}|B] = E[X] = m_1$$
$$E[N_q|B] = E[N_q]\rho = \lambda E[W]/\rho$$

have been used. Note that $X$ and $B$ are independent random variables.

For the event $I$, the waiting time becomes zero because the arriving customer is served immediately. Thus

$$E[W|I] = 0$$

and hence the second term on the right-hand side of eqn. 1 vanishes. Therefore, we obtain from eqn. 1

$$E[W] = \rho(E[R|B] + E[W])$$

or

$$E[W] = \frac{\rho E[R|B]}{1-\rho} \qquad (2)$$

Substituting the average residual service time $\overline{R}_B$ for $E[R|B]$ in this expression yields the desired Pollaczek–Khinchin formula.

$$E[W] = \frac{\rho E[R|B]}{1-\rho} = \frac{\rho}{1-\rho} \times \frac{m_2}{2m_1} = \frac{\lambda m_2}{2(1-\rho)} \qquad (3)$$

To calculate the mean number of customers in the system, we have to obtain the total system time $T$, which is the sum of the waiting time and the service time. It follows that

$$E[T] = E[W] + E[X]$$

Let $N$ be the mean number of customers in the system. By Little's formula, we have

$$\begin{aligned} N &= \lambda E[T] \\ &= \lambda\left[\frac{\lambda m_2}{2(1-\rho)} + m_1\right] \\ &= \rho + \frac{\lambda^2 m_2}{2(1-\rho)} \qquad (4) \end{aligned}$$

### *4.1 Mean waiting time in the M/G/1 queue with vacations*

Generally a queueing system will have busy periods with at least one customer present and idle periods with no customer present. A busy period is a time interval that begins when an arriving customer finds the system empty. An idle period is the period between two successive busy periods. Clearly, busy and idle periods occur alternatively and form a cycle.

Suppose that at the end of each busy period, the server goes on vacation for a random interval of time with first moment $v_l$ and second moment $v_2$. For computer communication networks, vacations correspond to transmissions of various kinds of control and record-keeping packets when there is little traffic or when the transmit queue is empty.

We shall derive an expression for the average waiting time in the M/G/1 queue with vacations. In this case the second term on the right-hand side of eqn. 1 is no longer zero and is simply the delay caused by the residual vacation interval equal to $(1 - \rho)v_2/2v_1$. Now from eqns. 1 and 2 and theorem 2 we obtain

$$E[W] = \rho\left[\frac{m_2}{2m_1} + E[W]\right] + (1-\rho)\frac{v_2}{2v_1}$$

Solving for $E[W]$ yields the mean waiting time for model $B$:

$$E[W] = \frac{\lambda m_2}{2(1-\rho)} + \frac{v_2}{2v_1} \qquad (5)$$

This expression may be regarded as an extension of the Pollaczek–Khinchin formula (eqn. 3) for the mean waiting time in the M/G/1 queue in discrete time to the case with vacations.

## 5 Application

Consider a polling system for transmission of cells from several cell-based streams into a statistical multiplexing system where the cell size for each stream is constant. This situation arises often in multiaccess channels.

As a typical example, we consider a communication channel that can be accessed by $m$ spatially separated stations. These $m$ stations are connected by cables in a unidirectional loop. Each station transmits the backlogged packets when it is polled. The interpolling times can be regarded as vacations. A station transmits packets when it receives the poll.

We will consider $m$ homogeneous traffic streams. Each consists of three Poisson arrival streams with rates $\lambda_v$, $\lambda_a$ and $\lambda_d$ which correspond to the arrival rate of video, audio and data packet streams, respectively. For simplicity, we assume the lengths of these packets are $l_v$, $l_a$ and $l_d$ cells for each type of stream (Fig. 3). The total packet arrival rate of the combined stream is

$$\lambda_\mathrm{s} = \lambda_\mathrm{a} + \lambda_\mathrm{v} + \lambda_\mathrm{d}$$

and the total cell arrival rate of the combined stream is

$$\lambda_\mathrm{c} = \lambda_\mathrm{a} l_\mathrm{a} + \lambda_\mathrm{v} l_\mathrm{v} + \lambda_\mathrm{d} l_\mathrm{d}$$

The total packet arrival rate of all the $m$ streams is then $m\lambda_s$.

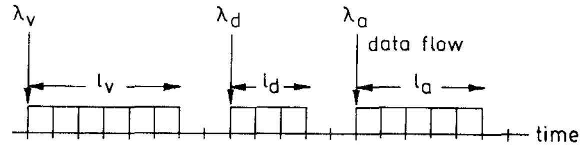

**Fig. 3** *Input packets for a single stream*

Consider the network as an M/G/1 queue. The mean residual service time is given by

$$\begin{aligned}\overline{R} &= \frac{\lambda_\mathrm{v} l_\mathrm{v}}{\lambda_\mathrm{v} l_\mathrm{v} + \lambda_\mathrm{a} l_\mathrm{a} + \lambda_\mathrm{d} l_\mathrm{d}} \times \frac{l_\mathrm{v}^2}{2l_\mathrm{v}} \\ &+ \frac{\lambda_\mathrm{a} l_\mathrm{a}}{\lambda_\mathrm{v} l_\mathrm{v} + \lambda_\mathrm{a} l_\mathrm{a} + \lambda_\mathrm{d} l_\mathrm{d}} \times \frac{l_\mathrm{a}^2}{2l_\mathrm{a}} \\ &+ \frac{\lambda_\mathrm{d} l_\mathrm{d}}{\lambda_\mathrm{v} l_\mathrm{v} + \lambda_\mathrm{a} l_\mathrm{a} + \lambda_\mathrm{d} l_\mathrm{d}} \times \frac{l_\mathrm{d}^2}{2l_\mathrm{d}} \\ &= \frac{\lambda_\mathrm{v} l_\mathrm{v}^2 + \lambda_\mathrm{a} l_\mathrm{a}^2 + \lambda_\mathrm{d} l_\mathrm{d}^2}{2(\lambda_\mathrm{v} l_\mathrm{v} + \lambda_\mathrm{a} l_\mathrm{a} + \lambda_\mathrm{d} l_\mathrm{d})} \\ &= \frac{m_2}{2m_1}\end{aligned}$$

where

$$m_1 = \frac{\lambda_\mathrm{v} l_\mathrm{v}}{\lambda_\mathrm{v} + \lambda_\mathrm{a} + \lambda_\mathrm{d}} + \frac{\lambda_\mathrm{a} l_\mathrm{a}}{\lambda_\mathrm{v} + \lambda_\mathrm{a} + \lambda_\mathrm{d}} + \frac{\lambda_\mathrm{d} l_\mathrm{d}}{\lambda_\mathrm{v} + \lambda_\mathrm{a} + \lambda_\mathrm{d}}$$

and

$$m_2 = \frac{\lambda_\mathrm{v} l_\mathrm{v}^2}{\lambda_\mathrm{v} + \lambda_\mathrm{a} + \lambda_\mathrm{d}} + \frac{\lambda_\mathrm{a} l_\mathrm{a}^2}{\lambda_\mathrm{v} + \lambda_\mathrm{a} + \lambda_\mathrm{d}} + \frac{\lambda_\mathrm{d} l_\mathrm{d}^2}{\lambda_\mathrm{v} + \lambda_\mathrm{a} + \lambda_\mathrm{d}}$$

Now we calculate the mean vacation time $v_1$ which is the sum of the other $m - 1$ mean busy periods and $m$ token releasing time $\tau$. Initially, there are $\lambda_s v_1$ packets in the M/G/1 queue. Consider the successive arrivals for a last-in first-out (LIFO) queueing discipline [8]. The mean busy period is equal to $\lambda_s v_1 B_{\mathrm{M/G/1}}$, where $B_{\mathrm{M/G/1}}$ is the mean busy period in the M/G/1 queue. From [[8], p. 233] we have

$$B_{\mathrm{M/G/1}} = \frac{E[X]}{1-\rho}$$

Then the mean vacation time can be expressed as

$$v_1 = (m-1)\lambda_\mathrm{s} v_1 B_{\mathrm{M/G/1}} + m\tau$$

or

$$\begin{aligned}v_1 &= \frac{m\tau}{1-(m-1)\lambda_\mathrm{s} B_{\mathrm{M/G/1}}} \\ &= \frac{m\tau}{1-(m-1)\frac{\rho}{1-\rho}} \\ &= \frac{m\tau(1-\rho)}{1-m\rho}\end{aligned}$$

where $\rho$ is the utilisation factor of the M/G/1 queue with arrival rate $\lambda_s$. For exponentially distributed vacation times with mean $v_1$, we have $v_2 = 2v_1^2$. From eqn. 5 we obtain the mean waiting time

$$\begin{aligned}E[W] &= \frac{\lambda_\mathrm{s} m_2}{2(1-\rho)} + \frac{v_2}{2v_1} \\ &= \frac{\lambda_\mathrm{s}\left(\frac{\lambda_\mathrm{v} l_\mathrm{v}^2}{\lambda_\mathrm{v}+\lambda_\mathrm{a}+\lambda_\mathrm{d}} + \frac{\lambda_\mathrm{a} l_\mathrm{a}^2}{\lambda_\mathrm{v}+\lambda_\mathrm{a}+\lambda_\mathrm{d}} + \frac{\lambda_\mathrm{d} l_\mathrm{d}^2}{\lambda_\mathrm{v}+\lambda_\mathrm{a}+\lambda_\mathrm{d}}\right)}{2(1-\rho)} \\ &+ \frac{m\tau(1-\rho)}{(1-m\rho)} \\ &= \frac{\lambda_\mathrm{v} l_\mathrm{v}^2 + \lambda_\mathrm{a} l_\mathrm{a}^2 + \lambda_\mathrm{d} l_\mathrm{d}^2}{2(1-\rho)} + \frac{m\tau(1-\rho)}{(1-m\rho)}\end{aligned}$$

## 6 Discussion

It is interesting to note that if model A is adopted, then using similar arguments as in Section 4 would result in the following expressions:

The Pollaczek–Khinchin formula in discrete time becomes

$$\begin{aligned}E[W] &= \frac{\rho}{1-\rho}\overline{R}_\mathrm{A} \\ &= \frac{\lambda(m_2-m_1)}{2(1-\rho)} \end{aligned} \tag{6}$$

The mean number of customers in the system is

$$N = \rho + \frac{\lambda^2(m_2-m_1)}{2(1-\rho)} \tag{7}$$

and the mean waiting time in the M/G/1 queue with vacations now becomes

$$E[W] = \frac{\lambda(m_2-m_1)}{2(1-\rho)} + \frac{v_2}{2v_1} \tag{8}$$

The results of eqns. 6, 7 and 8 for model A correspond to eqns. 3, 4 and 5 for model B, respectively.

## 7 Conclusions

In the past, most queueing analysis has been based on queueing phenomena in continuous time. Recently in the telecommunication industries, B-ISDN (Broadband Integrated Services Digital Network) has received considerable attention. B-ISDN can provide a common interface for future communication including the transmission of video, data and speech. Since information in B-ISDN is transported by means of discrete units of 53-octet ATM (Asynchronous Transfer Mode) cells, it appears that the analysis of a queueing system in discrete time is more natural.

In this paper, we derive the Pollaczek–Khinchin formula for the M/G/1 queue in discrete time with vacations. We simply use Little's formula and conditional expectation. This mean value analysis provides an alternative method of deriving performance measures for either continuous-time or discrete-time M/G/1 queues.

## 8 Acknowledgment

The authors wish to thank the referees for their careful reviews and constructive comments which improved the quality of the paper.

## 9 References

1 COOPER, R.B.: 'Queues served in cyclic order: Waiting times', *Bell Syst. Tech. J.*, 1970, **49**, pp. 399–413

2 LEVY, Y., and YECHIALI, U.: 'Utilization of idle time in an M/G/1 queueing system', *Manage. Sci.*, 1975, **22**, pp. 202–211

3 HEYMAN, D.P.: 'The T-policy for the M/G/1 queue', *Manage. Sci.*, 1977, **23**, pp. 775–778

4 SCHOLL, M., and KLEINROCK, L.: 'On the M/G/1 queue with rest period and certain service-independent queueing disciplines', *Oper. Res.*, 1983, **31**, pp. 705–719

5 FUHRMANN, S.W.: 'A note on the M/G/1 queue with server vacations', *Oper. Res.*, 1984, **32**, (6), pp. 1368–1373

6 FUHRMANN, S.W., and COOPER, R.B.: 'Stochastic decompositions in the M/G/1 queues with generalized vacations', *Oper. Res.*, 1985, **33**, (5), pp. 117–1129

7 BRUNEEL, H., and KIM, B.: 'Discrete-time models for communication systems including ATM' (Kluwer Academic, Boston, 1993)

8 KLEINROCK, L.: 'Queueing systems, vol. 1: theory' (John Wiley & Sons, New York, 1975)