Waiting time distribution for the *M/M/m* queue

W.-C. Chan and Y.-B. Lin

Abstract: A novel method is presented for the calculation of the waiting time distribution function for the $M/M/m$ queue. It is shown that the conditional waiting time obeys an Erlang distribution with rate $m\mu$, where μ is the service rate of a server. An explicit closed form solution is obtained by means of the probability density function of the Erlang distribution. The derivation of the result proved to be very simple. The significance of Khintchine's method and its close relation to the proposed method is pointed out. It is also shown that the waiting time distribution can be obtained from Takacs's waiting time distribution for the $G/M/m$ queue as a special case. This reveals some insight into the significance of Takacs's more general, but rather complex, result.

1 Introduction

Consider a queueing system with an unlimited waiting room and m servers. Suppose that customers arrive at the queueing system at the instants $\tau_0, \tau_1, \ldots, \tau_n, \ldots$, where by convention $\tau_0 = 0$, and for $n = 1, 2, 3, \ldots$, the interarrival times $\tau_{n+1} - \tau_n$ are identically distributed, independent, positive random variables with distribution function

$$\Pr[\tau_{n+1} - \tau_n \leq t] = F(t) = \begin{cases} 1 - e^{-\lambda t}, & \text{if } t \geq 0 \\ 0, & \text{if } t < 0 \end{cases} \quad (1)$$

Assume that the system is work conserving, i.e. there is no idle server if there is a waiting customer. Customers are served in their order of arrival and the service times are identically distributed, independent random variables with the distribution function

$$H(t) = \begin{cases} 1 - e^{-\mu t}, & \text{if } t \geq 0 \\ 0, & \text{if } t < 0 \end{cases} \quad (2)$$

which is independent of $\{\tau_n\}$. This queueing system is known as the Erlang delay system which has been investigated intensively in the literature [1–5]. The main concern of the waiting time is its stochastic behaviour and the determination of its distribution function. Khintchine [2] presented a very thoughtful approach and obtained an analytical closed form expression for the waiting time distribution function. This paper presents a simple and novel alternative method for the determination of the waiting time distribution function, which gives new insight into the waiting time for the Erlang delay system. For a more general study of the $G/G/m$ queue, the reader is referred to [6–13]. Note that in order to study the properties of the associated queueing process, these $G/G/m$ studies all resulted in having the problem of solving integral equations, and no analytic closed form solutions such as Takacs's for the $G/M/s$ queue were obtained.

IEE Proceedings online no. 20030274

doi:10.1049/ip-com:20030274

Paper first received 30th August 2002 and in revised form 7th January 2003

The authors are with the Department of Computer Science & Information Engineering, National Chiao Tung University, 1001 Ta Hsueh Road, Taiwan 30050, Republic of China

2 Some fundamental results

Since customers are served in the order of their arrival, the investigation of the stochastic behaviour of the waiting time can be reduced to that of the state of the system. We shall summarise some well known results that are needed for the investigation of waiting time.

2.1 Stationary probabilities for states of the M/M/m queue

Denote by $N(t)$ the total number of customers waiting or being served in the system at the instant t. We say that the system is in state k at the instant t if $N(t) = k$. Further, let

$$P_k(t) = \Pr[N(t) = k]$$

denote the probability that the system is in state k at the instant t. If $\lambda < m\mu$, then the limit

$$\lim_{t \to \infty} P_k(t) = p_k, \quad \text{for } k = 0, 1, 2, \ldots$$

always exists and is independent of the initial distribution $\{P_k(0)\}, k = 0, 1, 2, \ldots$ If the limiting distribution $\{p_k\}$ exists, then it is uniquely determined by the following system of linear equations

$$p_k = \sum_{j=k-1}^{\infty} p_{jk} p_j, \quad \text{where } p_{jk} = \int_{t=0}^{\infty} \pi_{jk}(t) dF(t)$$

and $\pi_{jk}(x)$ is the transition probability under the condition that the inter-arrival time is x. We shall quote the fundamental results for the limiting distribution $\{p_k\}$, $k = 0, 1, 2, \ldots$, as follows [2, 3, 4]:

$$p_k = \begin{cases} \left(\dfrac{a^k}{k!}\right) p_0, & \text{if } k \leq m \\ \left(\dfrac{a^k}{m! m^{k-m}}\right) p_0, & \text{if } k \geq m \end{cases} \quad (3)$$

and

$$p_0 = \left[\sum_{k=0}^{m-1} \frac{a^k}{k!} + \frac{a^m}{(m-1)!(m-a)} \right]^{-1} \quad (4)$$

where $a = \lambda/\mu$ is the offered traffic or traffic intensity.

2.2 Erlang distribution

The Erlang distribution of order k with parameter μ has the distribution function

$$E_k(x) = \begin{cases} 1 - \sum_{j=0}^{k-1} \left[\frac{(\mu x)^j}{j!}\right] e^{-\mu x}, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0 \end{cases} \quad (5)$$

whose probability density function is

$$e_k(x) = \left[\frac{\mu^k x^{k-1}}{(k-1)!}\right] e^{-\mu x}, \quad \text{where } x \geq 0, k \geq 1 \quad (6)$$

When $k = 1$, this probability density function reduces to the negative exponential density function. The Erlang random variable in (5) may be considered as a service time which is the sum of k independent random variables, each of which has an exponential distribution defined by (2). This service time was employed by Erlang in his method of stages. It is important to point out that the exponential distribution has the memoryless property. This means that the distribution of the remaining time, i.e. the residual time, for an exponentially distributed random variable is independent of the age of that random variable. This memoryless property plays an important role in the determination of the waiting time distribution function for $M/M/m$ and $G/M/m$ queueing systems.

3 The waiting time distribution function

Let W denote the waiting time. It would be more convenient and simpler to compute the probability $\Pr[W > t]$ than $\Pr[W \leq t]$, where $\Pr[W > t]$ denotes the probability that a test customer entering the system at a random moment has a waiting time greater than t. Furthermore, let $\Pr[W > t | N = k]$ be the probability of the event $\{W > t\}$ on the condition that on arrival the test customer finds the system in state k. Using the formula of total probability, we can write

$$\Pr[W > t] = \sum_{k=0}^{\infty} \pi_k \Pr[W > t | N = k] \quad (7)$$

where π_k is the probability that the system is in state k just prior to the arrival instant of the test customer. $\{\pi_k\}$, for $k = 0, 1, 2, \ldots$, is known as the arriving customer's distribution. Since the arrival process is Poisson, the arriving customer's distribution $\{\pi_k\}$ and the outside observer's distribution $\{p_k\}$ are equal. Then we have [3]

$$\pi_k = p_k, \quad k = 0, 1, 2, \ldots \quad (8)$$

Equation (8) implies the PASTA (Poisson arrivals see time averages) property.

Since $\Pr[W > t | N = k] = 0$ for $k < m$ and $t \geq 0$, (7) reduces to

$$\Pr[W > t] = \sum_{k=m}^{\infty} p_k \Pr[W > t | N = k] \quad (9)$$

It remains to determine the conditional waiting probabilities $\Pr[W > t | N = k]$ for $k \geq m$ because p_k is given by (3). Observe that during the waiting period of the test customer all m servers must be busy. Any one of these m busy servers can contribute a service rate μ. The resultant service rate of the m servers as a group is essentially $m\mu$. In other words, the whole group of m busy servers acts like a single server with an exponential service time of rate $m\mu$. At this point, $k-m+1$ customers are waiting for service in the system, and each of them will take an exponential service time of mean $1/m\mu$. It follows from the memoryless property of the exponential distribution that the waiting time of the test customer is composed of $k-m+1$ exponential service times, each of which has an exponential distribution with rate $m\mu$. Thus we can write the conditional waiting time as

$$W^*_{k-m+1} = R + \sum_{j=1}^{k-m} W_j, \quad k \geq m \quad (10)$$

where R is the residual (exponential) service time of the group of m customers being served on the arrival of the test customer, and W_j is the waiting time of the jth customer in the queue. Clearly, the conditional waiting time W^*_{k-m+1} can be regarded as an Erlang random variable with the following probability density function:

$$e_{k-m+1}(x) = \left[\frac{(m\mu)^{k-m+1} x^{k-m}}{(k-m)!}\right] e^{-m\mu x}, \quad x \geq 0, k \geq m \quad (11)$$

Then the conditional waiting time distribution is simply

$$\Pr[W > t | N = k] = \int_{x=t}^{\infty} e_{k-m+1}(x) dx = \int_{x=t}^{\infty} \left[\frac{(m\mu)^{k-m+1} x^{k-m}}{(k-m)!}\right] e^{-m\mu x} dx \quad (12)$$

Substituting (3) and (12) into (9) results in the waiting time distribution function

$$\begin{aligned} \Pr[W > t] &= \sum_{k=m}^{\infty} \left(\frac{a^k}{m! m^{k-m}}\right) p_0 \\ &\quad \times \int_{x=t}^{\infty} \left[\frac{(m\mu)^{k-m+1} x^{k-m}}{(k-m)!}\right] e^{-m\mu x} dx \\ &= m\mu p_m \int_{x=t}^{\infty} e^{-(m\mu - \lambda)x} dx \\ &= \left(\frac{m p_m}{m-a}\right) e^{-(m\mu - \lambda)t}, \quad t \geq 0 \end{aligned} \quad (13)$$

which is the desired solution.

4 Remarks

The simplicity of the integration in (13) results from the infinite sum, which yields an exponential function. If the $M/M/m$ queueing system has only a finite waiting room, then the upper limit of the summation becomes finite and no exponential function results. In this case, it would be simpler to use the distribution of (5) for $\Pr[W > t | N = k]$.

4.1 Khintchine's explanation

Khintchine [2] noticed that if $k \geq m$, there must be $k-m$ customers waiting for service in front of the test customer in the queue. If the queue discipline is first-come, first-served, then the arriving test customer finding $k-m$ customers waiting in front of him will obtain service after the $(k-m+1)$th departure. Thus the conditional probability $\Pr[W > t | N = k]$ is equal to the probability that during the time interval t after the arrival of the test customer, there will be at most $k-m$ departures of customers. It follows that

$$\Pr[W > t | N = k] = \sum_{j=0}^{k-m} f_j(t), \quad k \geq m \quad (14)$$

where $f_j(t)$ denotes the probability of exactly j departures in the time interval t, which remains to be determined. Since the service times are exponential, the probability that no departures occur during the time interval t from the arrival

instant of the test customer is equal to

$$f_0(t) = e^{-m\mu t}$$

which implies that the inter-departure times have exponential distribution with rate $m\mu$, and hence equivalently the number of departures during the time interval t obeys a Poisson process with rate $m\mu$. Therefore, we have

$$f_j(t) = \left[\frac{(m\mu t)^j}{j!}\right] e^{-m\mu t}, \quad j = 0, 1, 2, \ldots \tag{15}$$

This result indicates that when all m servers are busy in the time interval t, the process of departure follows a Poisson process with rate $m\mu$. Substituting (15) into (14) yields the conditional probability

$$\Pr[W > t | N = k] = e^{-m\mu t}\left[\sum_{j=0}^{k-m} \frac{(m\mu t)^j}{j!}\right], \quad k \geq m \tag{16}$$

which is the complementary distribution function of the Erlang distribution of order $k-m+1$ and parameter $m\mu$. Using (3) and (16) in (9), we find

$$\begin{aligned}\Pr[W > t] &= \left(\frac{e^{-m\mu t}}{m!m^{-m}}\right) p_0 \left\{\sum_{k=m}^{\infty}\sum_{j=0}^{k-m} \left(\frac{a}{m}\right)^k \left[\frac{(m\mu t)^j}{j!}\right]\right\} \\ &= \left(\frac{e^{-m\mu t}}{m!m^{-m}}\right) p_0 \left\{\sum_{j=0}^{\infty}\sum_{k=m+j}^{\infty} \left(\frac{a}{m}\right)^k \left[\frac{(m\mu t)^j}{j!}\right]\right\}\end{aligned} \tag{17}$$

Let $i = k-m-j$. Then (17) can be rewritten as

$$\begin{aligned}\Pr[W > t] &= \left(\frac{e^{-m\mu t}}{m!m^{-m}}\right) p_0 \\ &\times \left\{\sum_{j=0}^{\infty}\sum_{i=0}^{\infty} \left(\frac{a}{m}\right)^m \left(\frac{a}{m}\right)^i \left[\left(\frac{a}{m}\right)^j \frac{(m\mu t)^j}{j!}\right]\right\} \\ &= \left(\frac{e^{-m\mu t}a^m}{m!}\right) p_0 \left[\sum_{j=0}^{\infty} \frac{(\lambda t)^j}{j!}\right]\left[\sum_{i=0}^{\infty} \left(\frac{a}{m}\right)^i\right] \\ &= \left(\frac{mp_m}{m-a}\right) e^{-(m\mu-\lambda)t}, \quad t \geq 0\end{aligned} \tag{18}$$

It is interesting to note that Khintchine's method essentially makes use of the complementary Erlang distribution function of (16) to calculate the conditional probability

$$\Pr[W > t | N = k] = 1 - E_{k-m+1}(t) \tag{19}$$

4.2 Takacs's waiting time distribution for Palm input and exponential service times for the G/M/m queue

Takacs investigated the multiple server queueing process for Palm input and exponential service times. He obtained a more general but very complex formula for the waiting time distribution [14]. We quote his formula as follows:

$$\Pr[W \leq t] = 1 - \frac{Ae^{-m\mu(1-\omega)t}}{1-\omega}, \quad t \geq 0 \tag{20}$$

where ω is the only root of the functional equation

$$\omega = \phi(m\mu(1-\omega)) \tag{21}$$

in the unit circle, where

$$\phi(s) = \int_{t=0}^{\infty} e^{-st} dF(t) \tag{22}$$

and A is given by

$$A = \left\{\frac{1}{1-\omega} + \sum_{j=1}^{m}\left[\frac{\binom{m}{j}}{C_j(1-\phi_j)}\right]\left[\frac{m(1-\phi_j)-j}{m(1-\omega)-j}\right]\right\}^{-1} \tag{23}$$

and for $j = 0, 1, 2, \ldots,$

$$\phi_j = \phi(j\mu), \quad \text{and} \quad C_j = \prod_{i=1}^{j}\left(\frac{\phi_i}{1-\phi_i}\right)$$

From (20), we deduce the following formula:

$$\Pr[W > t] = \frac{Ae^{-m\mu(1-\omega)t}}{1-\omega}, \quad t \geq 0 \tag{24}$$

In principle, it is possible to obtain the waiting time distribution function (18) from (20) as special case when $F(t)$ is defined by (1). However, this task is not trivial. It is instructive to show that this is indeed the case. In this case, we have found that

$$\begin{aligned}&\phi(s) = \frac{\lambda}{s+\lambda}, \quad \omega = \frac{\lambda}{m\mu} = \frac{a}{m}, \quad C_j = \frac{a^j}{j!}, \\ &\text{and} \quad \left(\frac{1}{1-\phi_j}\right)\left[\frac{m(1-\phi_j)-j}{m(1-\omega)-j}\right] = 1\end{aligned} \tag{25}$$

Substituting (25) into (23), we have

$$\begin{aligned}A &= \left[\frac{m}{m-a} + \sum_{i=0}^{m-1}\left(\frac{m!}{a^m}\right)\left(\frac{a^i}{i!}\right)\right]^{-1} \\ &= \left(\frac{a^m}{m!}\right) p_0 \\ &= p_m\end{aligned} \tag{26}$$

From (25) and (26), then (24) can be rewritten as

$$\Pr[W > t] = \left(\frac{mp_m}{m-a}\right) e^{-(m\mu-\lambda)t}$$

as expected.

5 Conclusions

A novel method has been presented for the calculation of the waiting time distribution function $\Pr[W > t]$ for the $M/M/m$ queue. Although the result given by (13) is not new, the method of deriving the result is novel, and provides new insight into the conditional waiting time that has a complementary Erlang distribution of order $k-m+1$ and parameter $m\mu$. Also we have presented two alternative methods to obtain the same result. Khintchine's method is very logical and thoughtful, and offers a physical interpretation of the departure process of customers from a group of m busy servers. It is interesting to note that Khintchine's method implicitly makes use of the complementary Erlang distribution defined in (19).

It has also been shown that given Takacs's formula for the $G/M/m$ queue in (20), the same result (18) can be obtained as a special case. Since (20) is very complex, it appears that derivation of (13) from (20) is not a trivial task. To our knowledge, this task has not been carried out and hence is not available in the literature in the way presented in this paper. Also, this approach provides a much simpler conditional waiting time expression for use in performance modelling of wireless telephone networks [15].

6 Acknowledgments

The authors would like to thank the anonymous reviewers. Their comments have significantly improved the quality of this paper. This work was sponsored in part by MOE Program for Promoting Academic Excellence of Universities under the grant number 89-E-FA04-1-4, IIS, Academia Sinica, FarEastone, the Lee and MTI Center for Networking Research, NCTU, and National Science Council under contract NSC 90-2213-E-009-156.

7 References

1 MEDHI, J.: 'Stochastic models in queueing theory' (Academic Press, 1991)
2 KHINTCHINE, A.Y.: 'Mathematical methods in the theory of queueing' (Hafner, New York, 1969, 2nd edn.), (English translation from Russian)
3 COOPER, R.B.: 'Introduction to queueing theory' (Elsevier Science, New York, 1981, 2nd edn.)
4 KLEINROCK, L.: 'Queueing systems: Volume I-Theory', (Wiley, New York, 1976)
5 GROSS, D., and HARRIS, C. H.: 'Fundamentals of queueing theory' (John Wiley, 1998)
6 KENDALL, D.: 'Stochastic processes occuring in the theory of queues and their analysis by the method of imbedded Markov chains', *Ann. Math. Stat.*, 1954, **24**, pp. 338–354
7 KIEFER, J., and WOLFOWITZ, J.: 'On the theory of queues with many servers', *Trans. Am. Math. Soc.*, 1955, **78**, pp. 1–18
8 KARLIN, S., and McGREGOR, J.: 'The differential equations of birth-and-death processes and the Stieltjes moment problem', *Trans. Am. Math. Soc.*, 1957, **85**, pp. 489–546
9 KARLIN, S., and McGREGOR, J.: 'The classification of birth-and-death processes', *Trans. Am. Math. Soc.*, 1957, **86**, pp. 366–400
10 KARLIN, S., and McGREGOR, J.: 'Many server queueing processes with Poisson input and exponential service times', *Pac. J. Math.*, 1958, **8**, pp. 87–118
11 PRESMAN, E.: 'On the waiting time for many server queueing systems', *Theory Probab. Appl.*, 1965, **10**, pp. 63–73
12 DE SMIT, J.H.A.: 'Some general results for many server queues', *Adv. Appl. Probab.*, 1973, **5**, pp. 153–169
13 DE SMIT, J.H.A.: 'On the many server queue with exponential service times', *Adv. Appl. Probab.*, 1973, **5**, pp. 170–182
14 TAKACS, L.: 'Introduction to the theory of queues' (Greenwood Press, 1961)
15 TSAI, H.M., and LIN, Y.-B.: 'Modeling wireless local loop with general call holding times and finite number of subscribers', *IEEE Trans. Comput.*, 2002, **51**, (7), pp. 775–786