Expected lateness in an M/M/1/K queue

Abduh Sayid Albana, Ramzi Hammami, Yannick Frein

▶ **To cite this version:**

Abduh Sayid Albana, Ramzi Hammami, Yannick Frein. Expected lateness in an M/M/1/K queue. 2017. hal-01626006

HAL Id: hal-01626006

https://hal.univ-grenoble-alpes.fr/hal-01626006

Preprint submitted on 31 Oct 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Expected lateness in an M/M/1/K queue

Abduh Sayid Albana [1], Ramzi Hammami [2*], Yannick Frein [1]

[1] Univ. Grenoble Alpes, CNRS, Grenoble INP[+], G-SCOP, F-38000 Grenoble, France

[2] Rennes School of Business, R-SCOM research center, Rennes, France

{abduh-sayid.albana@grenoble-inp.fr, yannick.frein@g-scop.grenoble-inp.fr, hammami.ramzi@gmail.com}

* Corresponding author

Abstract - We calculate the expected lateness for a late job in an M/M/1/K queue (i.e., the expected waiting time in the system after a threshold lead time *l*, given that the waiting time is greater than *l*). The expected lateness for any given job can then be deduced immediately. Applications of our result include the calculation of the lateness penalty cost when a given lead time is quoted to the customers, such as in call centers or in some production systems.

Keywords: queuing theory; M/M/1/K; quoted lead time; expected lateness; lateness penalty cost.

1. Introduction

We consider an M/M/1/K queue, i.e., a queue in a system having a single server where the job service times have an exponential distribution and the arrival process of customers is determined by a Poisson process, and where the system has a finite batch size K. We provide a new performance that is not known in the literature for such a queue. Indeed, we calculate the expected lateness given that a job is late in an M/M/1/K queue.

Suppose that a company, modelled as an M/M/1/K queue, quotes the lead time *l* to its customers. A job is considered to be late when its waiting time in the system, denoted by *w*, is greater than the quoted lead time *l*. In this case, the job's lateness is given by *(w-l)*. We are interested here in the expected lateness for a late job. The expected lateness for any given job can be deduced immediately by multiplying our result by the probability that a job is late. Thus, given a threshold lead time *l*, we calculate the expected waiting time in an M/M/1/K queue after *l*, given that the waiting time is greater than *l*. In the particular case of *l=0*, the expected lateness given that a job is late is equal to the expected waiting time in the system, which is a known performance for an M/M/1/K.

[+] Institute of Engineering Univ. Grenoble Alpes

2. Motivation

The calculation of the expected lateness for a late job in an M/M/1/K queue permits to obtain the expected lateness penalty cost in a system modelled by an M/M/1/K. Indeed, as explained by Palaka et al. (1998), the expected lateness penalty cost can be given by: (penalty per job per unit's lateness) × (expected number of overdue clients) × (expected lateness given that a job is late). The penalty per job per unit's lateness is an input parameter, and the expected number of overdue clients is known for an M/M/1/K. Hence, obtaining the expected lateness for a late job will permit calculating the expected lateness penalty cost in an M/M/1/K queue, which may have many interesting applications.

A potential application of our result is in the call centers where we have to take into account the cost of not respecting the quoted lead time when we announce this lead time to a client. The examples of significant lateness penalty costs that are likely to impact the firms' decisions are also abundant in the industry. Savaçaneril et al., (2010) reported that the cost of late delivery in the FMC Wellhead Equipment Division may rise up to $250,000 per day and that the lateness penalties in the aircraft industry starts from $10,000-$15,000 and can go as high as $1,000,000 per day.

To illustrate the relevance of the consideration of the expected lateness penalty cost in the production systems, we consider the widely studied problem of a company, that is modelled as a make-to-order queue, and has to quote the right lead time to its customers. A shorter quoted lead time can lead to an increase in the demand but increases the risk of late delivery, implying an increase in the expected lateness penalty cost. In their pioneer paper, Palaka et al., (1998) addressed this problem while modelling the system as an M/M/1 queue. Later, most of the papers dealing with this type of problems adopted Palaka et al.'s framework (see e.g., Zhao et al., 2012; Boyaci and Ray, 2003, 2006; Ray and Jewkes, 2004). Our result can be used to generalize this stylized framework by considering an M/M/1/K queue instead of an M/M/1, which thus permits to consider a new policy where the clients are rejected whenever the queue is full, unlike the case of M/M/1 where all the customers are accepted. This may lead to different insights.

3. Problem statement and the result

We consider an M/M/1/K queue with mean service rate μ and mean arrival rate λ. The probability P_k of having k customers in the system (k=1, 2… K) for a random point in time and in steady state is given by equation (1) (see Gross et al. 2008). Therefore, P_K represents the probability of rejecting a customer, and $(1-P_K)$ is the probability that a customer is accepted. The throughput rate ($\overline{\lambda}$) is equal to the mean arrival rate multiplied by the probability of accepting a customer, as given in equation (2). The expected number of customers in the system for a random point in time and in steady state, denoted by L_s, is given by equation (3) (see Gross et al. 2008).

$$P_k = \frac{1-\rho}{1-\rho^{K+1}}\rho^k \text{ if } \rho \neq 1 \text{ and } P_k = \frac{1}{K+1} \text{ if } \rho = 1 \text{ with } \rho = \frac{\lambda}{\mu} \quad (1)$$

$$\overline{\lambda} = \lambda(1-P_K) \quad (2)$$

$$L_s = \frac{\rho}{1-\rho} - \frac{(K+1)\rho^{K+1}}{1-\rho^{K+1}} \quad (3)$$

We let the random variable *W* refers to the waiting time in the system, with probability density function $f_W(.)$ and cumulative distribution function $F_W(.)$. In an M/M/1/K queue, it is known that $F_W(.)$ is given by equation (4):

$$F_W(w) = 1 - \sum_{k=0}^{K-1} \frac{P_k}{1-P_K} \left(\sum_{i=0}^{k} \frac{(\mu w)^i}{i!} e^{-\mu w} \right) \quad (4)$$

The expected waiting time in an M/M/1/K queue is obtained by $L_s / \overline{\lambda}$, and its explicit expression can be deduced from equations (2) and (3).

Given a quoted lead time *l*, we denote by R_l the expected lateness for a late job. In other words, R_l represents the expected waiting time in the system beyond the threshold lead time l, given that the waiting time is greater than l. Our objective is to calculate R_l.

The probability of having a waiting time *W* between *w* and $w + dw$ given that *w* is greater than the quoted lead time *l*, is given by $f_W(w|w \geq l)dw$, and the lateness in this case is $(w-l)$. Hence, for a quoted lead time *l*, the expected lateness given that a job is late is: $\int_l^{\infty} (w-l) f_W(w|w \geq l) dw$. We will first determine the explicit expression of $f_W(w|w \geq l)$ and then turn to the calculation of the integral function R_l.

We have $f_W(w|w \geq l) = \frac{d}{dw} F_W(w|w \geq l)$ and $F_W(w|w \geq l) = \frac{\Pr(l \leq W \leq w)}{\Pr(W \geq l)} = \frac{F_W(w) - F_W(l)}{1 - F_W(l)}$.

Consequently, $f_W(w|w \geq l) = \frac{d}{dw}\left(\frac{F_W(w) - F_W(l)}{1 - F_W(l)} \right)$. In Lemma 1, we explicitly calculate $f_W(w|w \geq l)$.

Lemma 1. *Consider an M/M/1/K queue with mean service rate μ and mean arrival rate λ. We let W denote the waiting time in the system and* $f_W(.)$ *its probability density function. We denote by l a threshold lead time. The probability density function of the waiting time in the system given that this waiting time is greater than the threshold lead time l, is given by:*

$$f_W(w|w \geq l) = \frac{\mu e^{-\mu(w-l)}}{\sum_{k=0}^{K-1} P_k \left(\sum_{i=0}^{k} \frac{(\mu l)^i}{i!} \right)} \sum_{k=0}^{K-1} P_k \frac{(\mu w)^k}{k!}$$

Proof. $f_W(w|w \geq l) = \frac{d}{dw}\left(\frac{F_W(w) - F_W(l)}{1 - F_W(l)} \right)$ *where* $F_W(x) = 1 - \sum_{k=0}^{K-1} \frac{P_k}{1-P_K} \left(\sum_{i=0}^{k} \frac{(\mu x)^i}{i!} e^{-\mu x} \right)$.

Hence, $\frac{F_W(w) - F_W(l)}{1 - F_W(l)} = \frac{\sum_{k=0}^{K-1} \frac{P_k}{1-P_K} \left(\sum_{i=0}^{k} \frac{(\mu l)^i}{i!} e^{-\mu l} - \frac{(\mu w)^i}{i!} e^{-\mu w} \right)}{\sum_{k=0}^{K-1} \frac{P_k}{1-P_K} \left(\sum_{i=0}^{k} \frac{(\mu l)^i}{i!} e^{-\mu l} \right)} = 1 - e^{-\mu(w-l)} \frac{\sum_{k=0}^{K-1} P_k \left(\sum_{i=0}^{k} \frac{(\mu w)^i}{i!} \right)}{\sum_{k=0}^{K-1} P_k \left(\sum_{i=0}^{k} \frac{(\mu l)^i}{i!} \right)}$.

By deriving this function with respect to w, we obtain:

$$f_W(w|w \geq l) = \frac{1}{\sum_{k=0}^{K-1} P_k \left(\sum_{i=0}^{k} \frac{(\mu l)^i}{i!} \right)} \left[\mu e^{-\mu(w-l)} \sum_{k=0}^{K-1} P_k \left(\sum_{i=0}^{k} \frac{(\mu w)^i}{i!} \right) - e^{-\mu(w-l)} \sum_{k=0}^{K-1} P_k \left(\sum_{i=1}^{k} \frac{\mu (\mu w)^{i-1}}{(i-1)!} \right) \right]$$

$$= \frac{\mu e^{-\mu(w-l)} \sum_{k=0}^{K-1} P_k \left(\left(\sum_{i=1}^{k} \left(\frac{(\mu w)^i}{i!} - \frac{(\mu w)^{i-1}}{(i-1)!} \right) \right) + 1 \right)}{\sum_{k=0}^{K-1} P_k \left(\sum_{i=0}^{k} \frac{(\mu l)^i}{i!} \right)}.$$

Given that $\sum_{i=1}^{k} \left(\frac{(\mu w)^i}{i!} - \frac{(\mu w)^{i-1}}{(i-1)!} \right) = \frac{(\mu w)^k}{k!} - 1$, *we finally have*

$$f_W(w|w \geq l) = \frac{\mu e^{-\mu(w-l)}}{\sum_{k=0}^{K-1} P_k \left(\sum_{i=0}^{k} \frac{(\mu l)^i}{i!} \right)} \sum_{k=0}^{K-1} P_k \frac{(\mu w)^k}{k!} \blacksquare$$

Based on the result of Lemma 1, we now have $R_l = \int_l^{\infty} \frac{(w-l)\mu e^{-\mu(w-l)}}{\sum_{k=0}^{K-1} P_k \left(\sum_{i=0}^{k} \frac{(\mu l)^i}{i!} \right)} \sum_{k=0}^{K-1} P_k \frac{(\mu w)^k}{k!} dw = \left(\frac{1}{\sum_{k=0}^{K-1} P_k \left(\sum_{i=0}^{k} \frac{(\mu l)^i}{i!} \right)} \sum_{k=0}^{K-1} \frac{P_k}{k!} \right) \Psi_k(l)$, where $\Psi_k(l)$ is defined for a given *k* by: $\Psi_k(l) = \int_l^{\infty} (w-l)\mu e^{-\mu(w-l)} (\mu w)^k dw$. To find the expression of R_l, we thus need to calculate $\Psi_k(l)$.

The integral function $\Psi_k(l)$ can be written as follows: $\Psi_k(l) = \int_l^{\infty} w\mu e^{-\mu(w-l)} (\mu w)^k dw - \int_l^{\infty} l\mu e^{-\mu(w-l)} (\mu w)^k dw$.

Hence, $\Psi_k(l) = \frac{e^{\mu l}}{\mu} \int_l^{\infty} \mu e^{-\mu w} (\mu w)^{k+1} dw - l e^{\mu l} \int_l^{\infty} \mu e^{-\mu w} (\mu w)^k dw$, implying that $\Psi_k(l) = \frac{e^{\mu l}}{\mu} A_{k+1} - l e^{\mu l} A_k$, where $(A_k)_{k \in \mathrm{N}}$ is a sequence defined by $A_k = \int_l^{\infty} \mu e^{-\mu w} (\mu w)^k dw$.

Therefore, the next step consists in calculating A_k. The result is given in Lemma 2.

Lemma 2. $A_k = \int_l^{\infty} \mu e^{-\mu w} (\mu w)^k dw = \left(\sum_{i=0}^{k} \frac{k!}{i!} (\mu w)^i \right) e^{-\mu w}$

Proof. *We demonstrate the lemma by induction.*

For k=0, we verify that $A_0 = \int_l^{\infty} \mu e^{-\mu w} dw = \left[-e^{-\mu w} \right]_l^{\infty} = e^{-\mu l} = \left(\sum_{i=0}^{0} \frac{0!}{i!} (\mu l)^i \right) e^{-\mu l}$.

Suppose that $A_k = \left(\sum_{i=0}^{k} \frac{k!}{i!} (\mu l)^i \right) e^{-\mu l}$ *and let's demonstrate that* $A_{k+1} = \left[\sum_{i=0}^{k+1} \frac{(k+1)!}{i!} (\mu l)^i \right] e^{-\mu l}$.

We have $A_{k+1} = \int_l^{\infty} \mu e^{-\mu w} (\mu w)^{k+1} dw = \int_l^{\infty} g'(w) h(w) dw$ *where* $g'(w) = \mu e^{-\mu w}$ *and* $h(w) = (\mu w)^{k+1}$.

We use the partial integration to transform the expression of A_{k+1}.

Indeed, $A_{k+1} = \left[g(w)h(w)\right]_l^{\infty} - \int_l^{\infty} g(w)h'(w)dw = \left[-e^{-\mu w}(\mu w)^{k+1}\right]_l^{\infty} + (k+1)\int_l^{\infty} \mu e^{-\mu w}(\mu w)^k dw$.

Thus, $A_{k+1} = \left[-e^{-\mu w}(\mu w)^{k+1}\right]_l^{\infty} + (k+1)A_k$.

Given that $\lim_{w\to+\infty} e^{-\mu w}(\mu w)^{k+1} = 0$, *we deduce that* $A_{k+1} = e^{-\mu l}(\mu l)^{k+1} + (k+1)A_k$.

Since we supposed that $A_k = \left(\sum_{i=0}^{k} \frac{k!}{i!}(\mu l)^i\right)e^{-\mu l}$, *it comes that:*

$$A_{k+1} = e^{-\mu l}(\mu l)^{k+1} + (k+1)\left(\sum_{i=0}^{k} \frac{k!}{i!}(\mu l)^i\right)e^{-\mu l} = e^{-\mu l}\left[(\mu l)^{k+1} + (k+1)\left(\sum_{i=0}^{k} \frac{k!}{i!}(\mu l)^i\right)\right]$$

$$= e^{-\mu l}\left[(\mu l)^{k+1} + \sum_{i=0}^{k} \frac{(k+1)!}{i!}(\mu l)^i\right] = e^{-\mu l}\left[\sum_{i=0}^{k+1} \frac{(k+1)!}{i!}(\mu l)^i\right],$$ *which demonstrates the Lemma* ■

Using the above results, we can now provide the explicit expression of the expected lateness for a late job in an M/M/1/K queue. We remind the reader that the expected lateness for any given job can be easily obtained by multiplying our result by the probability that a job is late, which is equal to $(1 - F_W(l))$ in an M/M/1/K queue.

Theorem. *Consider an M/M/1/K queue with mean service rate μ and mean arrival rate λ. Given a quoted lead time l, the expected lateness for a late job (i.e., the expected waiting time in the system beyond l given that the waiting time is greater than l) is determined by:*

$$\frac{\sum_{k=0}^{K-1} \frac{P_k}{k!}\left[\frac{(\mu l)^{k+1}}{\mu} + \left(\frac{k+1}{\mu} - l\right)\sum_{i=0}^{k} \frac{k!}{i!}(\mu l)^i\right]}{\sum_{k=0}^{K-1} P_k \sum_{i=0}^{k} \frac{(\mu l)^i}{i!}}$$

Where, $P_k = \frac{1-\rho}{1-\rho^{K+1}}\rho^k$ *if* $\rho \neq 1$, $P_k = \frac{1}{K+1}$ *if* $\rho = 1$, *and* $\rho = \frac{\lambda}{\mu}$.

Proof. *Based on the result of Lemma 2,* it comes that

$$\Psi_k(l) = \frac{e^{\mu l}}{\mu}A_{k+1} - le^{\mu l}A_k = \sum_{i=0}^{k+1} \frac{(k+1)!}{i!}\frac{(\mu l)^i}{\mu} - \sum_{i=0}^{k} \frac{k!}{i!}(\mu l)^i l = \frac{(\mu l)^{k+1}}{\mu} + \left(\frac{k+1}{\mu} - l\right)\sum_{i=0}^{k} \frac{k!}{i!}(\mu l)^i.$$

Therefore, we finally obtain $R_l = \left(\frac{1}{\sum_{k=0}^{K-1} P_k \left(\sum_{i=0}^{k} \frac{(\mu l)^i}{i!} \right)} \sum_{k=0}^{K-1} \frac{P_k}{k!} \right) \Psi_k(l)$

$$= \frac{\sum_{k=0}^{K-1} \frac{P_k}{k!} \left[\frac{(\mu l)^{k+1}}{\mu} + \left(\frac{k+1}{\mu} - l \right) \sum_{i=0}^{k} \frac{k!}{i!} (\mu l)^i \right]}{\sum_{k=0}^{K-1} P_k \sum_{i=0}^{k} \frac{(\mu l)^i}{i!}}.$$

Particular cases:

- When the quoted lead time l tends to 0, the expected lateness given that a job is late becomes equal to the expected waiting time in the system. On the one hand, using our proposed formula of the expected lateness for a late job, denoted by R_l, we obtain $\lim_{l \to 0} R_l = \frac{\sum_{k=0}^{K-1} \frac{(k+1)}{\mu} P_k}{\sum_{k=0}^{K-1} P_k}$. On the other hand, it is known in the literature that the expected waiting time in an M/M/1/K queue is $L_s / \overline{\lambda} = \sum_{k=0}^{K-1} \frac{(k+1)}{\mu} \frac{\rho^k P_0}{1 - P_K}$ (see Sztrik, 2011). One can also demonstrate that $\frac{\sum_{k=0}^{K-1} \frac{(k+1)}{\mu} P_k}{\sum_{k=0}^{K-1} P_k} = \sum_{k=0}^{K-1} \frac{(k+1)}{\mu} \frac{\rho^k P_0}{1 - P_K}$. Thus, when l tends to 0, we verified that our formula of expected lateness for a late job effectively yields the expected waiting time in the system.
- For K=1, the base integral function that defines the expected lateness for a late job in an M/M/1/K (i.e., $\int_l^{\infty} (w - l) f_W(w | w > l) dw$) can be directly calculated, and gives $\frac{1}{\mu}$. It can be verified that our formula, proposed in the Theorem, also gives the same result when K=1.
- Given the complexity of our formula, we were not able to calculate directly $\lim_{K \to +\infty} R_l$. However, when K tends to infinity, the M/M1/K queue becomes equivalent to the M/M/1 queue. Therefore, when K tends to infinity, our result R_l gives the expected lateness for a

late job in an M/M/1, which is known to be equal to $\frac{1}{\mu-\lambda}$ (as the waiting time is exponentially distributed for an M/M/1). Thus, this proves that we have $\lim_{K\to+\infty} \frac{\sum_{k=0}^{K-1} \frac{P_k}{k!}\left[\frac{(\mu l)^{k+1}}{\mu} + \left(\frac{k+1}{\mu} - l\right)\sum_{i=0}^{k} \frac{k!}{i!}(\mu l)^i\right]}{\sum_{k=0}^{K-1} P_k \sum_{i=0}^{k} \frac{(\mu l)^i}{i!}} = \frac{1}{\mu-\lambda}$.

References

Boyaci, T., Ray, S. (2003). Product differentiation and capacity cost interaction in time and price sensitive markets. *Manufacturing & Service Operations Management*, 5, 18-36.

Boyaci, T., Ray, S. (2006). The impact of capacity costs on product differentiation in delivery time, delivery reliability, and price. *Production and Operations Management*, 15, 179-197.

Gross, D., Shortle, J.F., Thompson, J.M., Harris, C. (2008). *Fundamentals of queueing theory - Fourth Edition*, John Wiley & Sons, Inc., Hoboken, New Jersey.

Palaka, K., Erlebacher, S., and Kropp, D.H. (1998). Lead time setting, capacity utilization, and pricing decisions under lead time dependent demand. *IIE Transactions*, 30, 151–163.

Ray, S., Jewkes, E.M. (2004). Customer lead time management when both demand and price are lead time sensitive. *European Journal of Operational Research*, 153, 769–781.

Savaçaneril, S., Griffin, P. M., Keskinocak, P. (2010). Dynamic lead-time quotation for an M/M/1 base-stock inventory queue. *Operations research*, 58(2), 383-395.

Sztrik, J. (2011). *Basic Queueing Theory*. University of Debrecen: Faculty of Informatics, Debrecen.

Zhao, X., Stecke, K.E., and Prasad, A. (2012). Lead Time and Price Quotation Mode Selection: Uniform or Differentiated? *Production and Operations Management*, 21, 177–193.