Mean value analysis of single server retrial queues *

J.R. Artalejo
Department of Statistics and Operations Research
Faculty of Mathematics
Universidad Complutense of Madrid
28040 Madrid, Spain

J.A.C. Resing
Department of Mathematics and Computer Science
Eindhoven University of Technology
P.O. Box 513, 5600 MB Eindhoven, The Netherlands

Abstract

Mean value analysis is an elegant tool for determining mean performance measures in queueing models. We use the technique to analyze the $M/G/1$ retrial queue with exponential retrial times. We also show how the relations can be adapted to obtain mean performance measures in more advanced $M/G/1$-type retrial queues.

Keywords: retrial queue, mean value analysis

1 Introduction

Most papers on retrial queues deal with the steady-state distribution of the system state. In this context, the expected number of customers in orbit (or in the system) is typically obtained by differentiating the corresponding generating function. The use of this method implies the knowledge of the generating function (which can be calculated by using embedded Markov chains, supplementary variables, Markov renewal theory, etc.) and some algebraic manipulations (differentiation, L'Hôpital rule, etc.).

The goal of the mean value analysis technique is to provide an elegant alternative for obtaining the expected number of customers in orbit (and the expected waiting

*Corresponding author: J.A.C. Resing, Department of Mathematics and Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands, e-mail address: j.a.c.resing@tue.nl

time) by avoiding the use of generating functions. This significantly reduces the algebra. The mean value analysis technique is known as a powerful tool for determining mean performance measures in all kind of queueing models (see, e.g., Winands et al. [7] for a recent example of the use of the technique in the analysis of polling systems). However, as far as the authors know, it has not been applied to retrial queueing systems before.

The rest of the paper is organized as follows. In Section 2 we use mean value analysis to determine mean performance measures in the main $M/G/1$ retrial queue. Next, in Section 3 we show how the mean value relations can be adapted to obtain mean performance measures in more advanced $M/G/1$-type retrial queues studied in the literature.

2 The main $M/G/1$ retrial queue

In the main $M/G/1$ retrial queue, customers arrive according to a Poisson process with rate λ. Service times are generally distributed with mean $E(B)$ and mean residual service time $E(R) = E(B^2)/(2E(B))$. The retrial queue is characterized by the requirement that customers finding the server busy must leave the service area and join a group of unsatisfied customers called orbit. Customers in orbit retry after an exponentially distributed time with parameter μ. To ensure stability, we assume that $\rho = \lambda E(B) < 1$. With W we denote the steady-state waiting time of customers and with L we denote the steady-state orbit size.

The mean value relations for the system are given by

$$E(W) = E(L)E(B) + \rho\left(E(R) + \frac{1}{\mu}\right), \tag{1}$$

$$E(L) = \lambda E(W). \tag{2}$$

Combination of the two relations gives the expression for the mean waiting time

$$E(W) = \frac{\rho}{1-\rho}\left(E(R) + \frac{1}{\mu}\right). \tag{3}$$

The key relation is (1). Of course, (2) is just Little's law. The main ideas to obtain (1) are the following:

1. The steady-state waiting time W can be written as the sum of the *idle time of the server* during the waiting time, W_0, and the *busy time of the server* during the waiting time, W_1, i.e.,

$$W = W_0 + W_1. \tag{4}$$

2. The idle time of the server during the waiting time of a customer, given that the customer goes in orbit, is exponentially distributed with parameter μ. This fact follows from the memoryless property of the exponential distribution of the retrial times. The waiting time of a tagged customer going in orbit consists of successive busy and idle times of the server. The transitions from busy to

idle times are caused by service completions, the transitions from idle to busy times are caused by arrivals of other customers (external arrivals or retrials). These are external factors, they have nothing to do with the retrial times of our tagged customer. Now, cutting away the busy times of the server and glueing together the idle times of the server, a tagged customer going in orbit just sees one big idle time, where in each small interval of length Δ, he retries with probability $\mu\Delta$. Hence, we have that the total idle time of the server during the waiting time of the tagged customer, given that the customer goes in orbit, is exponentially distributed with parameter μ.

Because, due to the PASTA property, the blocking probability, P_b, that an arbitrary customer goes into orbit equals the fraction of time, ρ, that the server is busy, we obtain

$$E(W_0) = \frac{\rho}{\mu}. \tag{5}$$

Remark that the assumption of exponential retrial times is essential here.

3. Mean waiting times are the same for systems with random order of service (ROS) and systems with first come first served (FCFS) service discipline. The same holds for mean idle times and mean busy times of the server during these waiting times. In our retrial model, we mean the following with FCFS service discipline. On one hand, there is a FCFS discipline for the customers in orbit. Given that the number of customers in orbit is j, the retrial rate is $j\mu$. If a retrial occurs, the customer at the head of the orbit occupies the server. In addition, we assume that fresh arrivals finding the server free go to the orbit and, at that moment, the customer at the head of the orbit occupies the server. In the sequel we will call this the *strict FCFS service discipline*.

In the retrial system with strict FCFS service discipline, it is immediately clear that the total expected busy time of the server during the waiting time of a tagged customer is given by $E(L)E(B) + \rho E(R)$. Hence, we obtain

$$E(W_1) = E(L)E(B) + \rho E(R). \tag{6}$$

Now, (1) follows from a combination of (4), (5) and (6).

Remark: (Alternative proof of $E(W_0) = \rho/\mu$)
Introduce $E(L_0) = \sum_{j=0}^{\infty} jP_{0j}$ where P_{0j} is the steady-state probability that there are j customers in orbit and the server is idle. An alternative proof of $E(W_0) = \rho/\mu$ is given by combining Little's law $E(L_0) = \lambda E(W_0)$ with the relation $\mu E(L_0) = \lambda P_b$, which equals the rate at which customers leave the orbit to the rate at which customers enter the orbit.

Remark: (Alternative proof based on stochastic decomposition)
Since the $M/G/1$ retrial queue is a vacation model satisfying the classical assumptions given by Fuhrman and Cooper [5], we have

$$E(L) = E(L_\infty) + \frac{E(L_0)}{1-\rho},$$

where $E(L_\infty)$ represents the mean number of customer in queue (excluding the possible customer receiving service) in the standard $M/G/1$ queue.

From Little's law we have the equivalent expression

$$E(W) = E(W_\infty) + \frac{E(W_0)}{1-\rho}. \tag{7}$$

Since the mean value analysis for the standard $M/G/1$ queue gives

$$E(W_\infty) = \frac{\rho}{1-\rho} E(R), \tag{8}$$

we find, combining (5), (7) and (8), that $E(W)$ is given by (3).

Remark: (Proof of exponentiality of $(W_0|W_0 > 0)$**)**
Above, we heuristically argued that $(W_0|W_0 > 0)$ is exponential with parameter μ. A more formal proof of this fact is given below. Introduce the following notation:

$\varphi_{W_0}(s)$: the Laplace-Stieltjes transform (LST) of W_0,

$\varphi_j(s)$: the LST of the residual W_0 at departure epochs given that the number of customers in orbit is j (including the tagged customer),

k_i^x : the probability that i primary customers arrive during a residual service time given that the elapsed service time is x,

k_i : the probability that i primary customers arrive during a service time,

$P_{1j}(x)dx$: the steady-state probability that the server is busy, there are j customers in orbit and the elapsed service time of the customer in service $\in (x, x+dx)$.

We first notice that

$$\varphi_{W_0}(s) = \sum_{j=0}^{\infty} \int_0^{\infty} P_{1j}(x) \sum_{i=0}^{\infty} k_i^x \varphi_{j+1+i}(s) dx + 1 - \rho. \tag{9}$$

The following equations govern the dynamics of the LST $\varphi_j(s)$:

$$\begin{aligned} \varphi_{j+1+i}(s) &= \frac{\mu}{\lambda + (j+1+i)\mu + s} + \frac{\lambda}{\lambda + (j+1+i)\mu + s} \sum_{m=0}^{\infty} k_m \varphi_{j+1+i+m}(s) \\ &+ \frac{(j+i)\mu}{\lambda + (j+1+i)\mu + s} \sum_{m=0}^{\infty} k_m \varphi_{j+i+m}(s), \text{ for } j \geq 0,\ i \geq 0. \end{aligned} \tag{10}$$

Equation (10) describes the motion between two successive service completion epochs. For example, the term $\mu/(\lambda + (j+1+i)\mu + s)$ indicates that the next service time corresponds to the tagged customer. Then, we cut away the server busy times and use the memoryless property of the exponential law.

Inserting the constant (it does not depend on the orbit state) transform $\mu(\mu+s)^{-1}$ in the above equation, we automatically verify that it satisfies (10). Then, from (9) the desired result follows

$$\varphi_{W_0}(s) = \rho\frac{\mu}{\mu+s} + 1 - \rho.$$

3 Other advanced $M/G/1$-type retrial queues

The mean value analysis technique also leads to the expected values for the waiting times and orbit sizes in more advanced $M/G/1$-type retrial queues. This fact is illustrated in the sequel for a variety of systems including batch arrivals, priorities, impatience, network blocking, etc.

3.1 The model with batch arrivals

Consider the batch arrival $M/G/1$ retrial queue with exponential retrial times as discussed in Section 3.1 of Falin and Templeton [4]. The batch size of the arrivals is given by the random variable X with probability distribution

$$x_k = P(X = k), \qquad k \geq 1.$$

The mean value relations for the system are now given by

$$\begin{aligned} E(W) &= \left(E(L) + \sum_{k=1}^{\infty} r_k(k-1)\right) E(B) + \rho E(R) + \frac{1 - r_1(1-\rho)}{\mu}, \qquad (11) \\ E(L) &= \lambda E(X) E(W), \end{aligned}$$

where the server utilization ρ is given by

$$\rho = \lambda E(X) E(B),$$

and r_k is the probability that an arbitrary customer is the $k-$th customer served in his batch. The first two terms in the right-hand side of (11) correspond to the busy time of the server and the third term to the idle time of the server during the waiting time. Remark that the probability an arbitrary customer is going into orbit now equals $1 - r_1(1-\rho)$.

The probabilities r_k follow from standard renewal-type arguments and are given by (see, e.g., Adan and Resing [1], Section 10.4)

$$r_k = \frac{1}{E(X)} \sum_{n=k}^{\infty} x_n,$$

and hence it immediately follows that $r_1 = 1/E(X)$ and

$$\sum_{k=1}^{\infty} r_k(k-1) = \frac{E(X^2) - E(X)}{2E(X)}.$$

Substituting these two results in (11), we obtain the expressions for $E(W)$ and $E(L)$. In particular, we have

$$E(W) = \frac{1}{1-\rho}\left(\frac{E(X^2)-E(X)}{2E(X)}E(B) + \rho E(R)\right) + \frac{E(X)-1+\rho}{(1-\rho)\mu E(X)}.$$

3.2 The model with priority subscribers

We now deal with the model with priority subscribers as discussed in Section 3.2 of Falin and Templeton [4]. In this model, high priority customers form a queue (like in the standard $M/G/1$ queue) while low priority customers retry after an exponential time (like in the standard $M/G/1$ retrial queue). In this subsection the subscript 1 is associated with the high priority customers and the subscript 2 is associated with the low priority customers. The stability condition is given by $\rho_1 + \rho_2 < 1$, where $\rho_i = \lambda_i E(B_i)$, for $i = 1, 2$.

The mean value relations for the high priority customers are now given by

$$\begin{aligned} E(W_1) &= E(L_1)E(B_1) + \sum_{i=1}^{2} \rho_i E(R_i), \\ E(L_1) &= \lambda_1 E(W_1), \end{aligned}$$

leading to

$$E(W_1) = \frac{\sum_{i=1}^{2} \rho_i E(R_i)}{1-\rho_1},$$

and the mean value relations for the low priority customers are given by

$$\begin{aligned} E(W_2) &= \sum_{i=1}^{2} E(L_i)E(B_i) + \sum_{i=1}^{2} \rho_i E(R_i) + \rho_1 E(W_2) + \frac{\rho_1+\rho_2}{\mu}, \\ E(L_2) &= \lambda_2 E(W_2), \end{aligned} \tag{12}$$

leading to

$$E(W_2) = \frac{\sum_{i=1}^{2} \rho_i E(R_i)}{(1-\rho_1)(1-\rho_1-\rho_2)} + \frac{\rho_1+\rho_2}{(1-\rho_1-\rho_2)\mu}.$$

The last term in (12) corresponds to the total expected idle time of the server during the waiting time of a low priority customer. Remark that a low priority customer goes in orbit with probability $\rho_1+\rho_2$. The first three terms in (12) correspond to the total expected busy time of the server during the waiting time of a low priority customer. To obtain these three terms, we assume the strict FCFS policy for low priority customers, but we continue assuming that high priority customers have non-preemptive priority over low priority customers. The first two terms are the contribution due to the amount of work that a low priority customer finds upon arrival. The third term is the contribution of those forthcoming high priority customers arriving during W_2 and receiving service before the tagged customer.

3.3 The model with impatient subscribers

Consider the model with impatient subscribers as discussed in Section 3.3.2 of Falin and Templeton [4]. In this model, the probability that a customer retries after the first attempt equals H_1. On the other hand, the probability that a customer retries after the i-th attempt equals 1, for $i \geq 2$. The system is stable when $\rho H_1 < 1$. Observe that the blocking probability P_b in this model is the solution of the equation

$$\lambda(1 - P_b(1 - H_1))E(B) = P_b,$$

and hence given by

$$P_b = \frac{\rho}{1 + \rho(1 - H_1)}.$$

This follows from the PASTA property and the fact that $1 - P_b(1 - H_1)$ is the probability an arbitrary customer is eventually served.

The mean value relations are now given by

$$\begin{aligned} E(W_0) &= \frac{H_1 P_b}{\mu}, && (13)\\ E(W_1) &= E(L_0)E(B) + H_1 E(L_1)E(B) + H_1 P_b E(R), && (14)\\ E(L_0) &= \lambda E(W_0), && (15)\\ E(L_1) &= \lambda E(W_1), && (16) \end{aligned}$$

where $E(L_i) = \sum_{j=0}^{\infty} j P_{ij}$, with P_{0j} (resp. P_{1j}) the steady-state probability that there are j customers in orbit and the server is idle (resp. busy). Together these four relations determine the four unknowns $E(W_0), E(W_1), E(L_0)$ and $E(L_1)$.

Combining formulas (13)-(16), we get

$$E(W) = \frac{H_1 \rho}{1 - H_1 \rho} \left(\frac{E(R)}{1 + \rho(1 - H_1)} + \frac{1}{\mu} \right).$$

3.4 The model with general retrial times and only retrial of head of the orbit

In Gómez-Corral [6] a model with general retrial times is considered. Special features of the model are that those customers who find the server busy are queued in orbit and only the customer at the head of the orbit is allowed to conduct retrials. Besides, a new retrial time begins at a service completion epoch instead of at the moment of a service attempt failure.

Denote with the random variable A a retrial time and with $\alpha(s)$ its LST. The minimum between A and the exponential law with rate λ is denoted as ξ_A^{λ}. Then, the mean value relations are given by

$$\begin{aligned} E(W_0) &= (E(L) + \rho)E(\xi_A^{\lambda}), && (17)\\ E(W_1) &= E(L)E(B) + \rho E(R), \end{aligned}$$

leading to

$$E(W) = \frac{\rho(E(R) + E(\xi_A^\lambda))}{1 - \rho - \lambda E(\xi_A^\lambda)}.$$

In order to obtain (17) we now also use the strict FCFS service discipline to calculate the total expected *idle time of the server* during the waiting time. Now, using the fact that

$$E(\xi_A^\lambda) = \frac{1 - \alpha(\lambda)}{\lambda},$$

we obtain

$$E(W) = \frac{\rho}{\alpha(\lambda) - \rho} \left(E(R) + \frac{1 - \alpha(\lambda)}{\lambda} \right).$$

3.5 The model with two-phase service

We now consider the model with two-phase service studied in Artalejo and Choudhury [2]. In this model, the server provides an essential service B_1 (with residual service time R_1) to all customers. Some customers are provided with a second optional service B_2 (with residual service time R_2) , with probability p. Thus, we have

$$B = \begin{cases} B_1 + B_2, & \text{with prob. } p, \\ B_1, & \text{with prob. } 1 - p. \end{cases}$$

Note that $E(B) = E(B_1) + pE(B_2)$ and $E(B^2) = E(B_1^2) + pE(B_2^2) + 2pE(B_1)E(B_2)$. Furthermore, we denote $\rho_1 = \lambda E(B_1), \rho_2 = \lambda p E(B_2)$ and $\rho = \rho_1 + \rho_2 = \lambda E(B)$.

As before, P_{ij} denotes the steady-state probabilities of the system state. More concretely, i denotes the state of the server taking the values 0 (server idle), 1 (server busy with an essential service) and 2 (service busy with an optional service), while j denotes the number of customers in orbit. Furthermore, $E(L_i) = \sum_{j=0}^{\infty} j P_{ij}$, for $i = 1, 2, 3$.

Finally, let us introduce the following notation:

W_0 : the server idle time during the waiting time of an arbitrary tagged customer,

W_1^1 : the busy time of the server providing essential service during the waiting time of the tagged customer,

W_1^2 : the busy time of the server providing optional service during the waiting time of the tagged customer.

Because the model is a special case of the main $M/G/1$ retrial queue, we obtain, as in Section 2,

$$E(W_0) = \frac{\rho}{\mu}.$$

Now we assume the strict FCFS discipline and consider mean value analysis for $E(W_1^i)$, for $i = 1, 2$, which yields

$$
\begin{aligned}
E(W_1^1) &= E(L)E(B_1) + \rho_1 E(R_1), \\
E(W_1^2) &= pE(L)E(B_2) + \rho_1 pE(B_2) + \rho_2 E(R_2).
\end{aligned}
$$

Obviously, $E(L_i) = \lambda E(W_1^i)$, for $i = 0, 1, 2$. From the above formulas, we obtain

$$
E(W_1^1) = \frac{\rho_1 \rho}{1-\rho}\left(E(R) + \frac{1}{\mu}\right) + \rho_1 E(R_1),
$$

$$
E(W_1^2) = \frac{\rho_2 \rho}{1-\rho}\left(E(R) + \frac{1}{\mu}\right) + \rho_2 E(R_2) + \frac{\rho_1 \rho_2}{\lambda}.
$$

3.6 The model with network blocking

In this model, described in Falin [3], those incoming customers (primary arrivals and orbit customers) who find a free server can receive an engaged signal (due to dialing error, breakdown, wrong connection, etc.) with probability p. In that case, the customer will join the orbit. With probability $q = 1 - p$ customers who find a free server really enter the service station. As usual those customers who find the server busy also join the orbit.

The probability an arbitrary customer goes in orbit now equals $\rho + p(1-\rho) = p + q\rho$. The idle time of the server during the waiting time of a customer, given that the customer goes in orbit, now consists of a geometric number of exponential periods. After each exponential period with parameter μ, the idle time ends with probability q and continues with probability p. Hence, $(W_0 \,|\, W_0 > 0)$ is exponentially distributed with parameter μq. So, we conclude

$$
E(W_0) = \frac{p + q\rho}{\mu q}.
$$

Remark: (Alternative proof of $E(W_0) = (p + q\rho)/(\mu q)$)
The level crossing argument for the orbit now gives $\mu q E(L_0) = \lambda(p + q\rho)$, which in combination with $E(L_0) = \lambda E(W_0)$ also yields $E(W_0) = (p + q\rho)/(\mu q)$.

As in the main $M/G/1$ retrial queue, the mean value equation for the busy time of the server during the waiting time is given by

$$
E(W_1) = E(L)E(B) + \rho E(R).
$$

Now it is easy to get, using Little's law,

$$
\begin{aligned}
E(W_1) &= \frac{\rho}{1-\rho}E(R) + \frac{\rho(p + q\rho)}{\mu q(1-\rho)}, \\
E(W) &= \frac{\rho}{1-\rho}E(R) + \frac{p + q\rho}{\mu q(1-\rho)}.
\end{aligned}
$$

3.7 The model of two-way communication

In this model, we have two types of calls namely ingoing calls (i.e., calls made by regular customers as in the main $M/G/1$ retrial queue) and outgoing calls the behavior of which is as follows. An outgoing call is produced with rate α only when the server is idle (a possible motivation comes from the situation arising when the server is a telephone attended by a person who is also able to make his/her own calls). Here, we generalize the model described in [3] by assuming that ingoing calls and outgoing calls receive different service times. In the sequel B_1 represents the service time of an ingoing call (with residual service time R_1) and B_2 represents the service time of an outgoing call (with residual service time R_2).

Let us further introduce the following notation:

$\rho = \lambda E(B_1), \quad \sigma = \alpha E(B_2),$

W_1^0 : the total idle time of the server during the waiting time in orbit of an ingoing call,

W_1^1 : the total busy time of the server providing service to an ingoing call during the waiting time in orbit of an ingoing call,

W_1^2 : the total busy time of the server providing service to an outgoing call during the waiting time in orbit of an ingoing call.

The system is stable when $\rho < 1$. Once more, P_{ij} denotes the steady-state probabilities of the system state. More concretely, i denotes the state of the server taking the values 0 (server idle), 1 (server providing service to an ingoing call) and 2 (server providing service to an outgoing call), while j denotes the number of customers in orbit. We also consider the partial expectations, $E(L_i) = \sum_{j=0}^{\infty} jP_{ij}$, for $i = 0, 1, 2$.

Now it is straightforward to show (e.g., by applying Little's formula to the server) that the fraction of time the server is occupied by ingoing calls equals ρ, the fraction of time the server is occupied by outgoing calls equals $((1-\rho)\sigma)/(1+\sigma)$ and the fraction of time the server is idle equals $(1-\rho)/(1+\sigma)$.

Hence, the probability an ingoing call enters the orbit equals $\rho + ((1-\rho)\sigma)/(1+\sigma) = (\rho+\sigma)/(1+\sigma)$. So, we conclude

$$E(W_1^0) = \frac{\rho+\sigma}{(1+\sigma)\mu}.$$

The mean value equations for $E(W_1^1)$ and $E(W_1^2)$ are as follows:

$$\begin{aligned} E(W_1^1) &= E(L)E(B_1) + \rho E(R_1), \\ E(W_1^2) &= \frac{\sigma(1-\rho)}{1+\sigma} E(R_2) + \sigma E(W_1^0). \end{aligned}$$

Here, we assume that ingoing calls follow the strict FCFS service discipline. Outgoing calls may directly occupy the server. Using again Little's formula and the above mean value equations, we obtain

$$E(W) = \frac{\rho}{1-\rho}E(R_1) + \frac{\sigma}{1+\sigma}E(R_2) + \frac{\rho+\sigma}{(1-\rho)\mu}.$$

Acknowledgments: Jesus Artalejo thanks the support received from MEC, research project MTM2005-01248. The research of Jacques Resing was done within the framework of the European Network of Excellence Euro-NGI.

References

[1] I.J.B.F. Adan and J.A.C. Resing (2001). *Lecture Notes Queueing Theory*, Eindhoven University of Technology, http://www.win.tue.nl/∼iadan/queueing.pdf.

[2] J.R. Artalejo and G. Choudhury (2004). Steady state analysis of an $M/G/1$ queue with repeated attempts and two-phase service. *Quality Technology & Quantitive Management* **1**, 189–199.

[3] G.I. Falin (1986). Single-line repeated orders queueing systems. *Optimization* **17**, 649–667.

[4] G.I. Falin and J.G.C. Templeton (1997). *Retrial Queues*, Chapman and Hall, London.

[5] S.W. Fuhrmann and R.B. Cooper (1985). Stochastic decompositions in the $M/G/1$ queue with generalized vacations. *Operations Research* **33**, 1117–1129.

[6] A. Gómez-Corral (1999). Stochastic analysis of a single server retrial queue with general retrial times. *Naval Research Logistics* **46**, 561–581.

[7] E.M.M. Winands, I.J.B.F. Adan and G.J. van Houtum (2006). Mean value analysis for polling systems. *Queueing Systems* **54**, 35–44.