Transient Analysis of an M/M/1 Queue with Regularly Changing Arrival and Service Intensities

J.L. van den Berg*

PTT Research, P.O. Box 421, 2280 AK Leidschendam, The Netherlands

W.P. Groenendijk†

Centre for Mathematics and Computer Science, P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

Abstract

This paper presents a simple and elegant numerical method for computing the distribution of the number of customers as a function of time in an M/M/1 queue with regularly changing arrival and service intensities.

1 Introduction

The M/M/1 queue is one of the most intensively studied queueing systems. Obviously this is due to the fact that the model can be analysed relatively easily and often arises in situations where a queueing (sub-)network (via the use of Norton's theorem) is aggregated into a single-server queue. The M/M/1 queue has been completely solved analytically; see for example Cohen [4]. Yet, still many papers appear on the subject of the M/M/1 queue. A number of the more recent papers, cf. Abate and Whitt [1, 2, 3], have focussed on the transient behaviour of the M/M/1 queue, an aspect which is very important for practical applications. In most of the literature it is assumed that the parameters of the system are constant, i.e., do not vary over time. Notable exceptions are the cases where the arrival and/or service intensity depend on the number of customers in the system, see e.g. Gross and Harris [5], or where the arrival intensity is governed by an external Markov process (Markov Modulated Poisson Process, MMPP, cf. Heffes and Lucantoni [6]).

From a practical point of view, the transient behaviour of the M/M/1 queue where the arrival and service intensity are functions of time is very important. For instance the traffic pattern may show peak hours when considered over a certain period in time. The transient behaviour of this model has been analysed before, cf. the book by Saaty [8] and references contained therein. However, the solution given there is not very satisfactory due to its complexity and the corresponding difficulty in obtaining numerical results. In this paper we present another approach to the analysis of the transient behaviour of the M/M/1 queue with time dependent arrival and service intensity. This approach is based on embedding an external stochastic process in time; only at the embedding points the service and arrival intensity may change. We derive a simple and elegant numerical method for computing the distribution of the number of customers in the system as a function of time.

The rest of this paper is organized as follows. First, in Section 2, a formal model description is presented. The analysis is carried out in Section 3. In Section 4 we give a numerical example. Finally, in Section 5, conclusions and subjects for further research are given.

2 Model description

The model under consideration is a single-server queue in continuous time with an infinite buffer capacity. Required service times of customers are drawn from an exponential distribution with mean $\mu$. The customer arrival process is a non-homogeneous Poisson process with parameter $\lambda(t)$. The service intensity $\mu(t)$ of the server also is a function of time. Both the arrival intensity $\lambda(t)$ and the service intensity $\mu(t)$ are step functions which are built up as follows:

Consider the sequence $\{S_i, i = 0, 1, 2, \ldots\}$ of independent random variables, each with a negative exponential distribution with mean $s_i$. Let $S_0 := 0$. For

---

*Part of this research was done while the author worked at the Centre for Mathematics and Computer Science, Amsterdam, The Netherlands.
†Author's present address: Shell Research B.V., P.O. Box 3003, 1003 AA Amsterdam, The Netherlands.
\( \lambda(t) = \lambda_i \) if \( t \in \left[ \sum_{n=0}^{i-1} S_n, \sum_{n=0}^{i} S_n \right] \),
\( \mu(t) = \mu_i \)

where \( \lambda_i, \mu_i \) are (arbitrarily chosen) non-negative real numbers. The interval \( \left[ \sum_{n=0}^{i-1} S_n, \sum_{n=0}^{i} S_n \right], i = 1, 2, \ldots \), is called the 'i-th interval'. It is assumed that at time 0 the system starts with \( j \geq 0 \) customers present.

### Some further notation

The traffic offered to the system during the i-th interval is denoted by \( \rho_i \); clearly \( \rho_i = \lambda_i \beta / \mu_i \). The random variable \( X_i \) denotes the number of customers in the system at the end of the i-th interval.

In the next section we shall derive the generating function of the distribution of \( X_i, i = 1, 2, \ldots \).

The essential part in the approach, which will greatly facilitate the analysis, is the embedding of the external stochastic process \( \{ S_i, i = 0, 1, \ldots \} \) which governs the transitions in the arrival and service intensities. Our approach may still lead to an accurate representation of real-time behaviour; this is due to the fact that the sum of \( k \) intervals with the same mean length has a squared coefficient of variation equal to \( 1/k \), which can be made arbitrarily small by increasing the number of intervals \( k \).

### Remark

In many practical situations the behaviour of a system is (stochastically spoken) repeated during successive periods. For instance, think of a telephone information service which can be reached from, e.g., 9.00 A.M. until 6.00 P.M.. The behaviour of such a system during each period may be modelled by the M/M/1 queue described above but with a finite number, say \( N \), of successive intervals (i.e. each period consists of \( N \) intervals). If the behaviour of the system during a certain period is independent of its behaviour during the other periods, then it can be shown that the stationary number of customers in the system at an arbitrary epoch during an i-th interval is equal to the distribution of \( X_i \) - the number of customers in the system at the end of interval \( i \), \( i = 1, \ldots, N \) (note that the interval lengths are exponentially distributed).

### Analysis

In this section we analyse the model described in Section 2. In particular we derive an expression for the generating function \( G_i(\cdot) \) of the distribution of the number of customers \( X_i \) in the system at the end of interval \( i \), \( i = 1, 2, \ldots \). We have, for \( |r| < 1 \),

\[
G_{i+1}(r) := E\{ r^{X_{i+1}} \} := \sum_{j=0}^{\infty} r^j \Pr \{ X_{i+1} = j \} \]
\[
= \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} r^j \Pr \{ X_{i+1} = j | X_i = k \} \Pr \{ X_i = k \}, \quad i = 0, 1 \ldots .
\]

Now let \( P^{(i)}_{kj}(t), k,j = 0,1, \ldots, \) be the transition probabilities of the queue length process of a standard M/M/1 queue with arrival rate \( \lambda_i \) and service rate \( \mu_i \), i.e. \( P^{(i)}_{kj}(t) \) is the probability that starting with \( k \) customers in the M/M/1 system at time zero there are \( j \) customers in the system at time \( t \). It is easily seen that (note that both the service times and the interval lengths are exponentially distributed)

\[
Pr \{ X_{i+1} = j | X_i = k \} = P^{(i)}_{kj}(t) dt
\]
\[
\int_{t=0}^{\infty} P^{(i+1)}_{kj}(t) dPr \{ S_{i+1} < t \} =
\]
\[
(1/s_{i+1}) \int_{t=0}^{\infty} e^{-t/s_{i+1}} P^{(i+1)}_{kj}(t) dt, \quad i = 0, 1, \ldots .
\]

The integral in the right hand side of (2) can be viewed as the Laplace transform of \( P^{(i+1)}_{kj}(t) \) evaluated at \( 1/s_{i+1} \). Denoting this Laplace transform by \( \Pi^{(i+1)}_{kj}(\cdot) \) and substituting (2) into (1) we obtain

\[
G_{i+1}(r) = \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} r^j (1/s_{i+1}) \Pi^{(i+1)}_{kj}(1/s_{i+1}) \Pr \{ X_i = k \} =
\]
\[
(1/s_{i+1}) \sum_{k=0}^{\infty} Pr \{X_i = k\} \Pi_k^{(i+1)}(r, 1/s_{i+1}),
\]

where

\[
\Pi_k^{(i+1)}(r, \omega) := \sum_{j=0}^{\infty} r^j \Pi_k^{(i+1)}(\omega),
\]

\[Re \omega \geq 0, \ |r| < 1.\]

The generating function \(\Pi_k^{(i+1)}(r, \omega)\) of the sequence \(\Pi_k^{(i+1)}(\omega), j = 0, 1, \ldots,\) is known from the theory on the transient behaviour of the standard \(M/M/1\) queue. From the results in Cohen [4, Section I.4.4] it is found that, for \(i = 0, 1, \ldots,\)

\[
\Pi_k^{(i+1)}(r, \omega) = \frac{\beta}{\mu_{i+1}} \times
\]

\[
\frac{(1 - r)\{z_2(i + 1)\}^{k+1} - (1 - z_2(i + 1))r^{k+1}}{\rho_{i+1}\{1 - z_2(i + 1)\}\{r - z_2(i + 1)\}}.
\]

where \(z_1(i + 1)\) and \(z_2(i + 1)\) are the roots of the following equation in \(x:\)

\[\rho_{i+1}x^2 - (1 + \rho_{i+1} + \omega \beta/\mu_{i+1})x + 1 = 0.\]

It can be shown that for \(\omega \beta/\mu_{i+1} > 0\) one zero is in \([0,1]\) and the other in \([1,\infty)\) (see Cohen [4]); henceforth it will be assumed that \(z_1(.) \geq 1, 0 \leq |z_2(.)| < 1, i.e.,

\[
z_1(i + 1) = \frac{1 + \rho_{i+1} + \omega \beta/\mu_{i+1}}{2 \rho_{i+1}} + \frac{1}{2 \rho_{i+1}} \left\{\left(1 + \rho_{i+1} + \omega \beta/\mu_{i+1}\right)^2 - 4 \rho_{i+1}\right\}^{1/2},
\]

\[
z_2(i + 1) = \frac{1 + \rho_{i+1} + \omega \beta/\mu_{i+1}}{2 \rho_{i+1}} - \frac{1}{2 \rho_{i+1}} \left\{\left(1 + \rho_{i+1} + \omega \beta/\mu_{i+1}\right)^2 - 4 \rho_{i+1}\right\}^{1/2}.
\]

Substitution of (5) into (3) (take in (5) \(\omega = 1/s_{i+1}\)) leads after some algebra to the following recurrence relation for the \(G_i(.)\)'s:

\[
G_{i+1}(r) = \frac{1}{\lambda_{i+1}s_{i+1}\{r - z_1(i + 1)\}\{r - z_2(i + 1)\}} \times
\]

\[
\left\{\left(1 - r\right)z_2(i + 1)G_i(z_2(i + 1)) - rG_i(r)\right\},
\]

\[i = 0, 1, \ldots.\]

Now starting with \(G_0(r) \equiv r^i\) (i.e. \(j \geq 0\) customers in the system at time zero) \(G_1(r), G_2(r), \ldots\) can be successively obtained from this recurrence relation. For numerical calculation of these generating functions it is convenient to realize that relation (7) can be written as

\[
G_{i+1}(r) = \frac{P_{i+1}(r)}{Q_{i+1}(r)}, \quad i = 0, 1, \ldots,
\]

with \(P_{i+1}(r), Q_{i+1}(r)\) polynomials in \(r\) given by, for \(i = 0, 1, \ldots,\)

\[
P_{i+1}(r) = R_{i+1}^{(1)}(r)Q_i(r) - rP_i(r),
\]

\[
Q_{i+1}(r) = R_{i+1}^{(2)}(r)Q_i(r),
\]

where

\[
R_{i+1}^{(1)}(r) = \frac{(1 - r)z_2(i + 1)G_i(z_2(i + 1))}{1 - z_2(i + 1)},
\]

\[
R_{i+1}^{(2)}(r) = \lambda_{i+1}s_{i+1}\{r - z_1(i + 1)\}\{r - z_2(i + 1)\}.
\]

From (8)-(10) it is easily seen that a closed form expression for \(G_{i+1}(r)\) can be obtained by successive multiplication of polynomials in \(r\). Note that the denominator of \(G_{i+1}(r), Q_{i+1}(r),\) is simply given by

\[
Q_{i+1}(r) = \prod_{j=1}^{i+1} R_{j+1}^{(2)}(r),
\]

\[
\prod_{j=1}^{i+1} \lambda_j s_j \{r - z_1(j)\} \{r - z_2(j)\},
\]

which can be calculated beforehand (i.e. without the prior knowledge of \(G_1(r), \ldots, G_i(r),\) cf. (6)).

An important aspect of the above calculation scheme for \(G_{i+1}(r)\) is that multiplication of polynomials can
be very easily implemented on a computer using sym­bolic manipulation. Moreover, for obtaining mo­ments (or individual probabilities), the expressions
\[ P_{i+1}(r)/Q_{i+1}(r) \]
for \( G_{i+1}(r) \), \( i = 0, 1, \ldots \), can be easily differentiated.

It is easily seen that \( P_{i+1}(r) \) and \( Q_{i+1}(r) \) are polyno­mials of degree \( \max\{j + i + 1, 2i + 1\} \) and \( 2(i + 1) \)
respectively. It appears that the degree of these
polynomials can be reduced, which is convenient for
numerical evaluation. Indeed, from the fact that
\( G_{i+1}(r) \) is analytic within the unit circle and the ze­ros \( z_2(1), \ldots , z_2(i+1) \) of \( Q_{i+1}(r) \) are smaller than one
in this region (cf. the discussion above (6)) it follows
that also the numerator must be zero in these points.
Hence, \( G_{i+1}(r) \) can be written as

\[
G_{i+1}(r) = \frac{\hat{P}_{i+1}(r)}{\hat{Q}_{i+1}(r)}, \quad i = 0, 1, \ldots ,
\]

with \( \hat{P}_{i+1}(r) \) and \( \hat{Q}_{i+1}(r) \) polynomials of degree
\( \max\{j, i\} \) and \( i + 1 \) respectively,

\[
\hat{P}_{i+1}(r) = P_{i+1}(r)/\prod_{j=1}^{i+1}(r - z_2(j)),
\]

\[
\hat{Q}_{i+1}(r) = Q_{i+1}(r)/\prod_{j=1}^{i+1}(r - z_2(j)) =
\]

\[
\prod_{j=1}^{i+1} \lambda_j s_j \{r - z_1(j)\}.
\]

The expression for \( G_{i+1}(r) \) given by (12) is suitable
for carrying out a partial fraction expansion, which is
useful for inverting \( G_{i+1}(r) \) (for the inversion of prob­ability generating functions by the method of partial
fraction expansion see e.g. Kleinrock [7, Appendix 1]).
For example, if in (12) the zeros \( z_2(1), \ldots , z_2(i+1) \)
of \( Q_{i+1}(r) \) (given by (6)) are all distinct and the de­gree of \( \hat{P}_{i+1}(r) \) is smaller than that of \( \hat{Q}_{i+1}(r) \), then
\( G_{i+1}(r) \) can be written as

\[
G_{i+1}(r) = \frac{\hat{P}_{i+1}(r)}{\hat{Q}_{i+1}(r)} =
\]

\[
\frac{a_1}{r - z_1(1)} + \frac{a_2}{r - z_1(2)} + \ldots + \frac{a_{i+1}}{r - z_1(i+1)},
\]

with \( a_j = \hat{P}_{i+1}(z_2(j))/\hat{Q}_{i+1}(z_2(j)), \ j = 1, \ldots , i + 1. \)
So, in that case the distribution of \( X_{i+1} \) (determined
by \( G_{i+1}(r) \)) is given by:

\[
Pr\{X_{i+1} = j\} =
\]

\[
- \left\{ \frac{a_1}{(z_1(1))^{j+1}} + \ldots + \frac{a_{i+1}}{(z_1(i+1))^{j+1}} \right\},
\]

\[ j = 1, 2, \ldots . \]

The structure of the generating functions \( G_{i+1}(r) \), \( i = 0, 1, \ldots \), is such that we can always use (possibly a variant of) the above inversion method (see e.g. Klein­rock [7]).

**Remark**

In most problems of interest the difficult part of the
inversion method using partial fraction expansion is
to find the zeros of the denominator polynomial; in
our case the zeros \( z_1(1), \ldots , z_1(i+1) \) of the denomi­nator \( Q_{i+1}(r) \) are explicitly
given by (6).

In the next section we shall discuss an example.

### 4 Numerical example

Consider a service facility with one server, so that the
system can be modelled as a single-server queue. As­sume that the customer arrival process is a Poisson
process and required service times of customers have
a negative exponential distribution. For simplicity, let
\( \beta = 1, \mu(t) = 1, t \geq 0. \)
We are interested in the number of customers in the
system as a function of time over a particular period,
say one day. Applying the model as described in Sec­tion 2, we split this one day into, e.g., 48 intervals of
on average ten minutes each (making a total of eight
hours). Suppose the arrival intensity is as depicted
in Figure 1; on the horizontal axis the intervals are
given. The \( \lambda_i \)'s are sampled from the corre­sponding intervals. Note that between the 20-th and the 28-th
interval the stability limit is exceeded.
From the analysis in the previous section (in partic­ular from (8)-(10) with the simplification suggested
by (12) and (13)) we have calculated the mean and
variance of the number of customers over the period
of one day. The results are presented in Figure 2.
The whole calculation of the generating functions
\( G_1(r), \ldots , G_{48}(r) \) and their moments took only a few
seconds on a personal computer.
optimization problem for the $\mu_i$'s. Because the generating functions of the number of customers in the system are given by a recursive scheme a global optimization of the $\mu_i$'s isn't feasible; we'll have to do a one step look ahead and formulate the problem as a Markov decision process which can be solved by dynamic programming.

We are also investigating a limiting process for the intervals $s_i$. If we let the mean interval length go to zero, then under some assumptions for the $\lambda_i$'s and $\mu_i$'s we obtain a very interesting integral equation for the generating function of the number of customers in the system at time $t$.

This integral equation is currently being investigated.

References


