# Approximate Analysis of an M/G/1 Priority Queue with Priority Changes due to Impatience

P. de Waal

Department of Operations Reasearch, Statistics, and System Theory

**BS-R9202 1992**

# Approximate Analysis of an M/G/1 Priority Queue with Priority Changes due to Impatience

Peter de Waal
*CWI*
*P.O.Box 4079, 1009 AB Amsterdam, The Netherlands*

We discuss the approximate analysis of an M/G/1 queue where low priority customers may flow over to become high priority customers due to impatience. The first approximation method uses a standard M/G/1 priority queue in which the arrival rates are adjusted according to an estimation of the fraction of overflowing customers. The second approximation method consists of a brute force calculation of the equilibrium distribution of the queueing system with phase type service time distributions, using the method of stages and truncation of the state space.

*AMS Mathematics Subject Classification:* 60K25, 90B22
*Keywords & Phrases:* Impatient customers, priority queues, deadlines
*Note:* This author is supported by a grant of Koninklijke/Shell Laboratorium, Shell Research B.V., Amsterdam.

## 1 Introduction

In this paper we present a queueing model for corrective and preventive maintenance of a large number of components in a technical installation like for instance a plant at an oil refinery or an offshore platform. For such an installation an estimation of the maintenance workload is made on the basis of, among other things, design information of the installation, various repair strategies and operational regimes. This estimation should account for both unplannable or emergency maintenance jobs and plannable or preventive maintenance. We thus distinguish between these two different types of jobs, corrective maintenance (CM) and preventive maintenance (PM). CM corresponds to repair of components that have broken down, and PM is performed on a component to prevent it from breaking down.

The maintenance jobs are carried out by a maintenance crew. Since corrective maintenance on a component is usually associated with unavailability of the installation, it is carried out at a higher priority than preventive maintenance. The manning level of the maintenance crew determines the waiting time of maintenance jobs, which is usually referred to as *backlog.* This backlog should not become too large, since that may cause components that are waiting for preventive maintenance to break down before the job is executed. A long backlog for preventive maintenance may also lead to the violation of safety requirements, forcing an immediate execution of the job at a priority comparable to corrective maintenance. When

Report BS-R9202
ISSN 0924-0659
CWI
P.O.Box 4079, 1009 AB Amsterdam, The Netherlands

this happens, then the waiting time of other preventive maintenance jobs may increase, and this can eventually lead to almost all PM jobs turning into CM.

A large backlog may thus have a significant impact on the balance between PM and CM, and thus on the unavailability of the installation. Another reason to keep the backlog sufficiently small is the fact that CM, as opposed to PM, is by its nature very 'unplannable' and adjustments of the maintenance manpower may have a considerable impact on total costs.

For the performance of corrective and preventive maintenance we consider an M/G/1 queueing model with two priority classes. The high priority class corresponds to CM and the low priority customers represent PM jobs. Both types of jobs are generated by two independent Poisson sources. The maintenance crew is represented by a single server, whose speed is proportional to the number of crew members. CM jobs are served with non-preemptive or preemptive-resume priority over PM. Each PM job has a deadline associated with it. If the deadline of a PM job expires before his service has begun, then the job becomes a CM job, i.e. it leaves the waiting queue of PM and joins the CM queue.

The category of queueing models with impatient or *reneging* customers was first introduced by PALM [12]. In more recent literature we can classify two groups of papers on queueing models with impatient customers. The first group deals with the evaluation of various performance measures of queues with impatient customers, while the second group focuses on optimal control problems in queues with impatient customers. Examples of the first group are BACCELLI *et al.* [1], and STANFORD [15, 16]. They present a thorough analysis of for instance ergodicity conditions and waiting time distributions. In BACCELLI AND TRIVEDI [2] a transient analysis is presented of a system which stops as the first customer becomes impatient. Many practically oriented papers in this group arise in situations involving impatient telephone customers, especially in the area of overload control. Examples are DOSHI AND HEFFES [8], FAYOLLE AND BRUN [9], FORYS [10] and SZE [17]. An early overview of various queueing models with impatience can also be found in SAATY [14].

The second group of papers deals with various issues of optimal control of a queue with impatient customers. DE WAAL [20], [21, Chapter 4] and BLANC *et al.* [5] discuss the problem of optimal admission to a FCFS queue in order to maximize the discounted and longrun average reward associated with the departure of successful customers. Another topic in optimal control of queues with impatient customers concerns the scheduling of customers. In fact DOSHI AND HEFFES [8] and FORYS [10] already touch on this problem, since they compare various scheduling disciplines as overload control mechanisms in telephone switches. In PANWAR *et al.* [13] it is shown that the *Shortest Time to Extinction* (STE) policy is optimal for a class of continuous and discrete time nonpreemptive M/G/1 queues. Similar results are presented in BHATTACHARYA AND EPHREMIDES [3, 4]. Typical for most of the models is that reneging customers either leave the system or are allowed to be removed. In CHEN AND TOWSLEY [6] the optimal scheduling policy is given for a model in which all customers have to be served eventually.

An important difference of the model in this paper with the existing literature is the fact that a customer who reneges can affect the waiting time of other customers of the same type *including customers who arrived before him.* This phenomenon makes it, for instance, impossible to derive a recursive formula for the virtual waiting time as in BACCELLI *et al.* [1] or STANFORD [15]. Moreover, since deadlines may be random, even the order of PM jobs that change into CM jobs is not necessarily preserved.

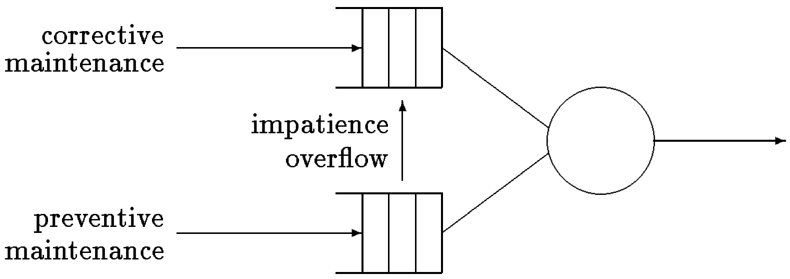

FIGURE 1. The queueing system

In the case of exponentially distributed service times we can write down a set of functional equations for the generating function of the joint queue length distribution. This set of differential equations for the bivariate generating function involves a partial derivative due to the impatience overflow. This problem can be reduced to a boundary value problem of Riemann–Hilbert type, that we have not attempted to solve yet. Instead we have developed two approximation methods for the computation of various performance measures. The first method uses a standard M/G/1 priority queue without overflow where the overflow process of the real model is approximated by adjusting the arrival rates. The second approximation method, developed for the original model with phase type service time distributions, uses brute force calculation of the equilibrium joint queue length distribution with a truncated state space.

This paper is organised as follows. In Section 2 we give a description of the model. The two approximation methods are introduced in Sections 3 and 4. Numerical results are presented in Section 5 and conclusions and future research directions are sketched in Section 6.

## 2 DESCRIPTION OF THE MODEL

Consider the queueing system as depicted in Figure 1. We distinguish two types of customers, type $C$ denoting corrective maintenance tasks and type $P$ denoting preventive maintenance tasks. We assume that tasks are generated by two independent Poisson processes with arrival rates $\lambda_C$ and $\lambda_P$, respectively.

The service times of customers are independent and they may have a distribution of arbitrary type. Let $B_C$ and $B_P$ denote the distribution functions of the service times of types $C$ and $P$. We also introduce $\mathfrak{B}_C$ and $\mathfrak{B}_P$ as the Laplace-Stieltjes transform of $B_C$ and $B_P$. Furthermore we let $\beta_C$ and $\beta_P$ denote the mean service times and $\beta_C^{(2)}$ and $\beta_P^{(2)}$ the second moments.

Customers are handled by one server that works according to either a non-preemptive or preemptive resume priority rule, where type $C$ has a higher priority than type $P$. Customers of the same priority are served in the order of arrival.

As was described in the Introduction preventive maintenance tasks are indeed served at a low priority provided that their service is not delayed for too long. When the waiting time of a preventive maintenance task exceeds some threshold, however, then the execution of that task becomes urgent and it should thus get high priority. We model this as follows. The value of the threshold, which we shall call *deadline*, is assumed to be a random variable $D$, that has distribution function $F_D$ and mean $\beta_D$. Deadlines of type $P$ customer are mutually independent and independent of the interarrival and service times. Consider now an arbitrary customer of type $P$ who arrives at time $t$ and has deadline $D$. If the service of this particular customer has not commenced at time $t + D$, then at that time the customer will leave the queue of low priority customers. Subsequently he joins the high priority queue at the end, i.e. behind all the customers that are already present there. He will there become a regular type $C$ customer, so his service time will have distribution function $B_C$. This switching of type will be called *impatience overflow* or simply *overflow*.

## 3 Approximation Method I

In this section we suggest a method to approximate the stationary fraction of type $P$ customers whose service is delayed for too long and who thus become type $C$. We denote this fraction as $q$, the *overflow probability*. The key idea behind the approximation method is that some aspects of the stationary behaviour of the model of Section 2 resemble those of an ordinary M/G/1 priority queue without deadlines, where the overflow process is replaced by an additional Poisson source with rate $q\lambda_P$. More precisely, the approximate model will be an ordinary M/G/1 queue with (preemptive resume or non-preemptive) priority and the same service time parameters as the original model, but with arrival rates $\lambda'_C$ and $\lambda'_P$ chosen as

$$\lambda'_C = \lambda_C + q\lambda_P,$$

$$\lambda'_P = (1-q)\lambda_P.$$

For convenience we shall denote this queueing model as $M_q$, reflecting the dependence of the arrival parameters on $q$. Of course the Poisson assumption for the overflow process is an approximation, but we expect it to provide good approximations for a system that is only lightly loaded. We shall come back to this in the discussion of numerical results in Section 5.

Thus remains the problem of finding a good choice for $q$. For this we need some additional notation. Let $W_q$ denote the (stationary) waiting time for type $P$ customers in model $M_q$, $F_{W_q}$ its distribution function and let $\mathfrak{F}_{W_q}$ denote the Laplace-Stieltjes transform of $F_{W_q}$. Furthermore let $\epsilon$ be a sufficiently small positive number. The approximation for $q$ will be computed by the following iterative procedure. Let $q_i \in [0,1]$ be the $i$-th approximation for $q$. In model $M_{q_i}$ we compute the fraction $q_{i+1}$ of type $P$ customers whose waiting time is larger than the deadline $D$, i.e. $q_{i+1} = \mathbb{P}(W_{q_i} > D)$. If $q_{i+1}$ is sufficiently close to $q_i$, i.e. $|q_i - q_{i+1}| < \epsilon$, then we take $q_i$ as an approximation for $q$, otherwise we repeat the same procedure with $q_i$ replaced by $q_{i+1}$.

Algorithm

0. Let $q_0 \in [0,1]$, $i := 0$.

1. Compute in the M/G/1 priority queueing model $M_{q_i}$:

$$q_{i+1} := \mathbb{P}(W_{q_i} > D).$$

2. If $|q_i - q_{i+1}| < \epsilon$, then STOP, else $q_i := q_{i+1}$, $i := i + 1$, and go to 1.

We conclude the description of the approximation method with the details on the computation of $\mathbb{P}(W_q > D)$. This is presented for the case where $D$ has an Erlang distribution. A similar procedure can be constructed along the same lines for hyperexponential distributions, or mixtures of Erlang and hyperexponential. For deterministic deadlines we can use the approximations for the waiting time percentiles of TIJMS [19].

Assume that $D$ has an Erlang-$k$ distribution with mean $\beta_D = \frac{k}{\mu}$, for some $k \in \mathbb{N}$, $\mu > 0$, so

$$F_D(t) = 1 - \sum_{n=0}^{k-1} \frac{(\mu t)^n}{n!} e^{-\mu t}.$$

The computation of $\mathbb{P}(W_q > D)$ is now straightforward:

$$\begin{aligned} 1 - \mathbb{P}(W_q > D) &= \int_0^\infty (1 - F_D(t)) dF_{W_q}(t) \\ &= \int_0^\infty \sum_{n=0}^{k-1} \frac{(\mu t)^n}{n!} e^{-\mu t} dF_{W_q}(t) \\ &= \sum_{n=o}^{k-1} \frac{(-\mu)^n}{n!} \int_0^\infty (-t)^n e^{-\mu t} dF_{W_q}(t) \\ &= \sum_{n=o}^{k-1} \frac{(-\mu)^n}{n!} \frac{d^n}{d\mu^n} \mathfrak{F}_{W_q}(\mu) \end{aligned} \tag{1}$$

Note that the stationary waiting time $W_q$ — the time between the moment of arrival and the beginning of service — has the same distribution for both the non-preemptive and preemptive resume priority discipline. Furthermore, $\mathfrak{F}_{W_q}$ is well known from literature and can be found in many queueing theory textbooks, e.g. COHEN [7, p.450]:

$$\mathfrak{F}_{W_q}(\mu) = \frac{(1-a)[\mu + (1 - \zeta_C(\mu,1))/\alpha_C]\, \alpha_P}{-1 + \mu\alpha_P + \mathfrak{B}_P(\mu + (1 - \zeta_C(\mu,1))/\alpha_C)} \tag{2}$$

where

$$\alpha_C = \frac{1}{\lambda'_C},$$

$$\alpha_P = \frac{1}{\lambda'_P},$$

$a$ is the workload

$$a = \lambda'_C \beta_C + \lambda'_P \beta_P,$$

and $\zeta_C(\mu, 1)$ is the zero $z$, with the smallest absolute value, of

$$z - \mathfrak{B}_C(\mu + (1 - z)/\alpha_C),$$

i.e. $\zeta_C(\mu, 1)$ is the L.S.T. of the busy period of an M/G/1 queue with mean interarrival time $\alpha_C$ and service time L.S.T. $\mathfrak{B}_C$.

Note that this approximation method cannot be used to compute waiting time characteristics. Of course we can compute the moments of the waiting times in the approximate M/G/1 model with two Poisson arrival streams, but the overflow process in the original model will be so irregular, that a Poisson process will be a very bad approximation. Moreover, an overflow usually occurs when the waiting times are large, and the overflowing customer will increase the waiting time of the remaining PM jobs, since it will become a high priority customer. From this we may conjecture that impatient customers will, on the average, see large waiting times, and that the overflow process will be very bursty. Ordinary CM jobs will also experience longer waiting times than in the $M_q$ model. These assumptions have indeed been verified by test data from simulations.

Numerical results for this approximation method will be discussed in Section 5.1.

## 4 Approximation Method II

The second approximation method is proposed for the model of Section 2 in which the service times have a phase type distribution. Furthermore we assume that the deadlines of low priority (PM) jobs are exponentially distributed. We shall describe this method only for the non-preemptive priority discipline. The preemptive resume priority discipline can be dealt with in a similar manner by an appropriate adjustment of the state space.

It is known from literature (cf. for instance Kleinrock [11, Section 4.2]) that queueing models with phase type distribution functions can be represented as a Markov process by the method of stages. We shall explain this in our model for the case of Coxian distributions. Assume that the service time distributions of customer type $i$, $i = C$, $P$, have a Coxian distribution of order $k_i$, with transition rates $\nu_n^i$ and transition probabilities $p_n^i$, $n = 1, \ldots, k_i$. The state of the underlying Markov queueing process can be represented as

$$s = (l_C, l_P, m, j),$$

where $l_i \in \mathbb{N}$ denotes the number of waiting customers of type $i$, $i = C$, $P$, $m =$ the type of the customer in service, and $j$ the phase of this customers service. The extra notation $s = \emptyset$ denotes an empty system. In Table 1 we show all the possible transitions from state $s = (l_C, l_P, m, j)$ with the corresponding transition rates.

The method of stages allows an exact formulation as a Markov process. The approximation that we propose consists in truncating the state space of this Markov process. In particular we truncate the queue lengths of CM and PM jobs, by rejecting arrivals when the respective lengths exceed $K_C$ and $K_P$, for some $K_C$ and $K_P > 0$. In addition we prohibit the overflow of an impatient customer from the PM queue to the CM queue if the queue length $l_C$ is equal to $K_C$. The differences in the transitions as compared to the original model are summarized in Table 2.

One can expect that the errors introduced by this truncation will be small if in the original model most of the probability mass of the equilibrium distribution is concentrated in states that have a small queue length. We may therefore conclude that the approximation method will fail in heavily loaded queueing systems and will behave well under light or moderate loads. The equilibrium distribution is computed as the solution of the Kolmogorov equations using Gauss–Seidel iteration (see for instance Tijms [18, Appendix B]).

| new state | transition rate | transition type | conditions |
|---|---|---|---|
| $(l_C, l_P, m, j+1)$ | $\nu_j^m p_j^m$ | phase completion | |
| $(l_C - 1, l_P, C, 1)$ | $\nu_j^m (1 - p_j^m)$ | service completion | $l_C > 0$ |
| $(l_C, l_P - 1, P, 1)$ | $\nu_j^m (1 - p_j^m)$ | service completion | $l_C = 0, l_P > 0$ |
| $\emptyset$ | $\nu_j^m (1 - p_j^m)$ | service completion | $l_C = l_P = 0$ |
| $(l_C + 1, l_P, m, j)$ | $\lambda_C$ | arrival type $C$ | |
| $(l_C, l_P + 1, m, j)$ | $\lambda_P$ | arrival type $P$ | |
| $(l_C + 1, l_P - 1, m, j)$ | $l_P/\beta_D$ | impatience overflow | $l_P > 0$ |

TABLE 1. Transitions from state $(l_C, l_P, m, j)$ in the original model

| new state | transition rate | transition type | conditions |
|---|---|---|---|
| $(l_C + 1, l_P, m, j)$ | $\lambda_C$ | arrival type $C$ | $l_C < K_C$ |
| $(l_C, l_P + 1, m, j)$ | $\lambda_P$ | arrival type $P$ | $l_P < K_P$ |
| $(l_C + 1, l_P - 1, m, j)$ | $l_P/\beta_D$ | impatience overflow | $l_P > 0, l_C < K_C$ |

TABLE 2. New transitions from state $(l_C, l_P, m, j)$ in the approximation model

The method provides also an indication of the accuracy of the approximation. In the approximate model we can compute the total equilibrium probability mass of the set

$$\mathfrak{S} := \{(l_C, l_P, m, j) \mid l_C = K_C \text{ or } l_P = K_P\} \qquad (3)$$

and use this as an indication of the approximation error. If the equilibrium probability mass on the set is large, then $K_C$ and $K_P$ have to be increased.

We shall discuss numerical results for this approximation method in Section 5.2.

## 5 NUMERICAL RESULTS

In this section we present numerical results for the two approximation methods that we introduced in Section 3 and 4. Both methods were tested on a model in which the parameters are based on real life data of a small production platform. The arrival rate of corrective maintenance is 225 jobs per year, and for preventive maintenance 275 jobs per year. A CM job requires on average 10 manhours of work, a PM jobs on average 20 manhours. One manyear totals 2000 manhours, i.e. one repairman can work for 2000/365 hours per day. The mean of the deadline distribution is 30 days. We normalize the time axis by setting one day equal to one time unit, so $\lambda_C = 225/365$, $\lambda_P = 275/365$, $\beta_C = 3650/(r * 2000)$, $\beta_P = 3650/(r * 1000)$ and $\beta_D = 30$, where $r$ is the number of available repairmen. Note that the workload $a = 3.875/r$, so the minimally required manning level is $r = 4$. Recall that the manning level is reflected in the speed of the server, not in the number of servers. This is held constant equal to one.

### *5.1 Numerical results for approximation method I*

In the example that we used to test approximation method I we consider exponentially distributed service times. For the deadline distributions we consider the following four cases:

| | *Service time distribution* | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Exponential | | Erlang-2 | | Hyper-2-exp.(1) | | Hyper-2-exp.(2) | |
| $r$ | Approx. | Simul. | Approx. | Simul. | Approx. | Simul. | Approx. | Simul. |
| 4 | 2.27e-01 | 1.88e-01 | 1.81e-01 | 1.32e-01 | 2.58e-01 | 2.16e-01 | 3.34e-01 | 2.88e-01 |
| 5 | 7.43e-02 | 6.83e-02 | 2.79e-02 | 2.47e-02 | 9.48e-02 | 8.61e-02 | 1.52e-01 | 1.38e-01 |
| 6 | 3.53e-02 | 3.37e-02 | 7.78e-03 | 7.74e-03 | 4.67e-02 | 4.45e-02 | 8.22e-02 | 7.82e-02 |
| 7 | 2.07e-02 | 2.03e-02 | 3.17e-03 | 3.21e-03 | 2.78e-02 | 2.72e-02 | 5.07e-02 | 4.87e-02 |
| 8 | 1.36e-02 | 1.34e-02 | 1.59e-03 | 1.52e-03 | 1.84e-02 | 1.81e-02 | 3.44e-02 | 3.34e-02 |
| 9 | 9.70e-03 | 9.75e-03 | 9.10e-04 | 8.92e-04 | 1.31e-02 | 1.31e-02 | 2.48e-02 | 2.45e-02 |
| 10 | 7.20e-03 | 6.93e-03 | 5.68e-04 | 5.86e-04 | 9.80e-03 | 1.00e-02 | 1.87e-02 | 1.85e-02 |
| 11 | 5.60e-03 | 5.68e-03 | 3.77e-04 | 4.16e-04 | 7.60e-03 | 7.51e-03 | 1.46e-02 | 1.45e-02 |
| 12 | 4.50e-03 | 4.49e-03 | 2.63e-04 | 2.68e-04 | 6.10e-03 | 6.13e-03 | 1.17e-02 | 1.16e-02 |

TABLE 3. Fraction of overflowing customers based on approximation method I.

**Exponential** $F_D(t) = 1 - e^{-t/\beta_D}$,

**Erlang–2** $F_D(t) = 1 - e^{-2t/\beta_D} \times \left(1 + \frac{2t}{\beta_D}\right)$

**Hyper-2-exponential (1)** $F_D(t) = 1 - 0.1 \times e^{-t/4\beta_D} - 0.9 \times e^{-3t/2\beta_D}$,

**Hyper-2-exponential (2)** $F_D(t) = 1 - 0.1 \times e^{-t/7\beta_D} - 0.9 \times e^{-3t/\beta_D}$.

The results for the computation of $q$, the fraction of PM jobs that flow over to CM jobs, are presented in Table 3. They are compared with estimates of this fraction as derived from simulation. The approximations for large values for $r$ are fairly good: relative errors range from 5% for $r = 6$ to 2% for $r = 12$. For small values of $r$ (i.e. a high load) the maximum relative errors are 37% for $r = 4$ and 13% for $r = 5$ (both for Erlang-2 distributed deadlines). The number of iterations that we needed to satisfy the convergence criterion with $\epsilon = 0.0001$ varied from 30 for $r = 4$ to 3 for $r = 12$.

### *5.2 Numerical results for approximation method II*

For an evaluation of approximation method II we consider the same model of the production platform. In this example the deadlines have an exponential distribution with mean $\beta_D = 30$, and we vary the type of the service time distributions. We consider the following cases:

**Exponential** $B_i(t) = 1 - e^{-rt/\beta_i}$, $i = C, P$,

**Erlang-k** $B_i(t) = 1 - \sum_{n=0}^{k-1} \frac{(rt)^n}{n!(\beta_i)^n} e^{-rt/\beta_i}$, $k = 2, 5, 10$, $i = C, P$,

**Hyper-2-exponential (1)** $B_i(t) = 1 - 0.9 \times e^{-rt/0.1\beta_i} - 0.1 \times e^{-rt/9.1\beta_i}$, $i = C, P$,

**Hyper-2-exponential (2)** $B_i(t) = 1 - 0.9 \times e^{-rt/0.5\beta_i} - 0.1 \times e^{-rt/5.5\beta_i}$, $i = C, P$.

The fraction $q$ of overflowing customers, for various values of the manning level $r$, is presented in Tables 4 and 5. It is apparent from these tables that the approximation method performs very well for all values of $r$. We must remark here that the errors in the approximations for larger values of $r$ are mainly due to the difficulties in getting accurate figures from the simulations.

| | *Service time distribution* | | | | | |
|---|---|---|---|---|---|---|
| | Exponential | | Erlang-2 | | Erlang-5 | |
| $r$ | Approx. | Simul. | Approx. | Simul. | Approx. | Simul. |
| 4 | 1.88e-01 | 1.88e-01 | 1.64e-01 | 1.63e-01 | 1.47e-01 | 1.47e-01 |
| 5 | 6.85e-02 | 6.83e-02 | 5.47e-02 | 5.47e-02 | 4.56e-02 | 4.56e-02 |
| 6 | 3.41e-02 | 3.37e-02 | 2.64e-02 | 2.65e-02 | 2.16e-02 | 2.16e-02 |
| 7 | 2.03e-02 | 2.03e-02 | 1.55e-02 | 1.54e-02 | 1.26e-02 | 1.28e-02 |
| 8 | 1.35e-02 | 1.34e-02 | 1.03e-02 | 1.01e-02 | 8.28e-03 | 8.18e-03 |
| 9 | 9.60e-03 | 9.75e-03 | 7.28e-03 | 7.06e-03 | 5.87e-03 | 5.92e-03 |
| 10 | 7.19e-03 | 6.93e-03 | 5.44e-03 | 5.35e-03 | 4.38e-03 | 4.42e-03 |
| 11 | 5.59e-03 | 5.68e-03 | 4.22e-03 | 4.22e-03 | 3.39e-03 | 3.46e-03 |
| 12 | 4.47e-03 | 4.49e-03 | 3.37e-03 | 3.44e-03 | 2.71e-03 | 2.72e-03 |

TABLE 4. Fraction of overflowing customers based on approximation method II.

| | *Service time distribution* | | | | | |
|---|---|---|---|---|---|---|
| | Erlang–10 | | Hyper-2-exp.(I) | | Hyper-2-exp.(II) | |
| $r$ | Approx. | Simul. | Approx. | Simul. | Approx. | Simul. |
| 4 | 1.41e-01 | 1.42e-01 | 4.20e-01 | 4.25e-01 | 2.98e-01 | 3.02e-01 |
| 5 | 4.25e-02 | 4.24e-02 | 2.56e-01 | 2.62e-01 | 1.49e-01 | 1.47e-01 |
| 6 | 1.99e-02 | 2.00e-02 | 1.66e-01 | 1.64e-01 | 8.58e-02 | 8.47e-02 |
| 7 | 1.16e-02 | 1.17e-02 | 1.15e-01 | 1.16e-01 | 5.51e-02 | 5.52e-02 |
| 8 | 7.61e-03 | 7.70e-03 | 8.31e-02 | 8.13e-02 | 3.82e-02 | 3.83e-02 |
| 9 | 5.39e-03 | 5.34e-03 | 6.28e-02 | 6.18e-02 | 2.79e-02 | 2.83e-02 |
| 10 | 4.02e-03 | 4.04e-03 | 4.89e-02 | 4.90e-02 | 2.13e-02 | 2.11e-02 |
| 11 | 3.11e-03 | 3.13e-03 | 3.91e-02 | 3.97e-02 | 1.68e-02 | 1.71e-02 |
| 12 | 2.48e-03 | 2.54e-03 | 3.20e-02 | 3.23e-02 | 1.35e-02 | 1.39e-02 |

TABLE 5. Fraction of overflowing customers based on approximation method II.

There are two sources of inaccuracies in this approximation method, viz. errors from the the Gauss–Seidel algorithm plus errors caused by the state space truncation. The first type of error can be controlled efficiently by the convergence criterion of the Gauss–Seidel algorithm: in our implementation it was set to stop only when a seven-digit accuracy was achieved.

The second source of errors, i.e. those caused by the state space truncation, is much less under control. We can compute the total equilibrium distribution on the border of the state space ($\mathfrak{S}$ in (3)). From trial runs it appeared that we could obtain relative approximation errors in $q$, the fraction of overflowing customers, in the order of 1% by taking $K_C$ and $K_P$ so large that the probability mass of $\mathfrak{S}$ was smaller than 0.001. In the case of exponential and Erlang distributions this accuracy could be accomplished by taking $K_C = K_P = 20$, but for the two examples with hyperexponential distributions we had to resort to large truncations of the state space ($K_C = K_P = 40$) to obtain similar results. The number of iterations for the Gauss–Seidel algorithm varied around 600 for heavily loaded systems to around 200 for $r = 12$.

Besides the fraction of overflowing customers we can compute other performance measures. In Tables 6 and 7 we show the mean waiting time of all high priority customers, i.e. including the overflowing low priority customers. The mean waiting times are, as was to be expected, very sensitive to the type of the service time distribution. The figures for the H-2(II) example are about 15 times larger than those of the Erlang-10 example, regardless of the workload.

| | *Service time distribution* | | | | | |
|---|---|---|---|---|---|---|
| | Exponential | | Erlang–2 | | Erlang–5 | |
| $r$ | Approx. | Simul. | Approx. | Simul. | Approx. | Simul. |
| 4 | 1.08e+00 | 1.08e+00 | 8.00e-01 | 8.00e-01 | 6.35e-01 | 6.34e-01 |
| 5 | 6.40e-01 | 6.38e-01 | 4.75e-01 | 4.77e-01 | 3.78e-01 | 3.79e-01 |
| 6 | 4.23e-01 | 4.19e-01 | 3.15e-01 | 3.16e-01 | 2.51e-01 | 2.51e-01 |
| 7 | 3.00e-01 | 3.01e-01 | 2.24e-01 | 2.24e-01 | 1.78e-01 | 1.78e-01 |
| 8 | 2.24e-01 | 2.23e-01 | 1.67e-01 | 1.67e-01 | 1.33e-01 | 1.32e-01 |
| 9 | 1.73e-01 | 1.74e-01 | 1.29e-01 | 1.29e-01 | 1.03e-01 | 1.03e-01 |
| 10 | 1.38e-01 | 1.38e-01 | 1.03e-01 | 1.03e-01 | 8.23e-02 | 8.19e-02 |
| 11 | 1.13e-01 | 1.13e-01 | 8.42e-02 | 8.42e-02 | 6.72e-02 | 6.74e-02 |
| 12 | 9.38e-02 | 9.41e-02 | 7.00e-02 | 7.02e-02 | 5.59e-02 | 5.59e-02 |

TABLE 6. Mean waiting time over all type $C$ customers

| | *Service time distribution* | | | | | |
|---|---|---|---|---|---|---|
| | Erlang–10 | | Hyper-2-exp.(I) | | Hyper-2-exp.(II) | |
| $r$ | Approx. | Simul. | Approx. | Simul. | Approx. | Simul. |
| 4 | 5.80e-01 | 5.82e-01 | 9.81e+00 | 1.05e+01 | 3.99e+00 | 4.04e+00 |
| 5 | 3.46e-01 | 3.47e-01 | 5.90e+00 | 6.02e+00 | 2.29e+00 | 2.27e+00 |
| 6 | 2.29e-01 | 2.29e-01 | 3.90e+00 | 3.84e+00 | 1.50e+00 | 1.49e+00 |
| 7 | 1.63e-01 | 1.63e-01 | 2.76e+00 | 2.81e+00 | 1.05e+00 | 1.04e+00 |
| 8 | 1.22e-01 | 1.22e-01 | 2.05e+00 | 2.00e+00 | 7.79e-01 | 7.73e-01 |
| 9 | 9.44e-02 | 9.43e-02 | 1.58e+00 | 1.56e+00 | 5.99e-01 | 6.00e-01 |
| 10 | 7.53e-02 | 7.50e-02 | 1.25e+00 | 1.25e+00 | 4.75e-01 | 4.69e-01 |
| 11 | 6.15e-02 | 6.16e-02 | 1.02e+00 | 1.02e+00 | 3.86e-01 | 3.91e-01 |
| 12 | 5.12e-02 | 5.12e-02 | 8.41e-01 | 8.35e-01 | 3.20e-01 | 3.20e-01 |

TABLE 7. Mean waiting time over all type $C$ customers

To conclude this section we present some comparisons between the queueing models with and without impatience. In Tables 8 and 9 we show the server utilization of both type $C$ and $P$. The total server utilization is given in Table 10. As expected we see in the models with impatience that the workload shifts from preventive to corrective maintenance and that this shift becomes more prominent when the coefficient of variation of the service times increases. Note also that the total server utilization decreases in a system with impatience. This is no surprise, since the mean service time for a CM job is smaller than for a PM job. This does not necessarily mean that it is desirable to have a large fraction $q$, since, although this means a relatively small total server utilization, it leads to a high corrective server utilization, and this has a significant impact on availability.

In Tables 11 and 12 the mean waiting times for type $C$ customers are summarized. The figures for the model with impatience are computed by the approximation method. The differences between models with and without impatience are again most prominent in the case of hyperexponential distributions (an increase of 20% for H-2(I)). For exponential and Erlang distributed service times the increase in the waiting times is less than 5%.

| | *queueing system with impatience* | | | | | | *no impatience* |
|---|---|---|---|---|---|---|---|
| *r* | Expon. | Erl.-2 | Erl.-5 | Erl.-10 | H-2 (1) | H-2 (2) | *all dist.* |
| 4 | 3.46e-01 | 3.38e-01 | 3.32e-01 | 3.30e-01 | 4.23e-01 | 3.84e-01 | 2.81e-01 |
| 5 | 2.44e-01 | 2.40e-01 | 2.38e-01 | 2.37e-01 | 2.95e-01 | 2.66e-01 | 2.25e-01 |
| 6 | 1.95e-01 | 1.94e-01 | 1.92e-01 | 1.92e-01 | 2.26e-01 | 2.07e-01 | 1.88e-01 |
| 7 | 1.65e-01 | 1.64e-01 | 1.63e-01 | 1.63e-01 | 1.83e-01 | 1.72e-01 | 1.61e-01 |
| 8 | 1.43e-01 | 1.42e-01 | 1.42e-01 | 1.42e-01 | 1.55e-01 | 1.47e-01 | 1.41e-01 |
| 9 | 1.27e-01 | 1.26e-01 | 1.26e-01 | 1.26e-01 | 1.35e-01 | 1.29e-01 | 1.25e-01 |
| 10 | 1.14e-01 | 1.13e-01 | 1.13e-01 | 1.13e-01 | 1.19e-01 | 1.15e-01 | 1.13e-01 |
| 11 | 1.03e-01 | 1.03e-01 | 1.03e-01 | 1.03e-01 | 1.07e-01 | 1.04e-01 | 1.02e-01 |
| 12 | 9.43e-02 | 9.41e-02 | 9.41e-02 | 9.40e-02 | 9.74e-02 | 9.53e-02 | 9.38e-02 |

TABLE 8. Server utilization of type $C$ for queueing systems with and without impatience

| | *queueing system with impatience* | | | | | | *no impatience* |
|---|---|---|---|---|---|---|---|
| *r* | Expon. | Erl.-2 | Erl.-5 | Erl.-10 | H-2 (1) | H-2 (2) | *all dist.* |
| 4 | 5.58e-01 | 5.75e-01 | 5.86e-01 | 5.90e-01 | 3.99e-01 | 4.82e-01 | 6.88e-01 |
| 5 | 5.12e-01 | 5.20e-01 | 5.25e-01 | 5.27e-01 | 4.09e-01 | 4.68e-01 | 5.50e-01 |
| 6 | 4.43e-01 | 4.46e-01 | 4.48e-01 | 4.49e-01 | 3.82e-01 | 4.19e-01 | 4.58e-01 |
| 7 | 3.85e-01 | 3.87e-01 | 3.88e-01 | 3.88e-01 | 3.48e-01 | 3.71e-01 | 3.93e-01 |
| 8 | 3.39e-01 | 3.40e-01 | 3.41e-01 | 3.41e-01 | 3.15e-01 | 3.31e-01 | 3.44e-01 |
| 9 | 3.03e-01 | 3.03e-01 | 3.04e-01 | 3.04e-01 | 2.86e-01 | 2.97e-01 | 3.06e-01 |
| 10 | 2.73e-01 | 2.74e-01 | 2.74e-01 | 2.74e-01 | 2.62e-01 | 2.69e-01 | 2.75e-01 |
| 11 | 2.49e-01 | 2.49e-01 | 2.49e-01 | 2.49e-01 | 2.40e-01 | 2.46e-01 | 2.50e-01 |
| 12 | 2.28e-01 | 2.28e-01 | 2.29e-01 | 2.29e-01 | 2.22e-01 | 2.26e-01 | 2.29e-01 |

TABLE 9. Server utilization of type $P$ for queueing systems with and without impatience

## 6 Conclusion and remarks

In this paper we have described two approximation methods for the performance analysis of an M/G/1 queue with two customer types of different priority. We have compared the

| | *queueing system with impatience* | | | | | | *no impatience* |
|---|---|---|---|---|---|---|---|
| *r* | Expon. | Erl.-2 | Erl.-5 | Erl.-10 | H-2 (1) | H-2 (2) | *all dist.* |
| 4 | 9.04e-01 | 9.13e-01 | 9.18e-01 | 9.20e-01 | 8.22e-01 | 8.66e-01 | 9.69e-01 |
| 5 | 7.56e-01 | 7.60e-01 | 7.63e-01 | 7.64e-01 | 7.04e-01 | 7.34e-01 | 7.75e-01 |
| 6 | 6.38e-01 | 6.40e-01 | 6.40e-01 | 6.41e-01 | 6.08e-01 | 6.26e-01 | 6.46e-01 |
| 7 | 5.50e-01 | 5.51e-01 | 5.51e-01 | 5.51e-01 | 5.31e-01 | 5.43e-01 | 5.54e-01 |
| 8 | 4.82e-01 | 4.82e-01 | 4.83e-01 | 4.83e-01 | 4.70e-01 | 4.78e-01 | 4.85e-01 |
| 9 | 4.30e-01 | 4.29e-01 | 4.30e-01 | 4.30e-01 | 4.21e-01 | 4.26e-01 | 4.31e-01 |
| 10 | 3.87e-01 | 3.87e-01 | 3.87e-01 | 3.87e-01 | 3.81e-01 | 3.84e-01 | 3.88e-01 |
| 11 | 3.52e-01 | 3.52e-01 | 3.52e-01 | 3.52e-01 | 3.47e-01 | 3.50e-01 | 3.52e-01 |
| 12 | 3.22e-01 | 3.22e-01 | 3.23e-01 | 3.23e-01 | 3.19e-01 | 3.21e-01 | 3.23e-01 |

TABLE 10. Total server utilization for queueing systems with and without impatience

| | *Service time distribution* | | | | | |
|---|---|---|---|---|---|---|
| | Exponential | | Erlang–2 | | Erlang–5 | |
| *r* | Impatience. | No impat. | Impatience. | No impat. | Impatience. | No impat. |
| 4 | 1.08e+00 | 1.02e+00 | 8.00e-01 | 7.73e-01 | 6.35e-01 | 6.22e-01 |
| 5 | 6.40e-01 | 6.24e-01 | 4.75e-01 | 4.68e-01 | 3.78e-01 | 3.74e-01 |
| 6 | 4.23e-01 | 4.13e-01 | 3.15e-01 | 3.10e-01 | 2.51e-01 | 2.48e-01 |
| 7 | 3.00e-01 | 2.94e-01 | 2.24e-01 | 2.21e-01 | 1.78e-01 | 1.76e-01 |
| 8 | 2.24e-01 | 2.20e-01 | 1.67e-01 | 1.65e-01 | 1.33e-01 | 1.32e-01 |
| 9 | 1.73e-01 | 1.71e-01 | 1.29e-01 | 1.28e-01 | 1.03e-01 | 1.02e-01 |
| 10 | 1.38e-01 | 1.36e-01 | 1.03e-01 | 1.02e-01 | 8.23e-02 | 8.17e-02 |
| 11 | 1.13e-01 | 1.11e-01 | 8.42e-02 | 8.35e-02 | 6.72e-02 | 6.68e-02 |
| 12 | 9.38e-02 | 9.26e-02 | 7.00e-02 | 6.95e-02 | 5.59e-02 | 5.56e-02 |

TABLE 11. Mean waiting time of type $C$ in systems with and without impatience

approximation methods with numerical results from simulation. Both methods appear to provide fair to good approximations for $q$, the fraction of overflowing customers. The second method, based on state space truncation, gives better results and provides more performance measures than the first method. The first method, however, has the advantage that the computational efforts are considerably smaller. Both methods were implemented in Pascal on a Sun Sparcstation–1. A typical run of the first method takes less than 1 minute, while the second method for worst case — hyperexponentially distributed service times — may take up to 20 minutes.

A number of suggestions come to mind for future research topics. It seems worthwhile to include the computation of more performance measures in approximation method II. The busy period distribution is an interesting candidate in this matter.

We also want to include the following extension to the model. The preventive maintenance jobs are served by a PM-dedicated FCFS server, but when PM jobs become impatient, they flow over to a queue with an infinite number of servers. This model represents the situation where contractors can be hired externally for corrective maintenance. In this case we are for

| | *Service time distribution* | | | | | |
|---|---|---|---|---|---|---|
| | Erlang–10 | | Hyper-2-exp.(I) | | Hyper-2-exp.(II) | |
| $r$ | Impatience. | No impat. | Impatience. | No impat. | Impatience. | No impat. |
| 4 | 5.80e-01 | 5.71e-01 | 9.81e+00 | 7.34e+00 | 3.99e+00 | 3.19e+00 |
| 5 | 3.46e-01 | 3.43e-01 | 5.90e+00 | 4.85e+00 | 2.29e+00 | 2.01e+00 |
| 6 | 2.29e-01 | 2.27e-01 | 3.90e+00 | 3.35e+00 | 1.50e+00 | 1.34e+00 |
| 7 | 1.63e-01 | 1.62e-01 | 2.76e+00 | 2.42e+00 | 1.05e+00 | 9.55e-01 |
| 8 | 1.22e-01 | 1.21e-01 | 2.05e+00 | 1.82e+00 | 7.79e-01 | 7.14e-01 |
| 9 | 9.44e-02 | 9.38e-02 | 1.58e+00 | 1.41e+00 | 5.99e-01 | 5.54e-01 |
| 10 | 7.53e-02 | 7.49e-02 | 1.25e+00 | 1.13e+00 | 4.75e-01 | 4.43e-01 |
| 11 | 6.15e-02 | 6.12e-02 | 1.02e+00 | 9.23e-01 | 3.86e-01 | 3.62e-01 |
| 12 | 5.12e-02 | 5.10e-02 | 8.41e-01 | 7.68e-01 | 3.20e-01 | 3.01e-01 |

TABLE 12. Mean waiting time of type $C$ in systems with and without impatience

example interested in the influence of the size of the PM repair crew, i.e. the speed of the dedicated server, on the cost of externally hired personnel.

## REFERENCES

[1] F. BACCELLI, P. BOYER, AND G. HEBUTERNE (1984). Single-Server Queues with Impatient Customers. *Adv. Appl. Prob.* 16, 887–905.

[2] F. BACCELLI AND K.S. TRIVEDI (1985). A Single-Server Queue in a Hard-Real-Time Environment. *Oper. Res. Letters* 4, 161–168.

[3] P.P. BHATTACHARYA AND A. EPHREMIDES (1989). Optimal Scheduling with Strict Deadlines. *IEEE Trans. Autom. Control* AC-34, 721–728.

[4] P.P. BHATTACHARYA AND A. EPHREMIDES (1991). Stochastic Monotonicity Properties of Multiserver Queues with Impatient Customers. *J. Appl. Prob.* 28, 673–682.

[5] J.P.C. BLANC, P.R. DE WAAL, P. NAIN, AND D. TOWSLEY (1991). A New Device for the Synthesis Problem of Optimal Control of Admission to an M/M/c Queue. Report BS–R9101, CWI, Amsterdam, to appear in *IEEE Trans. Automat. Control.*

[6] S. CHEN AND D. TOWSLEY (1991). Scheduling Time Constrained Customers in a Non-Removal Real-Time System: A Free Lunch from the Removal System Model. Technical report, Dept. Comp. and Inform. Science, University of Massachusetts, Amherst.

[7] J.W. COHEN (1982). *The Single Server Queue.* North–Holland, Amsterdam.

[8] B.T. DOSHI AND H. HEFFES (1986). Overload Performance of Several Processor Queueing Disciplines for the M/M/1 Queue. *IEEE Trans. Comm.* COM-34, 538–546.

[9] G. FAYOLLE AND M.A. BRUN (1988). On a System with Impatience and Repeated Calls, in *Queueing Theory and its Applications, Liber Amicorum for J.W. Cohen*, 283–305, O.J. Boxma and R. Syski (eds.), North-Holland, Amsterdam.

[10] L.J. Forys (1983). Performance analysis of a new overload strategy, in *Proc. of the 10th International Teletraffic Congress*.

[11] L. Kleinrock (1975). *Queueing Systems, Vol. I: Theory.* Wiley, New York.

[12] C. Palm (1937). Etude des Delais d'Attente. *Ericsson Technics* 5, 37–56, (in French).

[13] S.S. Panwar, D. Towsley, and J.K. Wolf (1988). Optimal Scheduling Policies for a Class of Queues with Customer Deadlines to the Beginning of Service. *J. Assoc. Comput. Mach.* 35, 832–844.

[14] T.L Saaty (1961). *Elements of Queueing Theory with Applications.* Dover Publications.

[15] R.E. Stanford (1979). Reneging Phenomena in Single Channel Queues. *Math. Oper. Res.* 4, 162–178.

[16] R.E. Stanford (1990). On Queues with Impatience. *Adv. Appl. Prob.* 22, 768–769.

[17] D.Y. Sze (1986). A Queueing Model for Overload Analysis, in *IFIP WG 7.3 International Seminar on Computer Networking and Performance Evaluation*, 413–422, T. Hasegawa, H. Takagi, and Y. Takahashi (eds.), North-Holland, Amsterdam.

[18] H.C. Tijms (1986). *Stochastic Modelling and Analysis: A Computational Approach.* Wiley, Chichester.

[19] H.C. Tijms (1988). A Quick and Practical Approximation to the Waiting Time Distribution in the Multi-Server Queue with Priorities, in *Computer Performance and Reliability*, 161–169, G. Iazeolla, P.J. Courtois, and O.J. Boxma (eds.), North-Holland, Amsterdam.

[20] P.R. de Waal (1987). Performance Analysis and Optimal Control of an M/M/1/k Queueing System with Impatient Customers, in *Messung, Modellierung und Bewertung von Rechensystemen, 4. GI/ITG-Fachtagung, Erlangen*, 28–40, U. Herzog and M. Paterok (eds.), Springer-Verlag, Berlin.

[21] P.R. de Waal (1990). *Overload Control of Telephone Exchanges.* Ph.D. thesis, CWI, Amsterdam.