Mathematical Methods of Operations Research (1998) 47:341–353

The M/M/c with Critical Jobs

IVO ADAN

Eindhoven University of Technology, Department of Mathematics and Computing Science, Box 513, 5600 MB Eindhoven, The Netherlands
e-mail: iadan@win.tue.nl

GERARD HOOGHIEMSTRA

Delft University of Technology, Department of Mathematics and Informatics, Box 356, 2600 AJ Delft, The Netherlands
e-mail: g.hooghiemstra@twi.tudelft.nl

Abstract: We consider the $M/M/c$ queue, where customers transfer to a critical state when their queueing (sojourn) time exceeds a random time. Lower and upper bounds for the distribution of the number of critical jobs are derived from two modifications of the original system. The two modified systems can be efficiently solved. Numerical calculations indicate the power of the approach.

Key Words: $M/M/c$, priority queue, bounds, matrix methods

1 Introduction

We consider an $M(\lambda)/M(\mu)/c$ queue, where customers transfer to a critical state when their queueing (sojourn) time exceeds a random time. This time is exponentially distributed with parameter θ. Critical customers have preemptive priority over non-critical ones (hence the servers *never* attend non-critical customers if there are critical customers *waiting* in the queue).

In the application that we have in mind, the customers are repairjobs and the servers are repairmen (engineers). When the queueing time of a job exceeds a random time, the repairjob will be called critical and causes a slowdown of the entire installation from which the repairjobs originate. An example of such an installation is a sugarfactory (sugarhouse), where sugarbeets are refined. The technical staff of such a factory, who maintain the installation, consists of engineers working in full shift during the beetcampaign. This beetcampaign is a period of approximately 100 days during which the beets are harvested from the fields and refined in the factory. The management of the sugarhouse is interested in the delay of the refinery process caused by technical failures of the installation. We model the repairjobs and the engineers as a multi-server queue.

A repairjob becomes critical when its queueing time exceeds a given random time and it is then treated with priority. We arrive at the model described above by assuming that failures arrive according to a Poisson process, the repairwork is done with an exponential rate and jobs become critical with an exponential rate. Of course, such a model can only be used as a first approximation. The basic quantities of interest for the management are the total time during a campaign that the system contains *critical* repairjobs and the average number of *critical* repairjobs.

The system can be represented by a two-dimensional Markov process with states (m,n) where m is the number of non-critical jobs and n the number of critical jobs in the system. It is difficult to find an explicit solution for the stationary probabilities of this Markov process. We will not attempt to do this. Instead lower and upper bounds for the distribution of the number of critical jobs will be derived from two modifications of the original system, which are easier to solve. The number of non-critical jobs in these two systems is bounded by a certain threshold. In the lower bound model this is realized by rejecting a new job if the number of non-critical jobs has reached the threshold and in the upper bound model a new job becomes immediately critical in this case. The larger the threshold, the better the bounds will be, but also the more effort it takes to compute the bounds. Note that when there are many jobs in the original system, most of them will be critical. Hence one might expect that the bounds are tight for already moderate values of the threshold.

The reason why the lower and upper bound system are easier to handle than the original model is that the Markov processes describing these systems have only one unbounded variable, namely n. So they are essentially one-dimensional. In fact, these processes are so-called quasi-birth-death processes, which can be efficiently solved by using Neuts' matrix-geometric approach [10].

The proof of the bounds in based on a Markov reward technique similar to the ones used in [4, 5, 6, 7, 1, 2]. In these references first the Markov processes representing the original model and the lower and upper bound model are translated into equivalent Markov chains. Then it is shown by induction that for each finite number of periods the performance of the original model is sandwiched between the performances of the two bound models. Letting the number of periods tend to infinity yields the result for the average performance. The translation into a Markov chain is only possible if the transition rates are bounded. In our case, this holds for the lower and upper bound model, but *not* for the original model. Therefore we have to follow a slightly different road. First we prove that the number of critical jobs in the lower (upper) bound model stochastically increases (decreases) as the threshold increases. This is established by using the technique described above. Then the proof is finished by showing that the distributions (and also the means) of the number of critical jobs in the lower and upper bound models converge to that of the original model as the threshold tends to infinity. In fact, we prove more than in the references mentioned, in the sense that not only the bounds are proved, but also that they converge to each other.

The Markov reward technique used in the present paper to establish computable bounds is also a powerful tool to prove qualitative properties, like e.g. monotonicity properties in queueing networks (cf. [13]) or optimality of routing policies to parallel queues (cf. [8]).

The model with an additional input stream of jobs which are critical from the beginning has been studied by De Waal [12] and Van Rooij [11]. They use this model to describe corrective and preventive maintenance of components in an installation like, e.g., a plant at an oil refinery. In their model corrective maintenance jobs (i.e., the critical jobs) have priority over preventive maintenance jobs. But preventive maintenance of a component can change into corrective maintenance, namely when that component breaks down while it is waiting. They develop approximations for the fraction of preventive maintenance jobs that become corrective and the mean waiting time of corrective maintenance jobs.

The paper is organized as follows. In Section 2 we describe the models. The bounds are established in Section 3 and the matrix-geometric analysis of the lower and upper bound model is briefly described in Section 4. We present numerical results in Section 5. The final section is devoted to conclusions and comments.

2 The Models

We consider an $M(\lambda)/M(\mu)/c$ queue, where jobs become critical when their queueing time exceeds an exponential time with mean $1/\theta$. Critical jobs have preemptive priority over non-critical jobs. In the lower and upper bound model the arrival mechanism is modified such that the number of non-critical jobs never exceeds a (fixed) threshold T. If, due to an arrival, the number of non-critical jobs would exceed T, then in the lower bound model that job is rejected and in the upper bound model that job becomes instantaneously critical.

The three models are Markov processes. The state of the original system can be described by the pair (m,n), where m is the number of non-critical jobs in the system and n the number of critical jobs. From state (m,n) there are transitions to $(m+1,n)$ with rate λ (an arrival) and $(m-1,n+1)$ with rate $m\theta$ (transfer to critical state). There are two other transitions corresponding to service completions, namely to $(m,n-1)$ with rate $\min(n,c)\mu$ (departure of a critical job) and, if $n<c$, then also a transition to $(m-1,n)$ is possible with rate $\min(c-n,m)\mu$ (departure of a non-critical job). The states in the lower and upper bound systems are restricted to the pairs (m,n) with $m \leq T$. The transitions are the same as in the original system, except that in the states (T,n) the transitions to $(T+1,n)$ with rate λ are replaced by transition (with the same rate to (T,n) and $(T,n+1)$ in the lower and upper bound system, respectively.

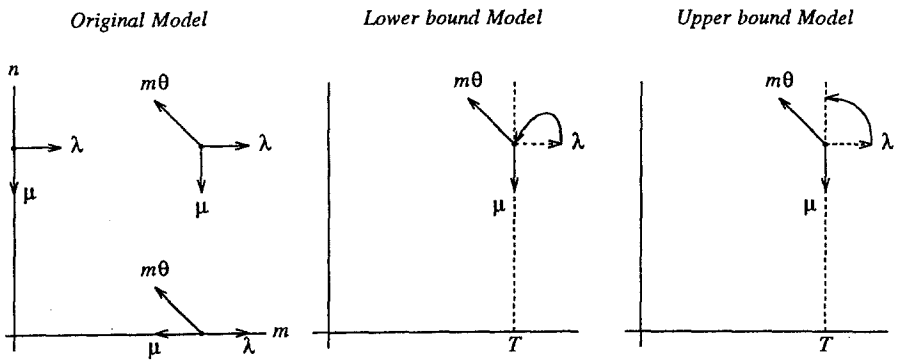

Fig. 1. Transition rates for the three models with $c = 1$

Since in the original and upper bound system the total number of jobs is the same as in the ordinary $M/M/c$, the condition $\lambda < c\mu$ is necessary and sufficient for these systems to be ergodic. The lower bound system destroys work, so it is ergodic if the original system is ergodic. The transition rates for the three models with $c = 1$ are depicted in Fig. 1. For the lower and upper bound model we only indicate the differences with respect to the original model.

In the next section we will prove that the number of critical jobs in the upper bound model is stochastically larger than in the original model. The proof for the lower bound model proceeds along the same lines, and it is therefore omitted.

3 Proof of the Upper Bounds

We first compare the upper bound models with thresholds T and $T+1$. Let $\underline{L}^u_T$ be a random variable denoting the (stationary) number of critical jobs in the model with threshold T. Then we will prove the following result.

Theorem 3.1: $\underline{L}^u_T \geq_{st} \underline{L}^u_{T+1}$ *for each* $T \geq 0$.

Let us fix $T \geq 0$ and $N \geq 0$. We now have to show that

$$P(\underline{L}^u_T \geq N) \geq P(\underline{L}^u_{T+1} \geq N)\,. \tag{1}$$

Let Q_T be the generator of the model with threshold T. Its equilibrium distri-

bution π_T satisfies $\pi_T Q_T = 0$. Related to this Markov process we introduce the Markov chain with transition matrix $I + \Delta Q_T$ where $\Delta > 0$, but sufficiently small for $I + \Delta Q_T$ to be nonnegative. Clearly the Markov chain has the same equilibrium distribution π_T as the Markov process. Hence, to establish (1) we can focus on the Markov chains $I + \Delta Q_T$ and $I + \Delta Q_{T+1}$ with $\Delta = 1/(\lambda + c\mu + (T+1)\theta)$. Along with these chains we introduce the one-step cost $c(m,n)$ defined as 1 if $n \geq N$ and 0 otherwise. Define $v_k(m,n)$ and $w_k(m,n)$ as the total expected cost over k periods for the models with thresholds T and $T+1$, respectively, and with (m,n) as initial state. Further we set $v_0 = w_0 = 0$. In the Appendix we prove by induction the following intuitively obvious inequalities for the functions w_k.

Lemma 3.2: For all $k \geq 0$ we have

(i) $w_k(m,n+1) \geq w_k(m,n), \quad 0 \leq m \leq T+1, n \geq 0;$
(ii) $w_k(m+1,n) \geq w_k(m,n), \quad 0 \leq m \leq T, n \geq 0;$
(iii) $w_k(m,n+1) \geq w_k(m+1,n), \quad 0 \leq m \leq T, n \geq 0.$

The inequalities (i) and (ii) state that it is preferable to start with less jobs in the system and (iii) states that it is attractive to change a critical job into a non-critical job. Note that the cost function also satisfies these inequalities. Lemma 3.2 is crucial for the proof of the following result (see the Appendix).

Lemma 3.3: For all $k \geq 0$ and all (m,n) with $0 \leq m \leq T$ and $n \geq 0$,

$$v_k(m,n) \geq w_k(m,n) \,. \tag{2}$$

From Lemma 3.3 we conclude that

$$P(\underline{L}_T^u \geq N) = \lim_{k\to\infty} \frac{v_k(m,n)}{k} \geq \lim_{k\to\infty} \frac{w_k(m,n)}{k} = P(\underline{L}_{T+1}^u \geq N) \,,$$

and so the proof of Theorem 3.1 is complete.

Next we show that the equilibrium distribution π_T of the upper bound model with threshold T converges weakly to the equilibrium distribution π of the original model as T tends to infinity.

Theorem 3.4: $\pi_T \xrightarrow{d} \pi$ as $T \to \infty$.

Proof. Set

$$\pi_T(k) = \sum_{m+n=k} \pi_T(m,n) .$$

Balancing the flow into and out the set of states (m,n) with $m+n \leq k$ yields that $\min(k+1,c)\mu\pi_T(k+1) = \lambda\pi_T(k)$ (note that for the lower bound model we have $\leq$ instead of $=$). Hence, it is immediate that for $\lambda < c\mu$ and independent of T the probabilities $\pi_T(k)$ decrease exponentially. It follows that the class of discrete probability measures $\{\pi_T, T = 0,1,\ldots\}$ on $S = \{(m,n) \in \mathbf{N}^2\}$ is tight. Consequently, by Prohorovs theorem (cf. [3], Theorem 6.1) the class $\{\pi_T\}$ is relatively compact, meaning that each subsequence $\pi_{T_n'}$ contains a further subsequence $\pi_{T_{n''}}$ that converges weakly to some discrete probability measure $\tilde{\pi}$ on S. The limit probability measure $\tilde{\pi}$ must satisfy the equilibrium equations of the original model and hence is equal to π. So each converging subsequence $\pi_{T_{n''}}$ has limit π and this implies the statement of the theorem. □

Let $\underline{L}$ be the number of critical jobs in the orginal system. Theorems 3.4 and 3.1 imply that the distribution of $\underline{L}_T^u$ converges to that of $\underline{L}$, monotonously.

Corollary 3.5: $P(\underline{L}_T^u \geq N) \downarrow P(\underline{L} \geq N)$ *as* $T \to \infty$ *for each* $N \geq 0$.

Denote the means of $\underline{L}_T^u$ and $\underline{L}$ by L_T^u and L, respectively. From the monotone convergence theorem we can conclude that the means also converge.

Corollary 3.6. $L_T^u \downarrow L$ *as* $T \to \infty$.

Similar results hold for $\underline{L}_T^l$, the number of critical jobs in the lower bound model (with, of course, obvious modifications such as $\uparrow$ instead of $\downarrow$ in the two corollaries formulated above).

4 Analysis of the Upper Bound Model

In this section we briefly describe the analysis of the upper bound model, which is based on the matrix geometric theory developed by Neuts [10]. The analysis of the lower bound model is identical (and therefore not included).

The upper bound model can be described by an irreducible Markov process with states (m,n), where m is the number of non-critical jobs in the system

and n the number of critical ones. The state space is restricted to the pairs with $m \leq T$. Let us define for $n = 0, 1, \ldots$, *level* n as the set of states $(0,n)$, $(1,n), \ldots, (T,n)$. Then we partition the state space into the levels $c, c+1, \ldots$ and we put together the levels $0, 1, \ldots, c-1$ with less regular transition behaviour in one set of boundary states. The states at a level are ordered lexicographically. For this partitioning the generator Q_T is of the form

$$Q_T = \begin{pmatrix} B_{00} & B_{01} & 0 & 0 & 0 & \cdots \\ B_{10} & A_1 & A_0 & 0 & 0 & \cdots \\ 0 & A_2 & A_1 & A_0 & 0 & \cdots \\ 0 & 0 & A_2 & A_1 & A_0 & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix}.$$

The blocks A_0, A_1 and A_2 are square matrices of order $T+1$. The matrices B_{00}, B_{01} and B_{10} are of dimension $c(T+1) \times c(T+1)$, $c(T+1) \times (T+1)$ and $(T+1) \times c(T+1)$, respectively.

For $\lambda < c\mu$ the system is ergodic. Then the equilibrium probability vector π_T exists. We partition π_T into the (large) vector $\pi_T^b = (\pi_T^0, \ldots, \pi_T^{c-1})$ of boundary states and into the sequence $\pi_T^c, \pi_T^{c+1}, \ldots$, where π_T^n is the equilibrium probability vector of level n. Note that the generator $A_0 + A_1 + A_2$ is irreducible, so we can conclude from (the continuous time version of) Theorem 1.5.1 in [10] that

$$\pi_T^n = \pi_T^c R^{n-c}, \quad n \geq c, \tag{3}$$

where the matrix R (the so-called *rate-matrix*) is characterized as the minimal nonnegative solution of the matrix quadratic equation

$$A_0 + RA_1 + R^2 A_2 = 0.$$

The vectors π_T^b and π_T^c follow from the boundary conditions

$$(\pi_T^b, \pi_T^c) \begin{pmatrix} B_{00} & B_{01} \\ B_{10} & A_1 + RA_2 \end{pmatrix} = 0$$

and the normalization equation

$$\sum_{n=0}^{c-1} \pi_T^n \mathbf{e} + \pi_T^c (I - R)^{-1} \mathbf{e} = 1,$$

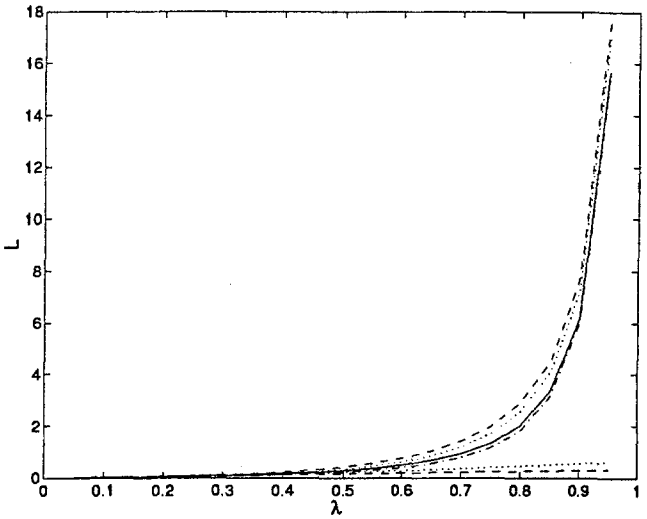

Fig. 2. Bounds for L as a function of λ for $\mu = c = 1$ and $\theta = 0.25$ with T varied as 2 and 3

where $\mathbf{e}$ is the column vector of length $T+1$ with all entries equal to one. For the determination of R efficient algorithms have been developed (see, e.g., [9]). From the equilibrium distribution π_T it is now straightforward to determine global performance characteristics. For instance, for L_T^u we find, by inserting (3), that

$$L_T^u = \sum_{n=0}^{\infty} n\pi_T^n \mathbf{e} = \sum_{n=0}^{c-1} n\pi_T^n \mathbf{e} + \pi_T^c (I-R)^{-2}\mathbf{e} + (c-1)\pi_T^c (I-R)^{-1}\mathbf{e}\,.$$

5 Numerical Results

This section is devoted to numerical results. In Fig. 2 we demonstrate the rate of convergence of the bounds for the mean number of critical jobs as a function of λ for the case $\mu = c = 1$ and $\theta = 0.25$. The dashed lines are the bounds for $T = 2$; the dotted ones for $T = 3$. The results show that especially the upper bound rapidly converges to the exact values, and that for high workload this bound is much better than the lower bound. Apparently, transforming a non-critical job somewhat earlier to a critical state has not much effect on L, but the impact of destroying work is considerable. Luckily, it is possible to produce with the upper bound model a (much better) lower bound for L as well. This will be explained below.

Table 1. Performance characteristics. In all examples we have set $c\mu = 1$

c	λ	θ	L	$P(\underline{L} = 0)$	f_c	T
1	.5	.1	.11353	.911353	.177295	16
		.5	.47101	.735504	.528992	9
	.9	.1	3.7234	.472336	.586293	29
		.5	7.4388	.219385	.867350	13
3	.5	.1	.40586	.682451	.266197	15
		.5	1.0792	.384542	.657668	9
	.9	.1	4.5008	.253397	.616969	28
		.5	8.4454	.074820	.893420	13

The number of jobs that becomes critical per unit of time should balance the number of critical jobs that leaves the system per unit of time. Hence

$$(L_{M/M/c} - L)\theta = \lambda f_c \,, \tag{4}$$

where $L_{M/M/c}$ is the mean number of jobs in an $M/M/c$ system and f_c is the fraction of jobs that becomes critical. Note that f_c satisfies

$$f_c = \frac{1}{\lambda} \sum_{n=1}^{c} P(\underline{L} \geq n)\mu \,,$$

where the sum at the right hand side is the number of critical jobs leaving the system per unit of time. We can use relation (4) to produce a lower (upper) bound for L from the upper (lower) bound for f_c. In other words, with one bound model we are able to produce a lower bound as well as an upper bound for L. In Fig. 2 we show the lower bound for L (the dash-dotted line) obtained from the upper bound for f_c which is produced by the upper bound model with $T = 3$. Clearly, this bound for L is much better than the one obtained from the lower bound model with $T = 3$.

In Table 1 we list for several values of c, λ and θ the mean number of critical jobs L, the probability $P(\underline{L} = 0)$ and the fraction f_c of jobs that becomes critical. In all examples in Table 1 we have set $\mu = 1/c$, so that λ is equal to the occupation rate of the servers. The value of T indicates the minimal threshold needed to obtain the performance measures with the accuracy (i.e., the number of digits) listed. The computation of L is based on the upper bound model only.

We see in Table 1 that the performance characteristics can be determined accurately for already moderate values of T. For each example the computation time on an ordinary 486 PC is at most a couple of seconds.

We conclude this section by comparing the mean number of critical jobs in an $M/M/1$ system where jobs become critical after an exponential time with

Table 2. Comparison of the mean number of critical jobs for an exponential and deterministic deadline

λ	θ	L_e	L_d
.5	.1	.1135	.0067
	.5	.4710	.3679
	1	.6559	.6065
.9	.1	3.723	3.311
	.5	7.439	7.369
	1	8.168	8.144

mean $1/\theta$ with a system where jobs become critical after a deterministic time $1/\theta$. Let us denote the mean number of critical jobs in the systems with an exponential and deterministic deadline by L_e and L_d, respectively. By Little's law we have that

$$L_d = \lambda C ,$$

where C is the expected time that a job is critical. Since the queueing time of a job is exponentially distributed with parameter $\mu - \lambda$, it follows that

$$C = e^{-(\mu-\lambda)/\theta} \frac{1}{\mu - \lambda} .$$

Hence,

$$L_d = e^{-(\mu-\lambda)/\theta} \frac{\lambda}{\mu - \lambda} .$$

In Table 2 we compare L_e and L_d for several values of λ and θ. The results show that the mean number of critical jobs is fairly insensitive to the distribution of the critical deadline, except when λ and θ are both small.

Note: If θ is very small ($\theta < \mu/10$), it seems sensible to bound the number of critical jobs instead of the non-critical ones. This can be realized by refusing the transfer of a job to a critical state, if due to that transfer the number of critical jobs would exceed a threshold T. Further, we have to bound the rate with which jobs become critical, the maximum rate is $M\theta$ say. It can be proved (along the same lines as in Section 3) that this model produces a lower bound for the distribution of critical jobs.

6 Conclusions

We have seen that it is possible to derive tight lower and upper bounds for the distribution of the number of critical jobs by comparing the original system with two modified systems. The lower and upper bound system are much easier to analyze than the original one, because they have a matrix-geometric solution.

It is straightforward to extend the analysis to the case where there is also an input stream of jobs which are critical from the beginning. As mentioned in the introduction, this model is considered in [12, 11]. Further it is also possible to derive bounds for the performance of a system with phase-type deadlines and/or service times. In this case, however, the state space is much larger than for the exponential system, since we have to include extra information of the status of the jobs and the service process in the state description.

Appendix

Proof of Lemma 3.2. The proof proceeds by induction. Since $w_0 = 0$ the inequalities are trivially satisfied for $k = 0$. Suppose that (i)–(iii) hold for k. Then we will establish them for $k+1$. The induction step is only worked out for (i), the other two inequalities can be treated similarly.

Case a: $m < T+1, n \geq c$. We have

$$\begin{aligned} w_{k+1}(m,n+1) &= c(m,n+1) + \Delta\lambda w_k(m+1,n+1) + \Delta m\theta w_k(m-1,n+2) \\ &\quad + \Delta c\mu w_k(m,n) + (1-\Delta(\lambda + c\mu + m\theta))w_k(m,n+1)\,, \\ w_{k+1}(m,n) &= c(m,n) + \Delta\lambda w_k(m+1,n) + \Delta m\theta w_k(m-1,n+1) \\ &\quad + \Delta c\mu w_k(m,n-1) + (1-\Delta(\lambda + c\mu + m\theta))w_k(m,n)\,. \end{aligned}$$

Comparing the right sides of the equations above we see that (i) for $k+1$ follows from the induction hypothesis (i).

Case b: $m < T+1, n < c, m \geq c-n$. Then

$$\begin{aligned} w_{k+1}(m,n+1) &= c(m,n+1) + \Delta\lambda w_k(m+1,n+1) + \Delta m\theta w_k(m-1,n+2) \\ &\quad + \Delta n\mu w_k(m,n) + \Delta\mu w_k(m,n) \\ &\quad + \Delta(c-n-1)\mu w_k(m-1,n+1) \\ &\quad + (1-\Delta(\lambda + c\mu + m\theta))w_k(m,n+1)\,, \end{aligned}$$

$$w_{k+1}(m,n) = c(m,n) + \Delta\lambda w_k(m+1,n) + \Delta m\theta w_k(m-1,n+1)$$
$$+ \Delta n\mu w_k(m,n-1) + \Delta\mu w_k(m-1,n)$$
$$+ \Delta(c-n-1)\mu w_k(m-1,n)$$
$$+ (1 - \Delta(\lambda + c\mu + m\theta))w_k(m,n) .$$

So from the induction assumptions (i)–(ii) we obtain $w_{k+1}(m,n+1) \geq w_{k+1}(m,n)$.

Case c: $m < T+1, n < c, m < c-n$. From

$$w_{k+1}(m,n+1) = c(m,n+1) + \Delta\lambda w_k(m+1,n+1) + \Delta m\theta w_k(m-1,n+2)$$
$$+ \Delta m\mu w_k(m-1,n+1) + \Delta n\mu w_k(m,n) + \Delta\mu w_k(m,n)$$
$$+ (1 - \Delta(\lambda + (m+n+1)\mu + m\theta))w_k(m,n+1) ,$$

$$w_{k+1}(m,n) = c(m,n) + \Delta\lambda w_k(m+1,n) + \Delta m\theta w_k(m-1,n+1)$$
$$+ \Delta m\mu w_k(m-1,n) + \Delta n\mu w_k(m,n-1) + \Delta\mu w_k(m,n)$$
$$+ (1 - \Delta(\lambda + (m+n+1)\mu + m\theta))w_k(m,n) ,$$

we get that $w_{k+1}(m,n+1) \geq w_{k+1}(m,n)$.

The proof of the three cases above with $m = T+1$ only needs obvious changes and it is therefore omitted.

Proof of Lemma 3.3. The proof is again by induction. For $k = 0$ inequality (2) is trivial. Suppose that (2) holds for k. In order to prove (2) for $k+1$ we have to distinguish between the cases $m < T$ and $m = T$, but the latter is the only interesting situation. For $m = T$ and $n \geq c$ we have (recall that v_k has threshold T),

$$v_{k+1}(T,n) = c(T,n) + \Delta\lambda v_k(T,n+1) + \Delta T\theta v_k(T-1,n+1)$$
$$+ \Delta c\mu v_k(T,n-1) + (1 - \Delta(\lambda + c\mu + T\theta))v_k(T,n)$$
$$w_{k+1}(T,n) = c(T,n) + \Delta\lambda w_k(T+1,n) + \Delta T\theta w_k(T-1,n+1)$$
$$+ \Delta c\mu w_k(T,n-1) + (1 - \Delta(\lambda + c\mu + T\theta))w_k(T,n) .$$

Hence

$$v_{k+1}(T,n) - w_{k+1}(T,n) \geq \Delta\lambda(v_k(T,n+1) - w_k(T+1,n))$$
$$\geq \Delta\lambda(v_k(T,n+1) - w_k(T,n+1)) \geq 0\,,$$

where the first and third inequality follow by induction; the second one follows from Lemma 3.2(iii). The case $m = T$ and $n < c$ follows in the same way.

References

[1] Adan IJBF, Van Houtum GJ, Van der Wal J (1994) Upper and lower bounds for the waiting time in the symmetric shortest queue system. Annals of Operat. Research 48:197–217

[2] ——— (1997) The symmetric longest queue system. Stochastic Models 13:105–120

[3] Billingsley P (1968) Convergence of probability measures. John Wiley & Sons, Chichester

[4] Van Dijk NM (1988) Simple bounds for queueing systems with breakdowns. Perf. Evaluation 8:117–128

[5] ——— (1988) A formal proof for the insensitivity of simple bounds for finite multi-server non-exponential tandem queues. Stochastic Processes 27:261–277

[6] Van Dijk NM, Lamond BF (1988) Simple bounds for finite single-server exponential tandem queues. Opns. Res. 36:470–477

[7] Van Dijk NM, Van der Wal J (1989) Simple bounds and monotonicity results for finite multi-server exponential tandem queues. QUESTA 4:1–16

[8] Hordijk A, Koole G (1992) On the assignment of customers to parallel queues. PEIS 6:495–511

[9] Latouche G, Ramaswami V (1993) A logarithmic reduction algorithm for quasi-birth-death processes. J. Appl. Prob. 30:650–674

[10] Neuts MF (1981) Matrix-geometric solutions in stochastic models. Johns Hopkins University Press, Baltimore

[11] Van Rooij MCJ (1995) Quantitative models for maintenance optimization. Master's Thesis, Tilburg University

[12] De Waal PR (1992) Approximate analysis of an M/G/1 priority queue with priority changes due to impatience. CWI Report BS-R9202

[13] Van der Wal J (1989) Monotonicity of the throughput of a closed exponential queueing network in the number of jobs. OR Spektrum 11:97–100

Received: April 1996
Revised version received: May 1997