The symmetric longest queue system

EINDHOVEN UNIVERSITY OF TECHNOLOGY
Department of Mathematics and Computing Science

Memorandum COSOR 94-33

The Symmetric Longest Queue System

I.J.B.F. Adan
G.J. van Houtum
J. van der Wal

Eindhoven, October 1994
The Netherlands

The Symmetric Longest Queue System

Ivo Adan, Geert-Jan van Houtum and Jan van der Wal
Eindhoven University of Technology, Eindhoven, The Netherlands

In this paper we derive the performance of the exponential symmetric longest queue system from two variants: a longest queue system with Threshold Rejection of jobs and one with Threshold Addition of jobs. It is shown that these two systems provide lower and upper bounds for the performance of the longest queue system. Both variants can be analyzed efficiently. Numerical experiments demonstrate the power of the approach.

1. INTRODUCTION

In this paper we study the symmetric longest queue system, which is characterized as follows. Consider a system with N types of jobs. Each type has its own queue. In each queue jobs arrive according to a Poisson stream with rate λ. The jobs are served by one server. The service times are exponentially distributed with mean $1/\mu$. If the server has completed a job, he picks the next job from the longest queue.

We encountered this model when studying the following problem. A company selling copiers has service contracts with its clients. If a system breaks down it has to be repaired within 24 hours. Repair usually means replacing one or more parts by spare parts. Often the defected parts can be repaired, think of printed circuit boards. The company has to decide how many spares are needed. Having too many spare parts leads to extra costs as these parts are never used and will become obsolete. On the other hand, having too few spare parts leads to the situation that too often a repair will take more than 24 hours. If there are many different types of spares that have to be repaired and the repair capacity is limited then it has to be decided which items have to be repaired first. A sensible strategy will be to choose those items which are likely to be needed first. Other examples of the problem described above are found in the repair of medical systems or airplane engines.

Simplifying the problem described above we arrive at the following model. Since the number of copiers is very large (thousands), it is reasonable to assume that defects occur according to homogeneous Poisson processes, one per part-type. Each type has its own queue. All defected parts of similar types are served by a single repairman. If he has completed a job, he has to choose the next job from one of the (his) queues. In this paper we will concentrate on a simplified version of the problem in which all Poisson processes have the same arrival rate and the repair times for all job types are exponentially distributed with the same mean. Because of the symmetry we also assume that the number of spares of each type is the same. Then a natural

Correspondence to: G.J. van Houtum, Eindhoven University of Technology, Department of Mathematics and Computing Science, P.O. Box 513, 5600 MB Eindhoven, The Netherlands.

repair strategy for the repairman is to always select his next job from the longest queue, since the number of good spares of that part type is the smallest. An important performance measure is the fraction of the time that all spares are in repair so no good spares are available. If during that time a failure in a copier occurs, repair of the copier within 24 hours will be unlikely. So what we want to obtain is the probability that there are more than a specified number of spares in repair. We will show that for the longest queue model this probability can be efficiently obtained.

Note that a realistic model will be far more complicated. The arrival rates and hence the number of spares per type will differ and the jobs will not be exponential and have distributions that differ per job type. In such a case the longest queue strategy may not be sensible. Also it might be more efficient to repair in batches.

We will approximate the longest queue model by two other models that are easier to handle and that provide upper and lower bounds for the queue length distribution. These two models exploit the aspect that in the longest queue system high imbalance in the queue lengths is not likely to occur. In the lower and upper bound model the difference in length between the longest and the shortest queue is limited to a prespecified threshold. In the lower bound model this is realized by rejecting an arrival in the longest queue if due to this arrival the difference between longest and shortest queue will exceed the threshold. It seems obvious that this rejection mechanism leads to shorter queues. If in the upper bound model due to an arrival in the longest queue the threshold will be exceeded, a job is added to the (all) shortest queue(s) as well. These extra arrivals clearly lead to longer queues. The larger the threshold, the less jobs will be rejected or added and so the bounds will be better. Since the server acts in a manner trying to balance the queue lengths, one might expect that the bounds will be tight for already moderate values of the threshold.

We will prove that these two model indeed produce upper and lower bounds for the queue length distribution of the longest queue system. The technique used in the proofs is similar to the ones used by Van der Wal [8], Van Dijk and Van der Wal [4], Van Dijk and Lamond [3] and Adan et al. [1]. First the Markov processes representing the three models are translated into equivalent Markov chains. Then we show by induction that for each finite number of periods the performance of one model is better than the performance of another one. Letting the number of periods tend to infinity yields the desired result for the average performance. The proofs are presented for the case of two queues only. The proofs for the case of more than two queues are essentially the same but notationally more complex and therefore omitted.

The longest queue model with two queues has been studied by Zheng and Zipkin [9] who assume that the longest queue policy is applied preemptively. The same model has also been analyzed by Flatto [5] using generating functions. Cohen [2] presents a generating function analysis of the longest queue model with two queues and generally distributed service times. To our knowledge, no results for the longest queue model with more than two queues are available in the literature.

This paper is organized as follows. In section 2 we describe the models and translate the continuous-time Markov processes to discrete-time Markov chains. Section 3 explains the technique used to prove that the performance of one model is better than that of another one. Monotonicity properties of the longest queue model are established in section 4. In the sections 5 and 6 it is proved that the threshold models indeed give lower respectively upper bounds for the longest queue model. In section 7 it is shown how the queue length distributions of the threshold models can be found. Numerical results are presented in section 8.

2. THE MODELS

In the longest queue model we consider N types of jobs. Each type has its own queue. The queues are numbered 1, ..., N. In each queue jobs arrive according to a Poisson stream with rate λ. The jobs are served by one server. The service times are exponentially distributed with mean $1/\mu$. If the server has completed a job he picks the next job from the longest queue. In the upper and lower bound model the arrival mechanism is modified to accomplish that the difference between the longest and the shortest queue is limited to a prespecified threshold L (≥ 1). If due to an arrival in the longest queue the difference between the longest and the shortest queue would exceed L, then in the upper bound model that job is rejected and in the lower bound model a job is added to all shortest queues as well.

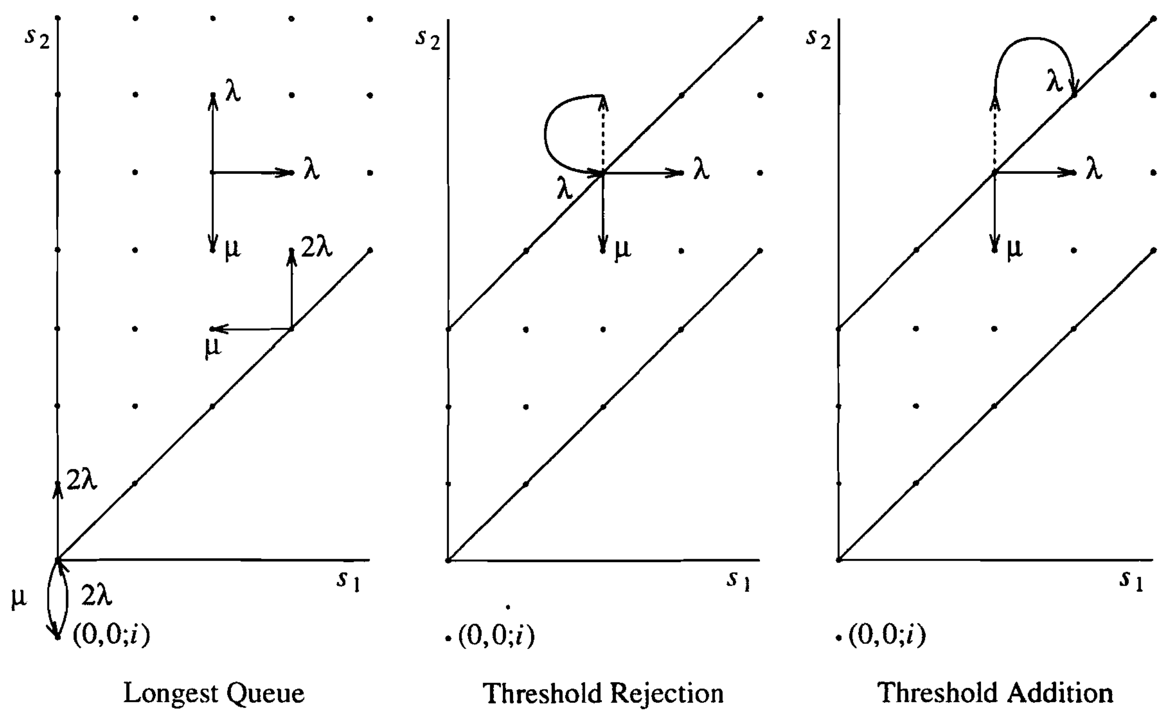
The state of the original longest queue system will be described by an ordered N-tuple of the queue lengths $s = (s_1, ..., s_N)$ with $s_1 \leq \cdots \leq s_N$. So s_1 is the length of the shortest queue, s_2 the length of the second shortest, and so on. Because of the symmetry we are not interested in the length of a specific queue. If a job is taken into service, then it is removed from its queue immediately. So in the queues there are waiting jobs only. If all queues are empty there are two possible states $(0, ..., 0; i)$ and $(0, ..., 0; b)$ where i stands for an idle server and b for a busy server. The state $(0, ..., 0; b)$ will be abbreviated as $(0, ..., 0)$. Note that this state description does not show which type of job is in service. Thinking of the repair of defected parts in copiers, then no essential information is lost if we assume that a part that is in repair will still be in time to replace a defected part. Of course, it is easy to include this type number in the state description, but it would seriously increase the number of states.

The state set in each of the two bound models is restricted to tuples s with $s_N - s_1 \leq L$.

The three models are Markov processes. In all states the output rate is less or equal to $N\lambda + \mu$. Without loss of generality we may take $N\lambda + \mu = 1$. The original longest queue model is ergodic if $N\lambda < \mu$. This condition is sufficient (but not necessary) for the lower bound model to be ergodic as well. For the upper bound model the ergodicity condition will be stronger (see section 8). In Figure 1 the transition-rate diagrams for the three models are depicted for $N = 2$ and $L = 3$. For the threshold models we only depicted the differences with respect to the original model. In the Threshold Rejection model the arrival arc with rate λ from state $(s_1, s_1{+}3)$ is redirected to the state itself and in the Threshold Addition model it is redirected to $(s_1{+}1, s_1{+}4)$.

Let Q be the generator of one of the three Markov processes we are dealing with. Then the corresponding equilibrium distribution p satisfies $pQ = 0$. Instead of studying the Markov process with generator Q we will consider the Markov chain with transition matrix $P = I + Q$. As $N\lambda + \mu = 1$ the matrix P is stochastic. Clearly the equilibrium distributions of the Markov chain and the Markov process are equal. Also mean costs per unit of time are easy to compare. If $c(s)$

Figure 1: The transition rates for the three models for $N = 2$ and $L = 3$.



is the cost rate in state s for the Markov process and we take $c(s)$ as the costs per period in the Markov chain, then the Markov process and the Markov chain will both have the same average costs per unit of time $\sum_s p(s)c(s)$. From now on we only consider the three Markov chains.

3. THE TECHNIQUE

In this section we will explain the technique used to prove that the performance of one model is better than that of another one. If we think of the repair of defected parts in copiers, then we are interested in avoiding the situation that for a certain type all spares are in repair. Therefore we take the cost function $c(s)$ equal to the number of queues with length equal to or longer than M. We can think of M as the maximum number of available spares. To prove that the lower and the upper bound model indeed give lower and upper bounds for the average costs we study the expected costs over a finite number of periods. Define $v_n(s)$ as the expected costs over n periods for the original longest queue model when starting in s. Similarly let u_n and w_n be the expected n-period costs for the lower and upper bound model respectively. Defining $u_0 = v_0 = w_0 = 0$ we will prove by induction that for all n and all s

$$u_n(s) \leq v_n(s) \leq w_n(s)\,. \tag{1}$$

Then it follows that the average costs are ordered in the same way.

To prove (1) we will first establish some obvious monotonicity results for the functions v_n. To keep notations simple we only consider the case $N = 2$. One easily sees that the results hold for more than 2 queues as well, the notations however become more complex in that case.

4. MONOTONICITY RESULTS FOR THE FUNCTIONS v_n

The monotonicity results that we need are the following intuitively obvious inequalities:

Lemma.

For all $n \geq 0$ we have

$$v_n(k,l+1) \geq v_n(k,l)\,,\ \ 0 \leq k \leq l \tag{2}$$

$$v_n(k+1,l) \geq v_n(k,l)\,,\ \ 0 \leq k < l \tag{3}$$

$$v_n(0,0) \geq v_n(0,0\,;i)\,. \tag{4}$$

These three inequalities state that it is preferable to start with smaller queues. Note that the cost function $c(s)$ (the number of queues with length at least M) also satisfies these inequalities, i.e.

$$c(k,l+1) \geq c(k,l)\,,\ \ 0 \leq k \leq l \tag{5}$$

$$c(k+1,l) \geq c(k,l)\,,\ \ 0 \leq k < l \tag{6}$$

$$c(0,0) \geq c(0,0\,;i)\,. \tag{7}$$

It will be clear from the proof of the lemma that the inequalities (2)-(4) are valid for any cost function $c(s)$ satisfying the properties (5)-(7).

Proof of the lemma.

The proof will be given by induction. Since $v_0 = 0$ the inequalities trivially hold for $n = 0$. Assuming (2)-(4) to hold for n we will establish them for $n+1$.

Proof of (2):

We have to distinguish between the three cases $0 \leq k < l$, $k = l > 0$ and $k = l = 0$.

Case a: $0 \leq k < l$.

We have

$$v_{n+1}(k,l+1) = c(k,l+1) + \lambda v_n(k,l+2) + \lambda v_n(k+1,l+1) + \mu v_n(k,l)\,, \tag{8a}$$

$$v_{n+1}(k,l) = c(k,l) + \lambda v_n(k,l+1) + \lambda v_n(k+1,l) + \mu v_n(k,l-1)\,. \tag{8b}$$

Comparing the right hand sides of (8a) and (8b) we immediately see that (2) for $n+1$ follows from (5) and the induction assumption for (2)-(4) for n.

Case b: $k = l > 0$.

Then

$$v_{n+1}(k,k+1) = c(k,k+1) + \lambda v_n(k,k+2) + \lambda v_n(k+1,k+1) + \mu v_n(k,k)\,,$$

$$v_{n+1}(k,k) = c(k,k) + 2\lambda v_n(k,k+1) + \mu v_n(k-1,k)\,.$$

So $v_{n+1}(k,k+1) \geq v_{n+1}(k,k)$.

Case c: $k = l = 0$.

From

$$v_{n+1}(0,1) = c(0,1) + \lambda v_n(0,2) + \lambda v_n(1,1) + \mu v_n(0,0)\,,$$

$$v_{n+1}(0,0) = c(0,0) + 2\lambda v_n(0,1) + \mu v_n(0,0;i) , \tag{9}$$

we directly get that $v_{n+1}(0,1) \geq v_{n+1}(0,0)$.

Proof of (3):

We have to distinguish between $k+1 < l$ and $k+1 = l$.

Case a: $k+1 < l$.

Then

$$v_{n+1}(k+1,l) = c(k+1,l) + \lambda v_n(k+1,l+1) + \lambda v_n(k+2,l) + \mu v_n(k+1,l-1) ,$$

$$v_{n+1}(k,l) = c(k,l) + \lambda v_n(k,l+1) + \lambda v_n(k+1,l) + \mu v_n(k,l-1) .$$

So $v_{n+1}(k+1,l) \geq v_{n+1}(k,l)$.

Case b: $k+1 = l$, so $(k+1,l) = (l,l)$ and $(k,l) = (l-1,l)$.

We have

$$v_{n+1}(l,l) = c(l,l) + 2\lambda v_n(l,l+1) + \mu v_n(l-1,l) ,$$

$$v_{n+1}(l-1,l) = c(l-1,l) + \lambda v_n(l-1,l+1) + \lambda v_n(l,l) + \mu v_n(l-1,l-1) .$$

Thus $v_{n+1}(l,l) \geq v_{n+1}(l-1,l)$.

Proof of (4):

From (9) and (2) and

$$v_{n+1}(0,0;i) = c(0,0;i) + 2\lambda v_n(0,0) + \mu v_n(0,0;i) ,$$

we immediately see that $v_{n+1}(0,0) \geq v_{n+1}(0,0;i)$.

5. THE LOWER BOUND MODEL; THRESHOLD REJECTION

Let $L\,(\geq 1)$ be the threshold. If an arrival in the longest queue leads to a difference of $L+1$ between the queue lengths then the job is rejected. Define the function $\delta_n := v_n - u_n$. Then we have to show that

$$\delta_n(s) \geq 0 \tag{10}$$

for all $n \geq 0$ and all s that are recurrent in the Threshold Rejection model. The proof follows by induction. For $n=0$ inequality (10) trivially holds. Assuming (10) to hold for n we prove it for $n+1$. We will distinguish five cases.

Case a: The states (k,l) with $0 \leq k < l < k+L$.

$$\delta_{n+1}(k,l) = \lambda\delta_n(k,l+1) + \lambda\delta_n(k+1,l) + \mu\delta_n(k,l-1) \geq 0 .$$

Case b: The states (k,k) with $k > 0$.

$$\delta_{n+1}(k,k) = 2\lambda\delta_n(k,k+1) + \mu\delta_n(k-1,k) \geq 0 .$$

Case c: The states $(k,k+L)$ with $k \geq 0$.

These are the only states in which the original longest queue model and the Threshold Rejection model differ. We have

$$u_{n+1}(k,k+L)=c(k,k+L)+\lambda u_n(k,k+L)+\lambda u_n(k+1,k+L)+\mu u_n(k,k+L-1)\,,$$

$$v_{n+1}(k,k+L)=c(k,k+L)+\lambda v_n(k,k+L+1)+\lambda v_n(k+1,k+L)+\mu v_n(k,k+L-1)\,.$$

So, using (2),

$$\begin{aligned}\delta_{n+1}(k,k+L)&=\lambda v_n(k,k+L+1)-\lambda u_n(k,k+L)+\lambda\delta_n(k+1,k+L)+\mu\delta_n(k,k+L-1)\\&\geq\lambda v_n(k,k+L)-\lambda u_n(k,k+L)+\lambda\delta_n(k+1,k+L)+\mu\delta_n(k,k+L-1)\\&=\lambda\delta_n(k,k+L)+\lambda\delta_n(k+1,k+L)+\mu\delta_n(k,k+L-1)\geq 0\,.\end{aligned}$$

Case d: The state (0,0).

$$\delta_{n+1}(0,0)=2\lambda\delta_n(1,0)+\mu\delta_n(0,0\,;i)\geq 0\,.$$

Case e: The state $(0,0\,;i)$.

$$\delta_{n+1}(0,0\,;i)=2\lambda\delta_n(0,0)+\mu\delta_n(0,0\,;i)\geq 0\,.$$

Conclusion

The Threshold Rejection model gives a lower bound for the average costs in the original longest queue model.

6. THE UPPER BOUND MODEL; THRESHOLD ADDITION

Let $L\,(\geq 1)$ be the threshold. If an arrival in the longest queue leads to a difference of $L+1$ between the queue lengths then a job is added to the (all) shortest queue(s) as well. The approach is identical to the one for the lower bound model in section 5. Define $\Delta_n := w_n - v_n$. We will show that

$$\Delta_n(s)\geq 0 \tag{11}$$

for all $n\geq 0$ and all s that are recurrent in the Threshold Addition model. The proof follows by induction. For $n=0$ inequality (11) trivially holds. Assuming (11) to hold for n we prove it for $n+1$. We have to distinguish the same five cases as in section 5. The only interesting situation is:

Case c: The states $(k,k+L)$ with $k\geq 0$.

$$w_{n+1}(k,k+L)=c(k,k+L)+\lambda w_n(k+1,k+L+1)+\lambda w_n(k+1,k+L)+\mu w_n(k,k+L-1)$$

$$v_{n+1}(k,k+L)=c(k,k+L)+\lambda v_n(k,k+L+1)+\lambda v_n(k+1,k+L)+\mu v_n(k,k+L-1)\,.$$

So, using (2),

$$\begin{aligned}\Delta_{n+1}(k,k+L)&=\lambda w_n(k+1,k+L+1)-\lambda v_n(k,k+L+1)+\lambda\Delta_n(k+1,k+L)+\mu\Delta_n(k,k+L-1)\\&\geq\lambda w_n(k+1,k+L+1)-\lambda v_n(k+1,k+L+1)+\lambda\Delta_n(k+1,k+L)+\mu\Delta_n(k,k+L-1)\\&=\lambda\Delta_n(k+1,k+L+1)+\lambda\Delta_n(k+1,k+L)+\mu\Delta_n(k,k+L-1)\geq 0\,.\end{aligned}$$

Conclusion

The Threshold Addition model gives an upper bound for the average costs in the original longest queue model.

7. ANALYSIS OF THE BOUND MODELS

From here on we consider the case with N (≥ 2) queues again. The model with Threshold Rejection will be analyzed first. This model can be described as an irreducible Markov chain with a state space consisting of the N-tuples $s=(s_1, ..., s_N)$ with $s_1 \leq \cdots \leq s_N \leq s_1+L$. So s_1 is the length of the shortest queue, s_2 is the length of the second shortest queue, and so on and the difference between the longest and the shortest queue is at most L. The state space can be partitioned into the single state $(0, ..., 0; i)$ and the levels 0, 1, ... where level l is defined as the set of states s with $s_1=l$. The states at a level are ordered lexicographically. For this partitioning the transition matrix P is of the form

$$P=\begin{bmatrix} B_{00} & B_{01} & 0 & 0 & 0 & \cdots \\ B_{10} & A_1 & A_0 & 0 & 0 & \cdots \\ 0 & A_2 & A_1 & A_0 & 0 & \cdots \\ 0 & 0 & A_2 & A_1 & A_0 & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots \end{bmatrix},$$

in which A_0, A_1 and A_2 are $m \times m$ matrices and B_{00}, B_{01} and B_{10} are 1×1, $1\times m$ and $m\times 1$ matrices respectively, where m is the number of states at a level, so

$$m=\binom{N+L-1}{L}.$$

We assume that the Markov chain P is ergodic, so that the equilibrium probability vector p exists. By partitioning the vector p into $p(0, ..., 0; i)$ and into the sequence of vectors $p_0, p_1, ...$ where p_l is the equilibrium probability vector of level l, we will show that

$$p_l = p_0 R^l, \quad l \geq 0, \tag{12}$$

for some nonnegative matrix R. Of course, this is well-known from the matrix-geometric theory developed by Neuts [6], where R is characterized as the minimal nonnegative solution of a quadratic matrix equation. But in our case the special structure of A_2 can be exploited to directly (without using results from [6]) obtain the matrix R in closed form (see also Ramaswami and Latouche [7]). Since it is only possible to jump from level l to level $l-1$ via state $(l, ..., l)$ to state $(l-1, l, ..., l)$ with probability μ, it follows that A_2 has only one nonzero entry, namely μ in the first row and kth column, where k denotes the position of $(l-1, l, ..., l)$ in level $l-1$. So A_2 can be written as

$$A_2 = \mu e_1 e_k^T,$$

where e_i is the ith unity column vector. Balancing the flow between level $l(\geq 0)$ and level $l+1$ yields

$$p_l A_0 e = p_{l+1} A_2 e = p_{l+1} \mu e_1 e_k^T e = p_{l+1} \mu e_1 ,$$

where e is the column vector with all its components equal to 1. Hence

$$p_{l+1} A_2 = p_{l+1} \mu e_1 e_k^T = p_l A_0 e e_k^T .$$

Substitution of this relation into the equilibrium equations at level l, i.e. into

$$p_l = p_{l-1} A_0 + p_l A_1 + p_{l+1} A_2 ,$$

leads to

$$p_l = p_{l-1} R , \tag{13}$$

where

$$R = A_0 (I - A_1 - A_0 e e_k^T)^{-1} ,$$

and I is the $m \times m$ identity matrix. The inverse of $I - A_1 - A_0 e e_k^T$ exists (and is nonnegative), since it is a substochastic matrix. From (13) the desired result (12) easily follows.

To finally complete the solution (12) the probability $p(0, ..., 0; i)$ and the vector p_0 have to be solved from the boundary conditions

$$p(0, ..., 0; i) = p(0, ..., 0; i) B_{00} + p_0 B_{10} ,$$

$$p_0 = p(0, ..., 0; i) B_{01} + p_0 (A_1 + R A_2) ,$$

together with the normalization equation

$$p(0, \cdots, 0; i) + \sum_{l=0}^{\infty} p_l e = 1 .$$

By substituting the form (12) for p_l, the sum of all $p_l e$ can be rewritten as

$$\sum_{l=0}^{\infty} p_l e = p_0 \sum_{l=0}^{\infty} R^l e = p_0 (I - R)^{-1} e .$$

The convergence of the series of powers R^l follows from the finiteness of the sum of all $p_l e$ (the Markov chain P is ergodic) and the fact that all components of p_0 are positive (the Markov chain P is irreducible). Hence, the normalization equation simplifies to

$$p(0, \cdots, 0; i) + p_0 (I - R)^{-1} e = 1 .$$

The model with Threshold Addition can be analyzed in the same way. The analysis only differs in the aspect of ergodicity. Since the model with Threshold Rejection destroys work, it is ergodic if the original model is ergodic. It is easily seen that the original model is ergodic if and only if $N\lambda < \mu$ by using that the total number of jobs in the system is stochastically the same as in the $M|M|1$ system with arrival rate $N\lambda$ and service rate μ. In the model with Threshold Addition extra work is created. So the condition $N\lambda < \mu$ is no longer sufficient, but the desired condition is formulated in [6] stating that the Markov chain is ergodic if and only if

$$\pi A_0 e < \pi A_2 e ,$$

where π is given by $\pi = \pi(A_0 + A_1 + A_2)$, $\pi e = 1$. The matrices A_0, A_1 and A_2 now correspond to the model with Threshold Addition (and are slightly different from the ones in the model with Threshold Rejection).

8. NUMERICAL RESULTS

To illustrate the method we present some numerical examples. Thinking of the repair of copiers we are interested in the service level $\beta(M)$ defined as the fraction of defects that can be repaired in time given that we have M spares of each part type. If at the time a defect occurs, at least M parts of the required type are in repair, then the defected part cannot be replaced by a good spare in time. So by using the PASTA property of Poisson arrivals it is easy to see that

$$\beta(M) = 1 - \frac{1}{N}\sum_{s} c(s)p(s),$$

where $c(s)$ is the number of queues with length at least M in state s. What we want to obtain is the minimal number M such that a given target service level β is satisfied. The two threshold models provide a lower and upper bound for the minimal M. Choosing the threshold L sufficiently large, the lower and upper bound will coincide, so that M can be determined exactly. In table 1 we list for increasing values of N and the workload ρ defined by $\rho = N\lambda/\mu$, the minimal M needed to satisfy the target service level β which is varied as 0.9, 0.95 and 0.99. L denotes the minimal threshold for which the lower and upper bound for M coincide and m is the number of states at a level for that value of L. The amount of work needed to solve the threshold models is roughly proportional to m^3. In the two examples with a * the service level may be slightly less than the target service level. The threshold models guarantee that $0.8999 \le \beta(6) \le 0.9004$ for $N=8$ and $\rho=0.95$ and $0.9499 \le \beta(2) \le 0.9501$ for $N=10$ and $\rho=0.8$.

Table 1: The minimal number of spares needed for each part type to satisfy the target service level.

β	ρ	$N=2$			$N=4$			$N=6$			$N=8$			$N=10$		
		M	L	m	M	L	m	M	L	m	M	L	m	M	L	m
0.90	0.50	2	2	3	2	3	20	1	2	21	1	2	36	1	2	55
	0.80	6	7	8	3	4	35	3	3	56	2	4	330	2	3	220
	0.90	11	6	7	6	5	56	5	6	462	4	4	330	3	4	715
	0.95	23	7	8	12	5	56	8	5	252	6*	6	1716	5	5	2002
0.95	0.50	3	3	4	2	2	10	2	2	21	2	2	36	1	3	220
	0.80	7	4	5	4	4	35	3	4	126	3	3	120	2*	5	2002
	0.90	15	8	9	8	5	56	6	5	252	5	5	792	4	4	715
	0.95	29	10	11	15	6	84	11	6	462	8	5	792	7	4	715
0.99	0.50	4	3	4	3	2	10	2	3	56	2	2	36	2	2	55
	0.80	11	5	6	6	4	35	5	5	252	4	3	120	3	4	715
	0.90	22	6	7	12	6	84	8	5	252	6	5	792	5	5	2002
	0.95	45	8	9	23	6	84	16	5	252	12	5	792	10	5	2002

From the results in table 1 we see that M can be determined exactly for already small values of the threshold L, even for high values of the workload ρ, and that the thresholds are smaller for larger systems. Surprisingly, the total number of spares NM appears to remain fairly constant as the number of different part types N varies.

Since we are able to efficiently evaluate the performance of the longest queue (LQ) policy, we can now compare the performance of this policy with the performance of the first-come first-served (FCFS) policy (which can be evaluated straightforwardly). In table 2 we display for several values of N, ρ and β the minimal number of spares needed of each part type to satisfy the target service level β under the LQ policy. The numbers in parentheses denote the extra spares needed for each part type under the FCFS policy.

Table 2: Comparison of the performance of the LQ and FCFS policy.

β	ρ	N=1	N=2	N=3	N=4	N=5	N=6	N=7	N=8	N=9	N=10
0.90	0.40	2 (0)	2 (-1)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)
	0.60	4 (0)	3 (0)	2 (0)	2 (0)	2 (0)	2 (0)	2 (0)	2 (-1)	1 (0)	1 (0)
	0.80	10 (0)	6 (0)	4 (0)	3 (0)	3 (0)	3 (0)	2 (1)	2 (0)	2 (0)	2 (0)
	0.90	21 (0)	11 (0)	8 (1)	6 (0)	5 (0)	5 (0)	4 (0)	4 (0)	3 (1)	3 (0)
	0.95	44 (0)	23 (0)	15 (1)	12 (0)	10 (0)	8 (1)	7 (1)	6 (1)	6 (0)	5 (1)
0.95	0.40	3 (0)	2 (0)	2 (0)	2 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)
	0.60	5 (0)	3 (0)	3 (0)	2 (0)	2 (0)	2 (0)	2 (0)	2 (0)	2 (0)	2 (0)
	0.80	13 (0)	7 (0)	5 (0)	4 (0)	4 (0)	3 (1)	3 (0)	3 (0)	3 (0)	2 (1)
	0.90	28 (0)	15 (0)	10 (1)	8 (0)	7 (0)	6 (0)	5 (1)	5 (0)	4 (1)	4 (0)
	0.95	58 (0)	29 (1)	20 (1)	15 (1)	12 (1)	11 (0)	9 (1)	8 (1)	7 (1)	7 (0)
0.99	0.40	5 (0)	3 (0)	3 (0)	2 (0)	2 (0)	2 (0)	2 (0)	2 (0)	2 (0)	2 (0)
	0.60	9 (0)	5 (0)	4 (0)	3 (1)	3 (0)	3 (0)	3 (0)	3 (0)	2 (1)	2 (1)
	0.80	20 (0)	11 (0)	8 (0)	6 (1)	5 (1)	5 (0)	4 (1)	4 (0)	4 (0)	3 (0)
	0.90	43 (0)	22 (1)	15 (1)	12 (1)	10 (1)	8 (1)	7 (1)	6 (2)	6 (1)	5 (2)
	0.95	89 (0)	45 (1)	30 (2)	23 (1)	19 (1)	16 (1)	14 (1)	12 (1)	11 (1)	10 (1)

We see that in most cases the LQ policy is better that the FCFS policy in the sense that less spares are needed. This may be explained by the property that for each part type the variance of the number of parts in repair under the LQ policy is (most likely) smaller than under the FCFS policy (the mean is of course the same). For the case of two queues this property has been proved by Zheng and Zipkin [9].

References

1. ADAN, I.J.B.F., HOUTUM, G.J. VAN, AND WAL, J. VAN DER, ''Upper and lower bounds for the waiting time in the symmetric shortest queue system,,'' *Annals of Operat. Research*, vol. 48, pp. 197-217, 1994.

2. COHEN, J.W., ''A two-queue, one-server model with priority for the longer queue,'' *Queueing systems*, vol. 2, pp. 261-283, 1987.

3. DIJK, N.M. VAN AND LAMOND, B.F., ''Simple bounds for finite single-server exponential tandem queues,'' *Opns. Res.*, vol. 36, pp. 470-477, 1988.

4. DIJK, N.M. VAN AND WAL, J. VAN DER, ''Simple bounds and monotonicity results for finite multi-server exponential tandem queues,'' *Queueing Systems*, vol. 4, pp. 1-16, 1989.

5. FLATTO, L., ''The longer queue model,'' *Prob. Engineer. Inform. Sci.*, vol. 3, pp. 537-559, 1989.

6. NEUTS, M.F., *Matrix-geometric solutions in stochastic models,* Johns Hopkins University Press, Baltimore, 1981.

7. RAMASWAMI, V. AND LATOUCHE, G., ''A general class of Markov processes with explicit matrix-geometric solutions,'' *OR Spektrum*, vol. 8, pp. 209-218, 1986.

8. WAL, J. VAN DER, ''Monotonicity of the throughput of a closed exponential queueing network in the number of jobs,'' *OR Spektrum*, vol. 11, pp. 97-100, 1989.

9. ZHENG, Y.S. AND ZIPKIN, P., ''A queuing model to analyze the value of centralized inventory information,'' *Opns. Res.*, vol. 38, pp. 296-307, 1990.