Upper and lower bounds for the waiting time in the symmetric shortest queue system

EINDHOVEN UNIVERSITY OF TECHNOLOGY
Department of Mathematics and Computing Science

Memorandum COSOR 92-09

Upper and lower bounds for the waiting time in the symmetric shortest queue system

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

Eindhoven, May 1992
The Netherlands

Eindhoven University of Technology
Department of Mathematics and Computing Science
Probability theory, statistics, operations research and systems theory
P.O. Box 513
5600 MB Eindhoven - The Netherlands

Secretariate: Dommelbuilding 0.03
Telephone: 040-47 3130

ISSN 0926 4493

UPPER AND LOWER BOUNDS FOR THE WAITING TIME IN THE SYMMETRIC SHORTEST QUEUE SYSTEM

Ivo Adan, Geert-Jan van Houtum, Jan van der Wal,
University of Technology Eindhoven, April 28, 1992.

ABSTRACT

In this paper we compare the exponential symmetric shortest queue system with two related systems: the shortest queue system with threshold jockeying and the shortest queue system with threshold blocking. The latter two systems are easier to analyse and are shown to give tight lower and upper bounds respectively for the mean waiting time in the shortest queue system. The approach also gives bounds for the distribution of the total number of jobs in the system.

1. INTRODUCTION

In job shop like production systems, flexible manufacturing systems and computer and communication networks with parallel server stations one might have to assign the jobs to a specific server immediately upon the time of arrival. Then a natural strategy is to assign an arriving job to the server with the shortest queue. However, the impact of the shortest queue assignment strategy on the system behaviour is not intuitively clear. Therefore, methods are being developed to evaluate its performance. In this paper we present an approach to efficiently calculate the mean waiting time for shortest queue routing at a single station.

The symmetric shortest queue system with two queues, Poisson arrivals and exponential service times has been extensively studied in the literature. Haight [13] introduced the problem. Kingman [16] and Flatto and McKean [9] use a uniformization technique to determine the generating function of the stationary queue length distribution. Another analytic approach is given in Cohen and Boxma [4] and Fayolle and Iasnogorodski [8]. They show that the analysis of the symmetric shortest queue system can be reduced to that of a Riemann-Hilbert boundary value problem. Though mathematically elegant, these analytical results offer no practical means for computing performance

characteristics. Recently, it has been shown that the symmetric shortest queue problem can be solved completely by the compensation approach introduced in Adan, Wessels and Zijm [1]. The advantage of this approach is that the analytical results are easily exploited for numerical computations. Many numerical studies have appeared on the present problem. Most studies however, deal with the evaluation of approximating models, see e.g. Gertsbakh [11], Grassmann [12], Rao and Posner [21] and Conolly [5]. Using linear programming, Halfin [14] obtains upper and lower bounds for the mean and the distribution of the number of jobs in the system. Heavy traffic diffusion approximations can be found in Foschini and Salz [10]. Knessl, Matkowsky, Schuss and Tier [17] derive asymptotic expressions for the stationary queue length distribution. These studies are all restricted to systems with two parallel queues.

For more than two queues no analytical results are available. Hooghiemstra, Keane and Van de Ree [15] develop a power series method to calculate the stationary queue length distribution for fairly general multidimensional exponential queueing systems. Their method is not restricted to systems with two queues, but applies equally well to systems with more queues. As far as the shortest queue system is concerned, Blanc [3] reports that the power series method is numerically satisfactory for practically all values of the work load for systems with up to 10 parallel queues. However, the theoretical foundation of this method is still incomplete. Nelson and Philips [18] derive an approximation for the mean response time for the shortest queue system with multiple queues. They report that their approximation has a relative error of less than 2 percent for systems with at most 16 queues and with service utilizations over the range from 0 to 0.99. A common disadvantage of the numerical methods mentioned is that in general no error bounds can be given.

In this paper we derive upper and lower bounds for the mean waiting time in the shortest queue system by comparing it with two other queueing models which are easier to evaluate.

The model that produces the lower bound for the waiting time is the shortest queue system with Threshold Jockeying, i.e., if the difference between the longest queue and the shortest queue exceeds a certain threshold, then one customer moves from the longest queue to the shortest queue.

This will improve the performance of the system somewhat because the situation in which one server is idle while there are waiting customers in another queue is less likely to occur. Note that for a threshold of 1 the system behaves as an $M \mid M \mid N$ queue with N the number of servers.

The unbalance in the queue lengths can only occur if temporarily the servicing in one

queue goes so much faster than the servicing in another queue that even sending all arrivals to the shorter queue can not compensate it. Therefore one might expect that for a somewhat larger threshold (3 or 4 maybe) the mean waiting time will be almost identical to the one in the original system.

A model giving an upper bound is the model with Threshold Blocking. In this model the server in the shortest queue is switched off as soon as the difference between the longest and the shortest queue reaches a certain threshold. If the difference drops below the threshold the servicing is resumed. It seems clear that blocking a server will have a negative impact on the performance of the system. And, as before, the larger the threshold is taken, the less blocking one should get and the better the bound should be.

We will prove that these two queueing models indeed provide lower and upper bounds for the mean waiting time in the shortest queue system. Even more, we will see that the total number of jobs in the system is stochastically smaller and larger, respectively, than in the original shortest queue system.

The line of proof is similar to the ones in Van der Wal [22], Van Dijk and Van der Wal [7] and Van Dijk and Lamond [6]. First of all we note that the three models lead to Markov processes that are equivalent to Markov chains. The mean performance characteristics of these Markov chains are the same as the ones of the Markov processes. In order to compare the Markov chains we look at finite period costs. By induction we show one model to be superior to another for each finite number of periods. Letting the number of periods go to infinity then yields the result for the average performance.
We will only give the proofs for the case of two queues. For more than two queues the proofs are essentially the same but notationally more complex.

In section 2 we discuss the translation of the Markov processes into the equivalent Markov chains. Section 3 discusses the technique of comparing the finite period models. Section 4 deals with two essential monotonicity results for the original shortest queue system: less customers in the system lead to lower costs, and a balanced system performs better than an unbalanced one. In sections 5 and 6 we prove that the models with threshold jockeying and threshold blocking lead to lower and upper bounds on the mean waiting time. In section we consider the stochastic monotonicity of the total number of jobs in the three systems. In section 8 it is shown how these two queueing models can be analyzed. Section 9 provides numerical results that show that the mean waiting time in the shortest queue system can be obtained from the two other systems using very moderate threshold values.

2. MARKOV PROCESSES AND MARKOV CHAINS

In all three Markov processes jobs arrive according to a Poisson process with rate λ. Immediately upon arrival a job is sent to the shortest queue. All service times are exponential with rate μ. The state of the Markov process is a vector $s=(s_1, \cdots, s_N)$, with N the number of servers. Exploiting the symmetry in the model we may assume that $s_1 \geq \cdots \geq s_N$. So s_1 is the number of jobs in the longest queue, s_2 is the number of jobs in the second longest queue and so on. The maximal output rate from a state is $\lambda+N\mu$. Without loss of generality we may take $\lambda+N\mu=1$.

The original shortest queue system and the model with threshold jockeying are ergodic if $N\mu>\lambda$. For the model with threshold blocking this condition is not sufficient. The condition under which the model with threshold blocking is ergodic will be formulated in section 7.

Let Q be the generator of one of the three Markov processes that we are dealing with, then the corresponding equilibrium distribution p satisfies $pQ=0$. Instead of studying the Markov process with generator Q we look at the Markov chain with transition matrix $P=I+Q$. Recall that $\lambda+N\mu=1$, so P is indeed a stochastic matrix. Clearly the equilibrium distribution p of the Markov process is the same as the one for the Markov chain. Also mean costs in the Markov process and the Markov chain are easy to compare. If $c(s)$ is the cost rate in the Markov process and we let $c(s)$ be the costs per period in the Markov chain, then the Markov process and the Markov chain will have the same average costs: $\Sigma_s\, p(s)c(s)$.

From now on we only consider the three Markov chains.

3. FINITE PERIOD MODELS

For the basic shortest queue system we define $v_n(s)$ as the expected n-period costs when starting in state s. Similarly u_n and w_n denote the expected n-period costs in the model with threshold jockeying and threshold blocking. Defining $u_0=v_0=w_0=0$ we will try to prove by induction that for all (relevant) states s and for all n

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

Of course this implies that the average costs for the three models are ordered in the same way.

If the cost function is the total number of jobs in the system, then we can conclude that the average number of jobs in the system for the three models are ordered. And using Little's formula the same holds for the mean waiting times.

If the costs are 1 if the total number of jobs in the system exceeds M and 0 otherwise, and the ordering result holds for all M, then the total number of jobs in the systems are stochastically ordered.

In order to be able to prove (1) we first have to establish some monotonicity results for the functions v_n.

4. MONOTONICITY OF THE FUNCTIONS v_n

From hereon we will consider the case of two queues only. The results hold for $N \geq 3$ as well, but the notations become more complex.

Because of the symmetry we only have to consider states $s=(i+l,i)$ with nonnegative integers i and l.

As cost function we use the total number of jobs in the system, so $c(i+l,i)=2i+l$.

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

Lemma 1

For all $n \geq 0$ we have

$$v_n(i+l,i) \geq v_n(i+l-1,i+1), \quad i \geq 0,\ l \geq 2 \tag{2}$$

$$v_n(i+l+1,i) \geq v_n(i+l,i), \quad i \geq 0,\ l \geq 0 \tag{3}$$

$$v_n(i+l,i+1) \geq v_n(i+l,i), \quad i \geq 0,\ l \geq 1. \tag{4}$$

So (2) states that more balance at the beginning leads to lower costs. Inequalities (3) and (4) say that removal of a job from the system reduces the expected costs.

Proof of Lemma 1:

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

In order to prove this lemma we will distinguish a number of cases.

Proof of (2):

Case a: $l \geq 3$.

We have

$$v_{n+1}(i+l,i)=2i+l+\lambda v_n(i+l,i+1) \tag{5a}$$
$$+\mu v_n(i+l-1,i)+\mu v_n(i+l,(i-1)^+)$$

and

$$v_{n+1}(i+l-1,i+1)=2i+l+\lambda v_n(i+l-1,i+2) \qquad (5b)$$
$$+\mu v_n(i+l-2,i+1)+\mu v_n(i+l-1,i)\,.$$

Now compare the right hand sides (RHS) of (5a) and (5b). The first terms are equal, the second term in (5a) is at least equal to the second one in (5b) because of (2), similarly the third and forth terms are ordered. For the forth term the case $i=0$ follows from (3).
So $v_{n+1}(i+l,i)\geq v_{n+1}(i+l-1,i+1)$.

Case b: $l=2$.

$$v_{n+1}(i+2,i)=2i+2+\lambda v_n(i+2,i+1)+\mu v_n(i+1,i)+\mu v_n(i+2,(i-1)^+) \qquad (6a)$$

$$v_{n+1}(i+1,i+1)=2i+2+\lambda v_n(i+2,i+1)+\mu v_n(i+1,i)+\mu v_n(i+1,i)\,. \qquad (6b)$$

The first, second and third terms in the RHS are equal, the forth terms are ordered because of (2) if $i>0$ or (3) if $i=0$.
So $v_{n+1}(i+2,i)\geq v_{n+1}(i+1,i+1)$.

Proof of (3):

Case a: $l\geq 1$.

$$v_{n+1}(i+l+1,i)=2i+l+1+\lambda v_n(i+l+1,i+1) \qquad (7a)$$
$$+\mu v_n(i+l,i)+\mu v_n(i+l+1,(i-1)^+)$$

$$v_{n+1}(i+l,i)=2i+l+\lambda v_n(i+l,i+1) \qquad (7b)$$
$$+\mu v_n(i+l-1,i)+\mu v_n(i+l,(i-1)^+)\,.$$

The first term in the RHS of (7a) is larger than the one in (7b), the other three terms are ordered by (3).
So $v_{n+1}(i+l+1,i)\geq v_{n+1}(i+l,i)$.

Case b: $l=0$.

$$v_{n+1}(i+1,i)=2i+1+\lambda v_n(i+1,i+1)+\mu v_n(i,i)+\mu v_n(i+1,(i-1)^+) \qquad (8a)$$

$$v_{n+1}(i,i)=2i+\lambda v_n(i+1,i)+\mu v_n(i,(i-1)^+)+\mu v_n(i,(i-1)^+)\,. \qquad (8b)$$

The first term in the RHS of (8a) is larger than the one in (8b), the second and third terms are ordered by (4) (and equal if $i=0$) and the forth terms are ordered by (3).
So $v_{n+1}(i+1,i)\geq v_{n+1}(i,i)$.

Proof of (4):

Case a: $l \geq 2$.

$$v_{n+1}(i+l,i+1) = 2i+l+1+\lambda v_n(i+l,i+2)+\mu v_n(i+l-1,i+1)+\mu v_n(i+l,i)\,. \quad (9a)$$

$$v_{n+1}(i+l,i) = 2i+l+\lambda v_n(i+l,i+1)+\mu v_n(i+l-1,i)+\mu v_n(i+l,(i-1)^+)\,. \quad (9b)$$

The first term in the RHS of (9a) is larger than in (9b), the other three terms are ordered by (4) (the forth terms are equal if $i=0$).
So $v_{n+1}(i+l,i+1) \geq v_{n+1}(i+l,i)$.

Case b: $l=1$.

$$v_{n+1}(i+1,i+1) = 2i+2+\lambda v_n(i+2,i+1)+\mu v_n(i+1,i)+\mu v_n(i+1,i) \quad (10a)$$

$$v_{n+1}(i+1,i) = 2i+1+\lambda v_n(i+1,i+1)+\mu v_n(i,i)+\mu v_n(i+1,(i-1)^+)\,. \quad (10b)$$

The first term in the RHS of (10a) is larger than in (10b), the second and third term are ordered by (3) and the forth terms are ordered by (4) (and equal if $i=0$).
So $v_{n+1}(i+1,i+1) \geq v_{n+1}(i+1,i)$.

5. THRESHOLD JOCKEYING

Let L be the threshold we consider. If a service completion leads to a difference of $L+1$ between the numbers of jobs in the longest and shortest queue, then one job moves from the longest to the shortest queue.
The n-period costs in the threshold jockeying model are denoted by u_n. It suffices to show that

$$u_n(s) \leq v_n(s) \quad (11)$$

for all states s that are recurrent in the jockeying 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 will distinguish four cases.

Case a: The states $(i+l,i)$ with $l=1, \cdots, L-1$.

$$u_{n+1}(i+l,i) = 2i+l+\lambda u_n(i+l,i+1)+\mu u_n(i+l-1,i)+\mu u_n(i+l,(i-1)^+) \quad (12a)$$

$$v_{n+1}(i+l,i) = 2i+l+\lambda v_n(i+l,i+1)+\mu v_n(i+l-1,i)+\mu v_n(i+l,(i-1)^+) \quad (12b)$$

So from (11) for n we get $u_{n+1}(i+l,i) \leq v_{n+1}(i+l,i)$.

Case b: The states (i,i).

$$u_{n+1}(i,i) = 2i + \lambda u_n(i+1,i) + \mu u_n(i, (i-1)^+)) + \mu u_n(i, (i-1)^+) \tag{13a}$$

$$v_{n+1}(i,i) = 2i + \lambda v_n(i+1,i) + \mu v_n(i, (i-1)^+)) + \mu v_n(i, (i-1)^+) \tag{13b}$$

So $u_{n+1}(i,i) \leq v_{n+1}(i,i)$.

Case c: The states $(i+L,i)$ with $i > 0$.

$$u_{n+1}(i+L,i) = 2i + L + \lambda u_n(i+L,i+1) + \mu u_n(i+L-1,i) + \mu u_n(i+L-1,i) \tag{14a}$$

$$v_{n+1}(i+L,i) = 2i + L + \lambda v_n(i+L,i+1) + \mu v_n(i+L-1,i) + \mu v_n(i+L,i-1) \tag{14b}$$

Comparing the RHS of (14a) and (14b) the only difficult term is the forth one. But by (2) we have $v_n(i+L,i-1) \geq v_n(i+L-1,i)$, so (11) gives $u_n(i+L-1,i) \leq v_n(i+L,i-1)$. Hence $u_{n+1}(i+L,i) \leq v_{n+1}(i+L,i)$ for $i > 0$.

Case d: The state $(L, 0)$.

$$u_{n+1}(L, 0) = L + \lambda u_n(L, 1) + \mu u_n(L-1,0) + \mu u_n(L, 0) \tag{15a}$$

$$v_{n+1}(L, 0) = L + \lambda v_n(L, 1) + \mu v_n(L-1,0) + \mu v_n(L, 0) \tag{15b}$$

So (11) immediately gives $u_{n+1}(L, 0) \leq v_{n+1}(L, 0)$.

This completes the proof of (11).

Conclusion

The jockeying model underestimates the mean number of jobs in the system, and hence gives a lower bound for the mean waiting time.

6. THRESHOLD BLOCKING

In the threshold blocking model the server in the shortest queue is blocked as long as the difference between the queue lengths in the longest and shortest queue equals the threshold L. In order to prove that threshold blocking yields an upper bound on the mean waiting time we will show that for all states s that are recurrent for the threshold blocking model and for all n we have

$$v_n(s) \leq w_n(s) \,. \tag{16}$$

As said before, w_n denotes the n-period costs for threshold blocking.
By definition $v_0 = w_0 = 0$, so (16) holds for $n = 0$. Assuming (16) to hold for n we will

prove it for $n+1$.

We will distinguish three cases.

Case a: The states $(i+l,i)$ with $l=1, \cdots, L-1$.

$$w_{n+1}(i+l,i)=2i+l+\lambda w_n(i+l,i+1)+\mu w_n(i+l-1,i)+\mu w_n(i+l,(i-1)^+) \quad (17a)$$

$$v_{n+1}(i+l,i)=2i+l+\lambda v_n(i+l,i+1)+\mu v_n(i+l-1,i)+\mu v_n(i+l,(i-1)^+)\,. \quad (17b)$$

So (16) immediately gives $v_{n+1}(i+l,i)\le w_{n+1}(i+l,i)$.

Case b: The states (i,i).

$$w_{n+1}(i,i)=2i+\lambda w_n(i+1,i)+\mu w_n(i,(i-1)^+)+\mu w_n(i,(i-1)^+) \quad (18a)$$

$$v_{n+1}(i,i)=2i+\lambda v_n(i+1,i)+\mu v_n(i,(i-1)^+)+\mu v_n(i,(i-1)^+) \quad (18b)$$

So $v_{n+1}(i,i)\le w_{n+1}(i,i)$.

Case c: The states $(i+L,i)$.

$$w_{n+1}(i+L,i)=2i+L+\lambda w_n(i+L,i+1)+\mu w_n(i+L-1,i)+\mu w_n(i+L,i) \quad (19a)$$

$$v_{n+1}(i+L,i)=2i+L+\lambda v_n(i+L,i+1)+\mu v_n(i+L-1,i)+\mu v_n(i+L,(i-1)^+)\,. \quad (19b)$$

Using (16), and for the forth term (4), we get $v_{n+1}(i+L,i)\le w_{n+1}(i+L,i)$.

This completes the proof of (16) for $n+1$.

Conclusion

Threshold blocking gives upper bounds for the mean number of jobs and the mean waiting time in the original shortest queue system.

7. STOCHASTIC MONOTONICITY

In sections 4-6 we were interested in the mean number of jobs in the system. Using the same approach one may establish the stochastic monotonicity of the total number of jobs in the system. Let M be any integer, and define the cost function $c(s)=1$ if $\Sigma s_i \ge M$ and 0 otherwise. It is easily seen that for this cost function the monotonicity results in the sections 4-6 hold as well. So, writing $F_{TJ}(i)$, $F_{SQ}(i)$ and $F_{TB}(i)$ for the probability of having at least i jobs in the jockeying system, the standard shortest queue system and the blocking system respectively, we have for all i

$$F_{TJ}(i) \geq F_{SQ}(i) \geq F_{TB}(i) .$$

8. ANALYZING THE TWO THRESHOLD MODELS

So far we have shown that the mean number of jobs and the mean waiting time of the shortest queue model can be bounded between the corresponding quantities of two threshold models. In this section it will be shown that the two threshold models are easier to evaluate than the shortest queue model.

From now on we consider the general situation with N servers again. The states of the two threshold models are characterized by the vectors $s=(s_1, \cdots, s_N)$ where by symmetry we may assume that $s_1 \geq \cdots \geq s_N$. So s_1 is the length of the longest queue, s_2 is the length of the second longest queue, and so on. The analysis can be restricted to the recurrent states, which are the ones with $s_1 \leq s_N+L$. This form of state space suggests to use the matrix-geometric approach, as developed by Neuts [19]. Indeed, it appears that this approach is very well suited for the analysis of the two models. The advantage of this approach is that the problem of solving infinitely many equilibrium equations is reduced to that of solving finitely many.

8.1 Threshold blocking

We first consider the model with threshold blocking. Application of the matrix-geometric approach requires a partitioning of the state space. Let us first define level l as the set of states s with $s_1 = l$. Then we partition the state space into the levels $0, 1, \cdots, L, L+1, \cdots$ and put together the levels $0, 1, \cdots, L-1$ with less regular behaviour in one set. The states at each level may be ordered lexicographically. For this partitioning the generator Q is of the form

$$Q = \begin{bmatrix} B_{00} & B_{01} & 0 & 0 & 0 & \cdots \\ B_{10} & B_{11} & A_0 & 0 & 0 & \cdots \\ 0 & A_2 & A_1 & A_0 & 0 & \cdots \\ 0 & 0 & A_2 & A_1 & A_0 & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots \end{bmatrix} . \tag{20}$$

The blocks A_0, A_1 and A_2 are of order m where m is the number of states at a level $\geq L$, so

$$m = \begin{bmatrix} N+L-1 \\ L \end{bmatrix} .$$

The Markov process Q is irreducible and, since two states at levels $>L$ can reach each other via paths not passing through levels $\leq L$, the generator $A_0+A_1+A_2$ is also irreducible. So theorem 1.7.1 in Neuts [19] can readily be applied. As mentioned in section 2 the condition $N\mu > \lambda$ is not sufficient for ergodicity, but the desired condition is given by theorem 1.7.1 stating that Q is ergodic if and only if

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

where e is the column vector of ones and π is the solution of

$$\pi(A_0+A_1+A_2)=0 , \quad \pi e = 1 .$$

By partitioning the equilibrium probability vector p into the vector $(p_0, \cdots, p_{L-1})$ and into the sequence of vectors $p_L, p_{L+1}, \ldots$ where p_l is the equilibrium probability vector of level l, we conclude from the same theorem that if Q is ergodic, then

$$p_l = p_L R^{l-L} , \quad l > L , \tag{21}$$

where the matrix R is the minimal nonnegative solution of the matrix quadratic equation

$$A_0 + RA_1 + R^2 A_2 = 0 . \tag{22}$$

We shall now show that due to the special matrix structure of A_0, the matrix R can be determined explicitly, once its maximal eigenvalue is known. Since it is only possible to jump from level l to level $l+1$ via state $(l, \cdots, l)$, it follows that all rows of A_0 are zero, except for the last row. Thus A_0 is of the form

$$A_0 = \begin{bmatrix} 0 \\ x \end{bmatrix}$$

where the vector $x=(x_1, \cdots, x_m)$. Since the rows in R corresponding to the zero rows in A_0, are also zero (see e.g. the proof of theorem 1.3.4 in [19]), we conclude that R is also of the form

$$R = \begin{bmatrix} 0 \\ y \end{bmatrix} \tag{23}$$

where the vector $y=(y_1, \cdots, y_m)$. Obviously, the component y_m is the maximal eigenvalue of R. This eigenvalue can be evaluated, without computing R first, by using an algorithm suggested in [19]. The other components of y may be solved from the matrix equation (22), which by insertion of the special forms of R and A_0 simplifies to the following set of linear equations

$$x + y(A_1 + y_m A_2) = 0 ,$$

so

$$y=-x(A_1+y_mA_2)^{-1}\ . \tag{24}$$

The inverse of $A_1+y_mA_2$ exists, since this matrix can be interpreted as a transient generator (escape is possible at least from the last state). The computation of the maximal eigenvalue of R however, can be costly for large values of m. In such case it may be more efficient to solve y by successive substitutions from

$$y=x+y(I+A_1+y_mA_2)\ , \tag{25}$$

where I is the $m\times m$ identity matrix (cf. page 9 in [19]).

By substituting the form (23) in (21) the matrix-geometric solution simplifies to

$$p_l=y_m^{l-L-1}\,p(L,\ \cdots,L)\,y\ ,\quad l>L\ . \tag{26}$$

The remaining probability vectors p_L, p_{L-1}, ..., p_0 can be solved from the boundary conditions. In fact, again by exploiting the property that it is only possible to jump from level l to level $l+1$ via state $(l,\ \cdots,l)$, it is easily verified that these vectors can be solved recursively from the equilibrium equations for the levels L, $L-1$, $\cdots$, 0.

8.2 Threshold jockeying

The model with threshold jockeying can be treated in exactly the same way. In addition, for this model it can be shown that $N\mu>\lambda$ is necessary and sufficient for ergodicity (see Adan, Wessels and Zijm [2]) and the maximal eigenvalue y_m can be found explicitly by using the following balance argument.

Let V_l be the set of states with $s_1+\cdots+s_N=l$ and $P(V_l)$ be the equilibrium probability for the set V_l. By balancing the flow between the sets V_l and V_{l+1} it follows that for all $l>(N-1)L$

$$P(V_{l+1})N\mu=P(V_l)\lambda\ ,$$

and by applying this relation N times,

$$P(V_{l+N})=P(V_l)\left[\frac{\lambda}{N\mu}\right]^N\ . \tag{27}$$

On the other hand, the set V_{Nl} is a subset of the union of the levels l, $l+1$, $\cdots$, $l+N-1$, so it follows from (26) that for all $l>L$

$$P(V_{Nl})=Ky_m^l\ , \tag{28}$$

for some constant K being independent of l. Combining (27) and (28) yields

$$y_m=\left[\frac{\lambda}{N\mu}\right]^N\ .$$

Remark

The explicit solution of R for the two models is mainly due to the special matrix structure of A_0. In fact, Ramaswami and Latouche [20] show that if the generator Q has the form (20) and A_0 is given by $A_0 = a \cdot b$ where a is a column vector and b is a row vector, then R is explicitly determined, once its maximal eigenvalue is known.

9. NUMERICAL RESULTS

This section is devoted to some numerical results. In the tables 1-3 we list for systems with 2, 5 and 8 parallel servers the upper and lower bounds for the 'normalized' mean waiting time for increasing values of the threshold L and the workload ρ defined by

$$\rho = \frac{\lambda}{N\mu} .$$

(The normalized mean waiting time is defined as the quotient of the mean waiting time and the mean service time.)

As L tends to infinity, then the upper and lower bounds can be expected to converge to the mean waiting time of the shortest queue system.

The numerical effort to calculate the mean waiting time for the blocking model essentially consists of first solving the vector y from (25) by successive substitutions and then recursively solving the equilibrium equations at the levels L, $L-1$, $\cdots$, 0. For the jockeying model the vector y is solved from the linear equations (24) with $y_m = \rho^N$. The dimension m of the vector y and thereby the numerical effort increases fast in N.

$N=2$	Mean Waiting Time / Mean Service Time						
	Threshold Jockeying				Threshold Blocking		
ρ	$L=2$	$L=4$	$L=8$	$L=\infty$	$L=8$	$L=4$	$L=2$
0.20	0.065	0.066	0.066	0.066	0.066	0.066	0.067
0.50	0.409	0.426	0.426	0.426	0.426	0.427	0.487
0.80	1.889	1.949	1.956	1.956	1.956	2.009	3.460
0.90	4.382	4.461	4.475	4.475	4.477	4.829	30.265
0.95	9.379	9.468	9.486	9.486	9.501	11.319	–

Table 1: Bounds for the mean normalized waiting time for the case of 2 servers.

$N=5$	Mean Waiting Time / Mean Service Time						
	Threshold Jockeying				Threshold Blocking		
ρ	$L=2$	$L=4$	$L=6$	$L=\infty$	$L=6$	$L=4$	$L=2$
0.20	0.003	0.003	0.003	0.003	0.003	0.003	0.003
0.50	0.107	0.109	0.109	0.109	0.109	0.109	0.116
0.80	0.717	0.754	0.754	0.754	0.754	0.758	1.630
0.90	1.722	1.795	1.798	1.798	1.799	1.852	–
0.95	3.724	3.821	3.827	3.827	3.836	4.173	–

Table 2: Bounds for the mean normalized waiting time for the case of 5 servers.

$N=8$	Mean Waiting Time / Mean Service Time						
	Threshold Jockeying				Threshold Blocking		
ρ	$L=2$	$L=3$	$L=4$	$L=\infty$	$L=4$	$L=3$	$L=2$
0.20	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.50	0.042	0.042	0.042	0.042	0.042	0.042	0.043
0.80	0.440	0.455	0.456	0.456	0.456	0.466	0.828
0.90	1.087	1.128	1.134	1.134	1.145	1.278	–
0.95	2.347	2.409	2.422	2.424	2.514	3.510	–

Table 3: Bounds for the mean normalized waiting time for the case of 8 servers.

The numerical results in the tables 1-3 illustrate that the bounds are tight for already small values of L. Under heavy load the threshold L for the blocking model is larger than the one for the jockeying model in order to produce upper and lower bounds with the same accuracy. Apparently, temporarily switching off the servers in the blocking model under heavy load has an important effect on the mean waiting time.

10. CONCLUSION

As we have seen it is possible to derive tight bounds on the mean waiting time in the shortest queue system by comparing it with two similar systems that are easier to analyse: the shortest queue system with threshold jockeying and the one with threshold blocking. We can also get bounds for the distribution of the total number of jobs in the system.

Using these results it should possible to produce fairly simple approximations for the behaviour of shortest queue stations in queueing networks.

References

1. ADAN, I.J.B.F., WESSELS, J., AND ZIJM, W.H.M., "Analysis of the symmetric shortest queue problem," *Stochastic Models*, vol. 6, pp. 691-713, 1990.
2. ADAN, I.J.B.F., WESSELS, J., AND ZIJM, W.H.M., "Matrix-geometric analysis of the shortest queue problem with threshold jockeying," Memorandum COSOR 91-24, Eindhoven University of Technology, Dep. of Math. and Comp. Sci., 1991 (submitted for publication).
3. BLANC, J.P.C., "The power-series algoritm applied to the shortest-queue model," *Opns. Res.*, vol. 40, pp. 157-167, 1992.
4. COHEN, J.W. AND BOXMA, O.J., *Boundary value problems in queueing system analysis,* North-Holland, Amsterdam, 1983.
5. CONOLLY, B.W., "The autostrada queueing problem," *J. Appl. Prob.*, vol. 21, pp. 394-403, 1984.
6. 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.
7. DIJK, N. 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.
8. FAYOLLE, G. AND IASNOGORODSKI, R., "Two coupled processors: the reduction to a Riemann-Hilbert problem," *Z. Wahrsch. Verw. Gebiete*, vol. 47, pp. 325-351, 1979.
9. FLATTO, L. AND MCKEAN, H.P., "Two queues in parallel," *Comm. Pure Appl. Math.*, vol. 30, pp. 255-263, 1977.
10. FOSCHINI, G. J. AND SALZ, J., "A basic dynamic routing problem and diffusion," *IEEE Trans. Commun.*, vol. COM-26, pp. 320-327, 1978.
11. GERTSBAKH, I., "The shorter queue problem: A numerical study using the matrix-geometric solution," *Eur. J. Oper. Res.*, vol. 15, pp. 374-381, 1984.
12. GRASSMANN, W.K., "Transient and steady state results for two parallel queues," *OMEGA Int. J. of Mgmt Sci.*, vol. 8, pp. 105-112, 1980.

13. HAIGHT, F.A., "Two queues in parallel," *Biometrica*, vol. 45, pp. 401-410, 1958.

14. HALFIN, S., "The shortest queue problem," *J. Appl. Prob.*, vol. 22, pp. 865-878, 1985.

15. HOOGHIEMSTRA, G., KEANE, M., AND REE, S. VAN DE, "Power series for stationary distributions of coupled processor models," *SIAM J. Appl. Math.*, vol. 48, pp. 1159-1166, 1988.

16. KINGMAN, J.F.C., "Two similar queues in parallel," *Ann. Math. Statist.*, vol. 32, pp. 1314-1323, 1961.

17. KNESSL, C., MATKOWSKY, B.J., SCHUSS, Z., AND TIER, C., "Two parallel queues with dynamic routing," *IEEE Trans. Commun.*, vol. 34, pp. 1170-1175, 1986.

18. NELSON, R.D. AND PHILIPS, T.K., "An approximation to the response time for shortest queue routing," *Performance Evaluation Review*, vol. 17, pp. 181-189, 1989.

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

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

21. RAO, B.M. AND POSNER, M.J.M., "Algorithmic and approximation analysis of the shorter queue model," *Naval Res. Log.*, vol. 34, pp. 381-398, 1987.

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

List of COSOR-memoranda - 1992

Number	Month	Author	Title
92-01	January	F.W. Steutel	On the addition of log-convex functions and sequences
92-02	January	P. v.d. Laan	Selection constants for Uniform populations
92-03	February	E.E.M. v. Berkum H.N. Linssen D.A. Overdijk	Data reduction in statistical inference
92-04	February	H.J.C. Huijberts H. Nijmeijer	Strong dynamic input-output decoupling: from linearity to nonlinearity
92-05	March	S.J.L. v. Eijndhoven J.M. Soethoudt	Introduction to a behavioral approach of continuous-time systems
92-06	April	P.J. Zwietering E.H.L. Aarts J. Wessels	The minimal number of layers of a perceptron that sorts
92-07	April	F.P.A. Coolen	Maximum Imprecision Related to Intervals of Measures and Bayesian Inference with Conjugate Imprecise Prior Densities
92-08	May	I.J.B.F. Adan J. Wessels W.H.M. Zijm	A Note on "The effect of varying routing probability in two parallel queues with dynamic routing under a threshold-type scheduling"
92-09	May	I.J.B.F. Adan G.J.J.A.N. v. Houtum J. v.d. Wal	Upper and lower bounds for the waiting time in the symmetric shortest queue system