Operations Research Letters 13 (1993) 107–112
North-Holland

March 1993

Matrix-geometric analysis of the shortest queue problem with threshold jockeying

I.J.B.F. Adan

Eindhoven University of Technology, Department of Mathematics and Computing Science, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

J. Wessels

Eindhoven University of Technology, Department of Mathematics and Computing Science, P.O. Box 513, 5600 MB Eindhoven, The Netherlands, and International Institute for Applied Systems Analysis, A-2361 Laxenburg, Austria

W.H.M. Zijm

University of Twente, Department of Mechanical Engineering, P.O. Box 217, 7500 AE Enschede, The Netherlands

Received November 1991
Revised August 1992

In this paper we study a system consisting of c parallel servers with possibly different service rates. Jobs arrive according to a Poisson stream and generate an exponentially distributed workload. An arriving job joins the shortest queue, where in case of multiple shortest queues, one of these queues is selected according to some arbitrary probability distribution. If the maximum difference between the lengths of the c queues exceeds some threshold value T, then one job switches from the longest to the shortest queue, where in case of multiple longest queues, the queue loosing a job is selected according to some arbitrary probability distribution. It is shown that the matrix-geometric approach is very well suited to find the equilibrium probabilities of the queue lengths. The interesting point is that a proper choice for the state space partitioning depends on the aspect one is interested in. Using one partitioning of the state space an explicit ergodicity condition can be derived from Neuts' mean drift condition and using another partitioning the associated R-matrix can be determined explicitly. Moreover, both partitionings used are different from the one suggested by the conventional way of applying the matrix-geometric approach. Therefore, the paper can be seen as a plea for giving more attention to the question of the selection of a partitioning in the matrix-geometric approach.

jockeying; matrix-geometric solution; queues in parallel; shortest queue

1. Introduction

The matrix-geometric approach, as introduced by Neuts in his book [8], has proved to be a powerful tool for the analysis of Markov processes with large and complicated state spaces, particularly the ones that appear when modeling queueing or maintenance systems. One stage in the approach is a partitioning of the state space. Usually this stage does not get much attention. Users tend to use so-called natural partitionings which are suggested by the way of modeling or, if such a natural partitioning is not available, they select a partitioning that fits most elegantly with the boundary behavior

Correspondence to: I.J.B.F. Adan, Eindhoven University of Technology, Department of Mathematics and Computing Science, P.O. Box 513, 5600 MB Eindhoven, The Netherlands.

of the process. In a previous paper [1], the authors demonstrate already that in the case of the shortest queue problem with threshold jockeying for 2 parallel queues it is more effective to apply a partitioning based on the behavior of the process in the interior of the state space than based on the boundary behavior as has been proposed by Gertsbakh [4].

The present paper investigates the shortest queue problem with threshold jockeying for c parallel queues. The results in this case are even more striking, since, against the belief that the matrix-geometric approach is not useful in obtaining insight in the case $c > 2$ (cf. Zhao and Grassmann [10]), we show that, by using the right criterion for selecting a partitioning, two partitionings can be constructed that help in solving the equilibrium equations completely.

Actually, this paper has two goals. In the first place it presents a simple way of showing that the shortest queue problem with threshold jockeying for c parallel queues has a nice and elegant solution. In the second place it demonstrates that the use of a proper criterion for selecting a partitioning of the state space can be crucial for the success of the matrix-geometric approach.

Consider a queueing system consisting of c parallel servers. The service rate of server i is γ_i, $i = 1,\ldots,c$, where, for simplicity of notation, it is supposed that $\gamma_1 + \cdots + \gamma_c = c$. Jobs arrive according to a Poisson stream with rate $c\rho$ and generate an exponentially distributed workload with unit mean. An arriving job joins the shortest queue. In case of multiple shortest queues, one of these queues is selected according to some arbitrary probability distribution. If the maximal difference between the lengths of the c queues exceeds some threshold value T, then one job switches from the longest to the shortest queue. In case of multiple longest queues, the queue loosing a job is selected according to some arbitrary probability distribution. This system can be represented by a continuous time Markov process whose state space S consists of the vectors $(n_1, n_2, \ldots, n_c)$ where n_i is the length of queue i, $i = 1,\ldots,c$. Due to the threshold jockeying the state space S is restricted to those vectors for which $|n_i - n_j| \leq T$ for all i and j.

For $c = 2$ this model has been analyzed by Gertsbakh [4] and by Adan, Wessels and Zijm [1]. For arbitrary c Zhao and Grassmann [10] use the concept of modified lumpability of continuous time Markov chains to find the equilibrium distribution of the queue lengths. The special case of *instantaneous* jockeying ($T = 1$) has been analyzed by Haight [6] for $c = 2$ and by Disney and Mitchell [2], Elsayed and Bastani [3], Kao and Lin [7] and Grassmann and Zhao [5] for arbitrary c.

In this paper it is shown that the matrix-geometric approach developed by Neuts [8] is very well suited to analyze this problem. In Section 2 we show for a suitably chosen partitioning of the state space that an explicit ergodicity condition, which is obviously $\rho < 1$, can be derived from Neuts' mean drift condition. However, for that partitioning the associated R-matrix cannot be determined explicitly. Therefore, in Section 3 we propose another partitioning for which the associated R-matrix can easily be determined explicitly. Actually, the latter choice generalizes the choice used in [1] for $c = 2$, which was suggested by a more direct way of solving the equilibrium equations. Gertsbakh [4] uses the matrix-geometric approach for $c = 2$, but his choice for the partitioning does not lead to an explicit solution for the associated R-matrix.

2. Necessary and sufficient ergodicity condition

Application of the matrix-geometric approach requires a partitioning of the state space. First define for $l = 0, 1, \ldots,$ *sublevel* l as the set of states $(n_1,\ldots,n_c) \in S$ satisfying $n_1 + \cdots + n_c = l$. Then, for each state $(n_1,\ldots,n_c)$ at sublevel $l > (c-1)T$, none of the queues is empty and the transition rates from this state are identical to the rates from the corresponding state $(n_1 + 1,\ldots,n_c + 1)$ at sublevel $l + c$. This suggests to define for all $l = 0, 1,\ldots,$ *level* l as the union of the sublevels lc, $lc + 1,\ldots, lc + c - 1$ and to partition the state space S according to these levels with $l = T, T + 1,\ldots,$ and to put the levels $0, 1,\ldots, T - 1$ with less regular behavior into one set. The states at level l are ordered by sublevel, states from each sublevel being ordered lexicographically. For this partitioning the generator Q is of the

following form, where the first class corresponds to the group of levels $0,\dots,T-1$,

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

Corresponding to the partitioning of level $l \geq T$ into the sublevels lc, $lc+1,\dots,lc+c-1$, the square matrices A_0, A_1 and A_2 are of the form

$$A_1=\begin{pmatrix} A_{0,0} & A_{0,1} & 0 & 0 & \cdots & \cdots & 0 \\ A_{1,0} & A_{1,1} & A_{1,2} & 0 & & & \vdots \\ 0 & A_{2,1} & A_{2,2} & A_{2,3} & & & \\ \vdots & & \ddots & \ddots & \ddots & & \vdots \\ & & & A_{c-3,c-4} & A_{c-3,c-3} & A_{c-3,c-2} & 0 \\ \vdots & & & 0 & A_{c-2,c-3} & A_{c-2,c-2} & A_{c-2,c-1} \\ 0 & \cdots & \cdots & 0 & 0 & A_{c-1,c-2} & A_{c-1,c-1} \end{pmatrix}, \tag{1}$$

$$A_0=\begin{pmatrix} 0 & \cdots & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & & \vdots \\ A_{c-1,0} & 0 & \cdots & 0 \end{pmatrix}, \qquad A_2=\begin{pmatrix} 0 & \cdots & 0 & A_{0,c-1} \\ \vdots & & 0 & 0 \\ \vdots & & & \vdots \\ 0 & \cdots & \cdots & 0 \end{pmatrix}.$$

The Markov process Q is irreducible and, since two states at levels $> T$ can reach each other via paths not passing through levels $\leq T$, the generator $A_0+A_1+A_2$ is also irreducible. Thus Theorem 1.7.1 in Neuts' book [8] can readily be applied. Specifically the Markov process Q is ergodic if and only if

$$\pi A_0 e < \pi A_2 e, \tag{2}$$

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

$$\pi(A_0+A_1+A_2)=0, \qquad \pi e = 1. \tag{3}$$

By partitioning π as $(\pi_0,\dots,\pi_{c-1})$ and the column vector e as

$$e=\begin{pmatrix} e_0 \\ \vdots \\ e_{c-1} \end{pmatrix},$$

corresponding to the form (1) of A_0, A_1 and A_2, we get as a result that inequality (2) reduces to

$$\pi_{c-1}A_{c-1,0}e_0 < \pi_0 A_{0,c-1}e_{c-1} \tag{4}$$

and (3) to

$$\begin{aligned} &\pi_{c-1}A_{c-1,0}+\pi_0 A_{0,0}+\pi_1 A_{1,0}=0, \\ &\pi_{i-1}A_{i-1,i}+\pi_i A_{i,i}+\pi_{i+1}A_{i+1,i}=0, \qquad i=1,\dots,c-2, \\ &\pi_{c-2}A_{c-2,c-1}+\pi_{c-1}A_{c-1,c-1}+\pi_0 A_{0,c-1}=0. \end{aligned} \tag{5}$$

Since the flow from a state at sublevel $l>(c-1)T$ to sublevel $l+1$ is $c\rho$ (an arrival) and the flow to sublevel $l-1$ is c (a service completion) it follows that

$$\begin{aligned} &A_{0,1}e_1=c\rho e_0, \qquad A_{0,0}e_0=-c(\rho+1)e_0, \qquad A_{0,c-1}e_{c-1}=ce_0, \\ &A_{i,i+1}e_{i+1}=c\rho e_i, \qquad A_{i,i}e_i=-c(\rho+1)e_i, \qquad A_{i,i-1}e_{i-1}=ce_i, \ i=1,\dots,c-2. \\ &A_{c-1,0}e_0=c\rho e_{c-1}, \qquad A_{c-1,c-1}e_{c-1}=-c(\rho+1)e_{c-1}, \qquad A_{c-1,c-2}e_{c-2}=ce_{c-1}. \end{aligned}$$

Hence inequality (4) simplifies to

$$c\rho\pi_{c-1}e_{c-1} < c\pi_0 e_0 \tag{6}$$

and multiplying the set (5) with e_i leads to

$$\begin{aligned}
&c\rho\pi_{c-1}e_{c-1} - \rho(c+1)\pi_0 e_0 + c\pi_1 e_1 = 0,\\
&c\rho\pi_{i-1}e_{i-1} - \rho(c+1)\pi_i e_i + c\pi_{i+1}e_{i+1} = 0, \qquad i = 1,\ldots,c-2,\\
&c\rho\pi_{c-2}e_{c-2} - \rho(c+1)\pi_{c-1}e_{c-1} + c\pi_0 e_0 = 0.
\end{aligned}$$

By the symmetry of these equations it follows that $\pi_0 e_0 = \cdots = \pi_{c-1}e_{c-1}$ and thus from (6) we can conclude that:

Theorem 1. *The Markov process Q is ergodic if and only if $\rho < 1$.*

So, for the chosen partitioning of the state space, the mean drift condition (2) easily leads to the desired ergodicity condition. The associated R-matrix, however, cannot be determined explicitly. In the next section we adapt the definition of the levels and show for this new partitioning that the associated R-matrix can be determined explicitly.

3. Explicit determination of R

We now adapt the definition of level l in such a way that level l is the set of states $(n_1,\ldots,n_c) \in S$ satisfying $\max(n_1,\ldots,n_c) = l$. The state space S is partitioned into the sequence of levels T, $T+1$, ... The levels $0, 1,\ldots,T-1$ with less regular behavior are put together in one set. The states at each level are ordered lexicographically. For this partitioning the generator Q is of the form

$$Q = \begin{pmatrix}
C_{00} & C_{01} & 0 & 0 & 0 & \cdots\\
C_{10} & C_{11} & D_0 & 0 & 0 & \cdots\\
0 & D_2 & D_1 & D_0 & 0 & \cdots\\
0 & 0 & D_2 & D_1 & D_0 & \cdots\\
\vdots & \vdots & \vdots & \ddots & \ddots & \ddots
\end{pmatrix}. \tag{7}$$

The square matrices D_0, D_1 and D_2 are of dimensions $m \times m$ where m is the number of states at a level $\geq T$. The Markov process Q is irreducible and, since two states at levels $> T$ can reach each other via paths not passing through levels $\leq T$, the generator $D_0 + D_1 + D_2$ is also irreducible. Thus Theorem 1.7.1 in Neuts' book [8] can again be applied. By partitioning the equilibrium probability vector $\boldsymbol{p}$ into the vector $(\boldsymbol{p}_0,\ldots,\boldsymbol{p}_{T-1})$ and into the sequence of vectors $\boldsymbol{p}_T$, $\boldsymbol{p}_{T+1},\ldots$, where $\boldsymbol{p}_l$ is the equilibrium probability vector of level l, we then obtain

$$\boldsymbol{p}_l = \boldsymbol{p}_T R^{l-T},\ l > T,$$

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

$$D_0 + RD_1 + R^2 D_2 = 0. \tag{8}$$

The matrix D_0 has a special structure. Since it is only possible to jump from level l to level $l+1$ via state $(l, l,\ldots,l)$, it follows that all rows of D_0 are zero, except for the last row. Thus D_0 is of the form

$$D_0 = \begin{pmatrix} 0 \\ \boldsymbol{v} \end{pmatrix} \quad \text{where } \boldsymbol{v} = (v_0,\ldots,v_{m-1}). \tag{9}$$

This special matrix structure of D_0 can be exploited to determine R explicitly. In fact, in the more general case that Q has the form (7) and D_0 is given by $D_0 = \boldsymbol{x}\cdot\boldsymbol{y}$ where $\boldsymbol{x}$ is a column vector and $\boldsymbol{y}$ is a row vector, Ramaswami and Latouche [9] show that R can be determined explicitly once the maximal

eigenvalue η of R is known. Hence, in order to apply their result, it is necessary to compute η without having first computed R. In [9] two algorithms are presented. However, below it is demonstrated that in the present model it is possible to derive an explicit expression for η, and thereby, also one for R.

Since rows in R which correspond to zero rows in D_0, are also zero (see e.g. the proof of Theorem 1.3.4 in [8]), we conclude from (9) that R is also of the form

$$R = \begin{pmatrix} 0 \\ \boldsymbol{w} \end{pmatrix} \quad \text{where } \boldsymbol{w} = (w_0, \ldots, w_{m-1}). \tag{10}$$

Hence, denoting the probability for $(n_1, \ldots, n_c)$ by $p(n_1, \ldots, n_c)$, the matrix-geometric solution simplifies to

$$\boldsymbol{p}_l = p(T,T,\ldots,T)\boldsymbol{w}_{m-1}^{l-T-1}\boldsymbol{w}, \quad l > T. \tag{11}$$

From (10) it is easily seen that $\eta = w_{m-1}$. The component w_{m-1} can be found by using the following balance argument. Let V_l be the set of states $(n_1, \ldots, n_c) \in S$ satisfying $n_1 + \cdots + n_c = l$ 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 $l > (c-1)T$

$$P(V_{l+1})c = P(V_l)c\rho,$$

and by applying this relation c times, we get

$$P(V_{l+c}) = \rho^c P(V_l). \tag{12}$$

On the other hand, (11) implies that $p(n_1 + 1, \ldots, n_c + 1) = w_{m-1}p(n_1, \ldots, n_c)$ if $\max(n_1, \ldots, n_c) > T$, so it follows that for $l > cT$

$$P(V_{l+c}) = w_{m-1}P(V_l). \tag{13}$$

Combining (12) and (13) yields

$$w_{m-1} = \rho^c.$$

The remaining components of $\boldsymbol{w}$ are solved from equation (8), which, by insertion of the special forms of R and D_0, simplifies to

$$\boldsymbol{v} + \boldsymbol{w}(D_1 + w_{m-1}D_2) = 0. \tag{14}$$

Finally, substitution of $w_{m-1} = \rho^c$ into (14) leads to

$$\boldsymbol{w} = -\boldsymbol{v}(D_1 + \rho^c D_2)^{-1}$$

where the inverse of $D_1 + \rho^c D_2$ exists, since this matrix can be interpreted as a transient generator (escape is possible at least from the last state). These findings are summarized in the following theorem:

Theorem 2. *$R = \binom{0}{w}$ where $\boldsymbol{w} = -\boldsymbol{v}(D_1 + \rho^c D_2)^{-1}$.*

Given the solution (11) on the levels $T+1$, $T+2, \ldots$, the probability vectors $\boldsymbol{p}_T$, $\boldsymbol{p}_{T-1}, \ldots, \boldsymbol{p}_0$ 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 $(l, l, \ldots, l)$, it readily follows that these vectors can be solved recursively from the equilibrium equations for the levels $T, T-1, \ldots, 0$.

References

[1] I.J.B.F. Adan, J. Wessels and W.H.M. Zijm, "Analysis of the asymmetric shortest queue problem with threshold jockeying", *Stochastic Models* 7, 615–627 (1991).

[2] R.L. Disney and W.E. Mitchell, "A solution for queues with instantaneous jockeying and other customer selection rules", *Naval Res. Log.* **17**, 315–325 (1971).

[3] E.A. Elsayed and A. Bastani, "General solutions of the jockeying problem", *European J. Oper. Res.* **22**, 387–396 (1985).

[4] I. Gertsbakh, "The shorter queue problem: A numerical study using the matrix geometric solution", *European J. Oper. Res.* **15**, 374–381 (1984).
[5] W.K. Grassmann and Y. Zhao, "The shortest queue model with jockeying", *Naval Res. Log.* **37**, 773–787 (1990).
[6] F.A. Haight, "Two queues in parallel", *Biometrica* **45**, 401–410 (1958).
[7] E.P.C. Kao and C. Lin, "A matrix-geometric solution of the jockeying problem", *European J. Oper. Res.* **44**, 67–74 (1990).
[8] M.F. Neuts, *Matrix-geometric solutions in stochastic models*, Johns Hopkins University Press, Baltimore, MD 1981.
[9] V. Ramaswami and G. Latouche, "A general class of Markov processes with explicit matrix-geometric solutions", *OR Spektrum* **8**, 209–218 (1986).
[10] Y. Zhao and W.K. Grassmann, "Solving a parallel queueing model by using modified lumpability", Research paper, Queen's University, Dep. of Math. and Stat., 1991.