Matrix geometric analysis of the shortest queue problem with threshold jockeying

Citation for published version (APA):
Memorandum COSOR 91-24
Matrix-geometric analysis of the shortest queue problem with threshold jockeying
I.J.B.F. Adan
J. Wessels
W.H.M. Zijm

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

Eindhoven, October 1991
The Netherlands
MATRIX-GEOMETRIC ANALYSIS OF THE SHORTEST QUEUE PROBLEM WITH THRESHOLD JOCKEYING

I.J.B.F. Adan\textsuperscript{1),} J. Wessels\textsuperscript{1,2),} W.H.M. Zijm\textsuperscript{3)"

Abstract: 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 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, 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 for one partitioning of the state space an explicit ergodicity condition can be derived from Neuts' mean drift condition, whereas for 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.

Key Words: jockeying, matrix-geometric solution, queues in parallel, shortest queue.

1. Introduction

The matrix-geometric approach, as introduced by Neuts in his book \cite{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 behaviour of the process. In a previous paper \cite{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 behaviour of the process in the

\textsuperscript{1) Eindhoven University of Technology, Department of Mathematics and Computer Science, p.o. Box 513, 5600 MB - Eindhoven, The Netherlands.}
\textsuperscript{2) International Institute for Applied Systems Analysis, A-2361 Laxenburg, Austria.}
\textsuperscript{3) University of Twente, Department of Mechanical Engineering, p.o. Box 217, 7500 AE - Enschede, The Netherlands.}
interior of the state space than based on the boundary behaviour 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 provides any 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 \( \gamma_i \), \( i = 1, \ldots, c \), where, for simplicity of notation, it is supposed that \( \gamma_1 + \ldots + \gamma_c = c \). Jobs arrive according to a Poisson stream with rate \( cp \) 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 analyse 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 \( l = 0, 1, \ldots \) the set of states \( (n_1, \ldots, n_c) \in S \) satisfying \( n_1 + \ldots + n_c = l \). 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 \( l = 0, 1, \ldots \) the set of states \( (n_1, \ldots, n_c) \) at sublevel \( l \) 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 behaviour 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 \( 0, ..., T-1 \),

\[
Q = \begin{bmatrix}
B_{00} & B_{01} & 0 & 0 & 0 & \cdots \\
B_{10} & A_1 & A_0 & 0 & 0 & \cdots \\
0 & A_2 & A_1 & A_0 & 0 & \cdots \\
0 & 0 & A_2 & A_1 & A_0 & \cdots \\
\vdots & \vdots & \vdots & \ddots & \ddots & \ddots \\
\end{bmatrix}
\]

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

\[
A_1 = \begin{bmatrix}
A_{0,0} & A_{0,1} & 0 & 0 & \cdots & \cdots & 0 \\
A_{1,0} & A_{1,1} & A_{1,2} & 0 & \cdots & \vdots & \vdots \\
0 & A_{2,1} & A_{2,2} & A_{2,3} & \ddots & \ddots & \ddots \\
\vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \ddots \\
A_{c-3,c-4} & A_{c-3,c-3} & A_{c-3,c-2} & 0 & \cdots & \cdots & \cdots \\
0 & A_{c-2,c-2} & A_{c-2,c-1} & A_{c-2,c-1} & \cdots & \cdots & \cdots \\
0 & 0 & \cdots & \cdots & 0 & A_{c-1,c-2} & A_{c-1,c-1} \\
\end{bmatrix}
\]

\[
A_0 = \begin{bmatrix}
0 & \cdots & \cdots & \cdots & 0 \\
0 & 0 & \cdots & \cdots & 0 \\
A_{-1,0} & 0 & \cdots & \cdots & 0 \\
\end{bmatrix}, \quad A_2 = \begin{bmatrix}
0 & \cdots & 0 & A_{0,c-1} \\
\vdots & \ddots & \ddots & \ddots \\
0 & 0 & 0 & 0 \\
\end{bmatrix}
\]

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

\[
\pi A_0 \pi < \pi A_2 \pi , \quad (2)
\]
where \( e \) is the column vector of ones and \( \pi \) is the solution of
\[
\pi(A_0 + A_1 + A_2) = 0, \quad \pi e = 1.
\]  
By partitioning \( \pi \) as \((\pi_0, \ldots, \pi_{c-1})\) and the column vector \( e \) as
\[
e = \begin{bmatrix}
\pi_0 \\
\vdots \\
\pi_{c-1}
\end{bmatrix},
\]
corresponding to the form (1) of \( A_0, A_1 \) and \( A_2 \), we get as a result that inequality (2) reduces to
\[
\pi_{c-1}A_{c-1}e_0 < \pi_0 A_{0,c-1}e_0
\]  
and the equations (3) to
\[
\begin{align*}
\pi_{c-1}A_{c-1}e_0 + \pi_0 A_{0,0} + \pi_1 A_{1,0} &= 0, \\
\pi_{i-1}A_{i-1}e_i + \pi_i A_{i,i} + \pi_{i+1}A_{i+1,i} &= 0, \quad i = 1, \ldots, c-2, \\
\pi_{c-2}A_{c-2,c-1} + \pi_{c-1} A_{c-1,c-1} + \pi_0 A_{0,c-1} &= 0.
\end{align*}
\]  
Since the flow from a state at sublevel \( i > (c-1)T \) to sublevel \( i+1 \) is \( cp \) (an arrival) and the flow to sublevel \( i-1 \) is \( c \) (a service completion) it follows that
\[
A_{0,0}e_0 = -c(\rho + 1)e_0, \quad A_{0,c-1}e_{c-1} = ce_0,
\]
\[
A_{i,i+1}e_{i+1} = cpe_i, \quad A_{i,i}e_i = -c(\rho + 1)e_i, \quad A_{i,i-1}e_{i-1} = ce_i, \quad i = 1, \ldots, c-2,
\]
\[
A_{c-1,0}e_0 = cpe_{c-1}, \quad A_{c-1,c}e_{c-1} = -c(\rho + 1)e_{c-1}, \quad A_{c-1,c-2}e_{c-2} = ce_{c-1}.
\]
Hence inequality (4) simplifies to
\[
cp\pi_{c-1}e_{c-1} < c\pi_0 e_0
\]  
and multiplying the set of equations (5) with \( e_i \) leads to
\[

cp\pi_{c-1}e_{c-1} - \rho(c + 1)\pi_0 e_0 + c\pi_1 e_1 = 0,
\]
\[

cp\pi_{i-1}e_{i-1} - \rho(c + 1)\pi_i e_i + c\pi_{i+1} e_{i+1} = 0, \quad i = 1, \ldots, c-2,
\]
\[

cp\pi_{c-2}e_{c-2} - \rho(c + 1)\pi_{c-1} e_{c-1} + c\pi_0 e_0 = 0.
\]
By the symmetry of these equations it follows that \( \pi_0 e_0 = \ldots = \pi_{c-1} e_{c-1} \) 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, but 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 $l$ as the set of states $(n_1, ..., n_c) \in S$ satisfying $\max(n_1, ..., n_c) = l$. The state space $S$ is partitioned into the sequence of levels $T, T+1, ...$. The levels $0, 1, ..., T-1$ with less regular behaviour are put together in one set. The states at each level are ordered lexicographically. For this partitioning the generator $Q$ is of the form

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

The square matrices $D_0, D_1, D_2$ 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 $p$ into the vector $(p_0, ..., p_{T-1})$ and into the sequence of vectors $p_T, p_{T+1}, ...$ where $p_l$ is the equilibrium probability vector of level $l$, we then obtain

$$p_l = p_TR^{l-T}, \quad l > T,$$

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

$$D_0 + RD_1 + R^2D_2 = 0. \quad (7)$$

The matrix $R$ can be determined explicitly due to the special matrix structure of $D_0$. Since it is only possible to jump from level $l$ to level $l+1$ via state $(l, l, ..., l)$, it follows that all rows of $D_0$ are zero, except for the last row. Thus $D_0$ is of the form

$$D_0 = \begin{bmatrix} 0 \\ v \end{bmatrix} \quad \text{where} \quad v = (v_0, ..., v_{m-1}).$$

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 that $R$ is also of the form

$$R = \begin{bmatrix} 0 \\ w \end{bmatrix} \quad \text{where} \quad w = (w_0, ..., w_{m-1}).$$

Hence the matrix-geometric solution simplifies to

$$p_l = p_T w_T w_{l-T-1}, \quad l > T. \quad (8)$$
Let $V_l$ be the set of states $(n_1, \ldots, n_c) \in S$ satisfying $n_1 + \ldots + n_c = l$. Then the component $w_{m-1}$ is determined by balancing the flow between $V_l$ and $V_{l+1}$, yielding for $l > (c-1)T$

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

and by applying this relation $c$ times,

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

(9)

On the other hand, the set $V_{cl}$ is a subset of the union of the levels $l, l+1, \ldots, l+T-1$, so it follows from (8) that for $l > T$

$$P(V_{cl}) = K w_{m-1},$$

for some constant $K$ being independent of $l$. Combining (9) and (10) yields

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

The remaining components of $w$ are solved from equation (7), which, by insertion of the special forms of $R$ and $D_0$, simplifies to

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

Finally, substituting $w_{m-1} = \rho^c$ leads to

$$w = -v(D_1 + \rho^cD_2)^{-1},$$

where the inverse of $D_1 + \rho^cD_2$ exists, since this matrix can be interpreted as a transient generator (escape is possible at least from the last state).

**Theorem 2**: $R = \begin{bmatrix} 0 \\ w \end{bmatrix}$ where $w = -v(D_1 + \rho^cD_2)^{-1}$.

**Remark**: The explicit solution of $R$ is mainly due to the special matrix structure of $D_0$. In fact, Ramaswami and Latouche [9] show that if the generator $Q$ can be partitioned as in the beginning of this section and $D_0$ is given by $D_0 = x'y$ where $x$ is a column vector and $y$ is a row vector, then $R$ is explicitly determined, once its maximal eigenvalue is known.

4. **Conclusion**

In this paper we studied the shortest queue problem with $c$ servers and threshold jockeying and we showed that the matrix-geometric approach is very well suited to analyse this problem. The interesting point was that a proper choice for the state space partitioning depends on where one is interested in. One partitioning leads to an explicit ergodicity condition and for another partitioning the associated $R$-matrix is determined explicitly. Another conclusion may be that, for the application of the matrix-geometric approach for solving the equilibrium equations of Markov processes, the aspect of selecting a proper partitioning is of crucial importance. In the
treated case the extension of Gertsbakh's partitioning to $c > 2$ does not provide any insight. However, two other partitionings, based on the process behavior in internal points, lead to a complete solution of the problem.

References


<table>
<thead>
<tr>
<th>Number</th>
<th>Month</th>
<th>Author(s)</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>91-01</td>
<td>January</td>
<td>M.W.I. van Kraaij, W.Z. Venema, J. Wessels</td>
<td>The construction of a strategy for manpower planning problems.</td>
</tr>
<tr>
<td>91-03</td>
<td>January</td>
<td>M.W.P. Savelsbergh</td>
<td>The vehicle routing problem with time windows: minimizing route duration.</td>
</tr>
<tr>
<td>91-04</td>
<td>January</td>
<td>M.W.I. van Kraaij</td>
<td>Some considerations concerning the problem interpreter of the new manpower planning system formasy.</td>
</tr>
<tr>
<td>91-06</td>
<td>March</td>
<td>R.J.G. Wilms</td>
<td>Properties of Fourier-Stieltjes sequences of distribution with support in [0,1).</td>
</tr>
<tr>
<td>91-07</td>
<td>March</td>
<td>F. Coolen, R. Dekker, A. Smit</td>
<td>Analysis of a two-phase inspection model with competing risks.</td>
</tr>
<tr>
<td>91-08</td>
<td>April</td>
<td>P.J. Zwietering, E.H.L. Aarts, J. Wessels</td>
<td>The Design and Complexity of Exact Multi-Layered Perceptrons.</td>
</tr>
<tr>
<td>91-09</td>
<td>May</td>
<td>P.J. Zwietering, E.H.L. Aarts, J. Wessels</td>
<td>The Classification Capabilities of Exact Two-Layered Perceptrons.</td>
</tr>
<tr>
<td>91-10</td>
<td>May</td>
<td>P.J. Zwietering, E.H.L. Aarts, J. Wessels</td>
<td>Sorting With A Neural Net.</td>
</tr>
<tr>
<td>91-11</td>
<td>May</td>
<td>F. Coolen</td>
<td>On some misconceptions about subjective probability and Bayesian inference.</td>
</tr>
<tr>
<td>Number</td>
<td>Month</td>
<td>Authors</td>
<td>Title</td>
</tr>
<tr>
<td>--------</td>
<td>---------</td>
<td>-------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>91-12</td>
<td>May</td>
<td>P. van der Laan</td>
<td>Two-stage selection procedures with attention to screening.</td>
</tr>
<tr>
<td>91-14</td>
<td>June</td>
<td>J. Korst, E. Aarts, J.K. Lenstra, J. Wessels</td>
<td>Periodic assignment and graph colouring.</td>
</tr>
<tr>
<td>91-16</td>
<td>July</td>
<td>P. Deheuvels, J.H.J. Einmahl</td>
<td>Approximations and Two-Sample Tests Based on P - P and Q - Q Plots of the Kaplan-Meier Estimators of Lifetime Distributions.</td>
</tr>
<tr>
<td>91-19</td>
<td>August</td>
<td>P. van der Laan</td>
<td>The efficiency of subset selection of an almost best treatment.</td>
</tr>
<tr>
<td>91-20</td>
<td>September</td>
<td>P. van der Laan</td>
<td>Subset selection for an -best population: efficiency results.</td>
</tr>
<tr>
<td>91-22</td>
<td>September</td>
<td>R.J.M. Vaessens, E.H.L. Aarts, J.H. van Lint</td>
<td>Genetic Algorithms in Coding Theory - A Table for ( A_t(n,d) ).</td>
</tr>
<tr>
<td>91-23</td>
<td>September</td>
<td>P. van der Laan</td>
<td>Distribution theory for selection from logistic populations.</td>
</tr>
<tr>
<td>91-24</td>
<td>October</td>
<td>I.J.B.F. Adan</td>
<td>Matrix-geometric analysis of the shortest queue problem with threshold jockeying.</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td></td>
<td></td>
<td>J. Wessels</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>W.H.M. Zijm</td>
<td></td>
</tr>
</tbody>
</table>