BOUNDS FOR PERFORMANCE CHARACTERISTICS; A SYSTEMATIC APPROACH VIA COST STRUCTURES

G.J. van Houtum, W.H.M. Zijm
University of Twente,
Department of Mechanical Engineering,
Box 217, 7500 AE - Enschede,
the Netherlands,
email: {g.j.j.a.n.vanhoutum,w.h.m.zijm}@wb.utwente.nl

I.J.B.F. Adan, J. Wessels
Eindhoven University of Technology,
Department of Mathematics and Computer Science,
Box 513, 5600 MB - Eindhoven,
the Netherlands,
email: {iadan,wessels}@win.tue.nl

ABSTRACT

In this paper we present a systematic approach for the construction of bounds for the average cost in Markov chains with infinitely many states. The technique to prove the bounds is based on dynamic programming. Most performance characteristics of Markovian systems can be represented by the average cost for some appropriately chosen cost structure. Therefore, the approach can be used to generate bounds for relevant performance characteristics. The approach is demonstrated for the shortest queue model. It is shown how for this model several bounds for the mean waiting time can be constructed. We include numerical results to demonstrate the quality of these bounds.

1 INTRODUCTION

In this paper we consider an irreducible N-dimensional Markov chain with states $\boldsymbol{m} = (m_1, \cdots, m_N)$, where each m_i is an integer, and transition probabilities $p(\boldsymbol{m}, \boldsymbol{n})$. Let π denote its equilibrium distribution (which we assume to exist). If $c(\boldsymbol{m})$ is the cost per period in state $\boldsymbol{m}$, then the average cost g is given by

$$g = \sum_{\boldsymbol{m}} c(\boldsymbol{m})\pi(\boldsymbol{m}).$$

To determine g, we need the distribution π, which in many cases is difficult to obtain exactly. Therefore we try to develop a systematic approach for the construction of bounds for g which can easily and efficiently be computed.

The approach tries to construct modifications of the original model which produce bounds for g. And of course these modifications should be easier to handle than the original model. The main contribution of the paper is that it is shown how such modifications may be found systematically: one should first try to identify precedences (with respect to cost) between states. Based on these precedences it appears to be easy to produce suitable modifications. In fact, the approach originates from earlier work in [19, 21, 20, 23, 24, 3, 26, 4], and attempts to unify the techniques used in these references.

The technique used in this paper to establish computable bounds, is also a powerful tool to prove qualitative properties, like e.g. monotonicity properties in queueing networks (cf. [19]) or optimality of routing policies to parallel queues (cf. [12]).

Many performance characteristics of Markovian systems, such as, e.g., mean queue lengths and mean waiting times in queueing problems, can be represented by the average cost for some appropriately chosen cost structure. Hence, the approach can be used to generate bounds for relevant performance characteristics.

We will demonstrate the approach for the shortest queue problem. To keep the presentation simple, we only consider the system with two queues. But, in fact, the power of the approach is that it also works well for more than two queues, because in that case there is no analytical solution available. The problem with two queues has been extensively studied in the literature. Exact analytical results can be found in [13, 7, 1]. There are also many papers analyzing approximations for the shortest queue problem, see [6, 8, 9, 10, 11, 15, 16, 18, 27]. It appears that the present approach leads to several models producing bounds for performance characteristics such as, e.g., the mean waiting time.

These models cover the ones in [6, 9, 10, 18, 27]. None of these references, however, rigorously proves that these models indeed produce bounds. But this is done in the present paper. An important property of the bounds presented here is that their quality can be controlled by some threshold (or truncation) parameter. The larger this parameter, the more accurate the bounds will be, but also the more effort it takes to compute them.

The paper is organized as follows. In the next section, we describe the shortest queue model. This model will be used throughout the paper to demonstrate the concepts and techniques. In Section 3, we introduce a modification of the original Markov chain, and subsequently we compare the average costs of the modified and original chain in Section 4. Section 5 deals with the proof of precedences. In Section 6, we present numerical results to demonstrate the quality of the bounds produced for the mean waiting time of the shortest queue model. Finally, Section 7 is devoted to conclusions.

2 BASIC EXAMPLE: SHORTEST QUEUE MODEL

The shortest queue model is characterized as follows. There are two identical parallel servers, each with its own queue (see Figure 1). The service times are exponential with rate μ. Jobs arrive according to a Poisson stream with rate λ and join the shortest queue. This system can be described by a Markov process with states $\boldsymbol{m} = (m_1, m_2)$ where m_1 and m_2 are the length of the shortest and longest queue, resp. (so $m_1 \leq m_2$). Without loss of generality we may take $\lambda + 2\mu = 1$ and assume that the servers always work (also when there is no job). But, service completion only leads to a departure if there is a job in the queue, otherwise it is fake (and, fake jobs will be interrupted as soon as a real job arrives). The artificial assumption of working on fake jobs implies that in each state the outgoing transition rates add up to 1 (i.e., we uniformized the Markov process). The flow diagram is shown in Figure 2. As cost rate we take the number of waiting jobs, so

$$c(\boldsymbol{m}) = \max\{m_1 - 1, 0\} + \max\{m_2 - 1, 0\}. \tag{1}$$

Then the average cost yields the mean number of waiting jobs in the system, and by Little's law, the mean *normalized* waiting time W. Here, the normalized waiting time is defined as the ratio of the waiting time and the mean service time ($= 1/\mu$).

The process observed at jumps is a Markov chain, and since the mean time between jumps is always 1, it has the same equilibrium distribution as the original Markov process. If we take $c(\boldsymbol{m})$ as cost per period, then it also has the same average cost. From now

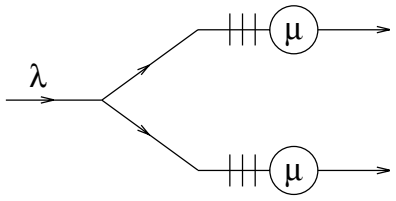

Figure 1: The shortest queue model

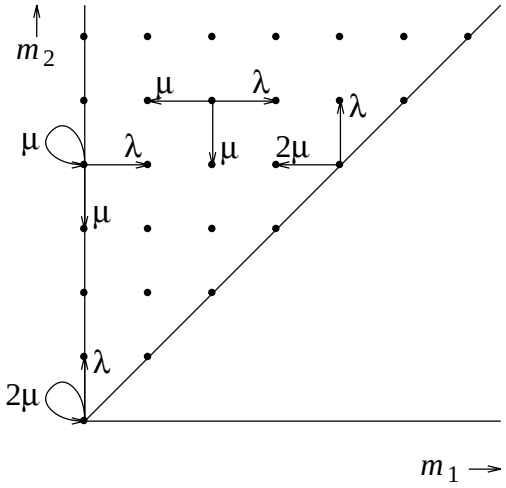

Figure 2: Flow diagram for the shortest queue model

on we only consider the Markov chain (at jumps) instead of the continuous-time Markov process. Note that the rates in Figure 2 correspond to transition probabilities in the jump process.

3 THE MODIFIED MODEL

We consider the following modification of the original model (Markov chain) introduced in Section 1. In some states $\boldsymbol{m}$ we redirect one or more outgoing transitions. This means that a transition from $\boldsymbol{m}$ to $\boldsymbol{n}$, say, is redirected to another state $\tilde{\boldsymbol{n}}$. Then the new transition probability to $\boldsymbol{n}$ is zero and to $\tilde{\boldsymbol{n}}$ it equals $p(\boldsymbol{m},\boldsymbol{n}) + p(\boldsymbol{m},\tilde{\boldsymbol{n}})$. Denote the new transition probabilities by $\tilde{p}(\boldsymbol{m},\boldsymbol{n})$. The costs per state are not altered. We assume that the modified chain is unichained (some states may now be transient) and that its equilibrium distribution exists. The average cost is denoted by $\tilde{g}$.

Of course, there are many possibilities to modify the system. Which transitions should be redirected and how? When does it lead to an upper or lower bound for g? How to prove this? In the next sections we will attempt to answer these questions.

Example 3.1 *(Modifications of the shortest queue model)*
Below we first describe some modifications of the shortest queue model. In each of these modifications we use a threshold parameter T, which is some fixed, but arbitrary positive integer. The modifications (for $T = 3$) are depicted in Figure 3.

Finite Buffers (FB): The simplest modification is obtained by redirecting only one transition in only one state, namely by redirecting the transition from (T, T) to $(T, T+1)$ (an arrival) to state (T, T) (reject the new job). Note that states with m_1 or $m_2 > T$ are now transient. This model corresponds to the situation where both servers have a finite buffer of size T. It has been analyzed e.g. in [6, 10].

Central Buffer (CB): For all states $(m_1, m_1 + 1)$ with $m_1 \geq T$ the transition to $(m_1 - 1, m_1 + 1)$ (a departure) is redirected to (m_1, m_1). This model has the following interpretation. Both servers have a finite local buffer of size T, and there is a central buffer with infinite capacity. On arrival a job is sent to the central buffer if there is no room in the local buffers. As soon as there is room again a job is released from the central (if there is any) to the local buffer. Hence, state $(T + 2, T + 2)$ means that both local buffers are full and 2 jobs are waiting in the central buffer.

Threshold Jockeying (TJ): For all states $(m_1, m_1 + T)$ with $m_1 > 0$ the transition to $(m_1 - 1, m_1 + T)$ (a departure in the shortest queue) is redirected to $(m_1, m_1 + T - 1)$. This means that a job switches from the longest to the shortest queue as soon as the difference between the queue lengths exceeds T. For an analysis of this model we refer to [2, 9, 27].

One Infinite Buffer (OIB): For all states (T, m_2) with $m_2 > T$ the transition to $(T+1, m_2)$ is redirected to (T, m_2+1). This model corresponds to the situation where one server has a finite buffer of size T and the other server has an infinite one. On arrival a job joins the shortest queue if there is room, and otherwise the longest one (in the infinite buffer). A matrix-geometric analysis of this model can be found in [18].

Threshold Killing (TK): For all states $(m_1, m_1 + T)$ with $m_1 > 0$ the transition to $(m_1 - 1, m_1 + T)$ is redirected to $(m_1 - 1, m_1 + T - 1)$. So when the difference in queue lengths exceeds T due to departure in the shortest queue, then the job in service in the longest queue is directly killed (and removed).

Threshold Blocking (TB): For all states $(m_1, m_1 + T)$ with $m_1 > 0$ the transition to $(m_1 - 1, m_1 + T)$ is redirected to $(m_1, m_1 + T)$. This means that when a job is completed in the shortest queue and its departure would lead to a difference in queue lengths greater

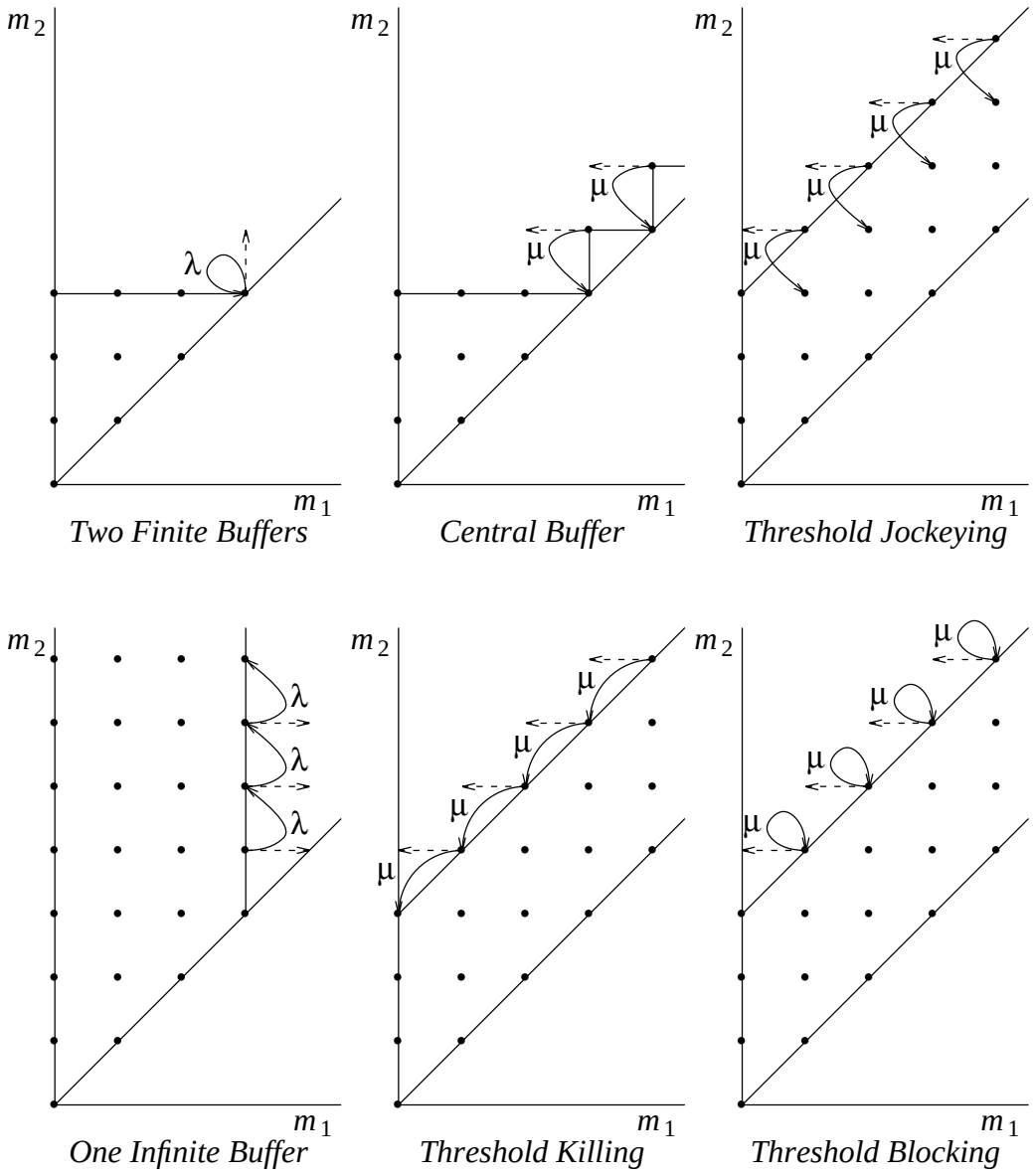

Figure 3: Flow diagrams for modifications of the shortest queue model

than T, then its departure is blocked and the job is served once more (with a new service time).

For each of the modifications described above, the resulting state space of recurrent states is finite (the FB model) or infinite in at most one direction (the other models). Therefore these modifications are much easier to handle than the original model. In fact, the FB model can be solved by a standard numerical procedure and the other ones can be efficiently solved using the matrix-geometric approach of Neuts [17] (e.g., by applying the algorithm in [14]). It may be intuitively clear which modifications lead to an upper bound and which ones to a lower bound for the mean normalized waiting time. In the next section, we present a technique to prove this and we develop a systematic approach to construct such modifications leading to bounds.

Remark that the central buffer and the three threshold models exploit the property that most of the probability mass in the shortest queue model is concentrated around the diagonal of the state space. So one might expect that these models produce tight bounds for already small values of the threshold T. □

4 COMPARISON OF AVERAGE COSTS

We now return to the general model. Suppose we want to show that $\tilde{g} \leq g$. To do so we study the expected cost over a finite number of periods. Define $u_n(\boldsymbol{m})$ and $v_n(\boldsymbol{m})$ as the expected cost over n periods for the modified and original model, resp., when starting in $\boldsymbol{m}$. Defining $u_0 = v_0 = 0$ we try to prove by induction that for all n and all recurrent states $\boldsymbol{m}$ in the modified model

$$u_n(\boldsymbol{m}) \leq v_n(\boldsymbol{m}). \tag{2}$$

Then it follows that the average costs are ordered in the same way. To establish (2) we first need *precedences* between states. We say that state $\boldsymbol{m}$ has precedence over, or is more attractive than state $\boldsymbol{n}$, if $\boldsymbol{m}$ and $\boldsymbol{n}$ satisfy the following *precedence relation*:

$$v_n(\boldsymbol{m}) \leq v_n(\boldsymbol{n}) \qquad \text{for all } n \geq 0. \tag{3}$$

In words, starting in $\boldsymbol{m}$ yields lower total expected cost than in $\boldsymbol{n}$. Now the first, and crucial step is the characterization of a set P of pairs $(\boldsymbol{m}, \boldsymbol{n})$ satisfying (3). These pairs are called *precedence pairs*. Usually precedence pairs are intuitively obvious, and of course, they depend on the one-period cost c (set $n = 1$ in (3)). The proof of these pairs is the

topic of the next section. Once a sufficiently rich set P has been characterized, the proof of (2) is easy as will be shown below.

The proof of (2) follows by induction. For $n = 0$ inequality (2) trivially holds. Assuming (2) holds for n we try to prove it for $n+1$. The expected cost over $n+1$ periods satisfies

$$v_{n+1}(\boldsymbol{m}) = c(\boldsymbol{m}) + \sum_{\boldsymbol{n}} p(\boldsymbol{m}, \boldsymbol{n}) v_n(\boldsymbol{n}).$$

It follows that $v_{n+1}(\boldsymbol{m}) \geq u_{n+1}(\boldsymbol{m})$ provided we have constructed the modified model by redirecting transitions to more attractive states (i.e. a transition to $\boldsymbol{n}$ is redirected to $\tilde{\boldsymbol{n}}$ only if $(\boldsymbol{n}, \tilde{\boldsymbol{n}}) \in P$). Namely, then we have

$$\begin{aligned} v_{n+1}(\boldsymbol{m}) &\geq c(\boldsymbol{m}) + \sum_{\boldsymbol{n}} \tilde{p}(\boldsymbol{m}, \boldsymbol{n}) v_n(\boldsymbol{n}) \\ &\geq c(\boldsymbol{m}) + \sum_{\boldsymbol{n}} \tilde{p}(\boldsymbol{m}, \boldsymbol{n}) u_n(\boldsymbol{n}) \\ &= u_{n+1}(\boldsymbol{m}), \end{aligned}$$

where the second inequality follows from the induction assumption. These findings are summarized in the following theorem.

Theorem 4.1 *Provided the modified chain has been constructed by redirecting transitions to more attractive states, it holds that* $\tilde{g} \leq g$.

The important conclusion is that based on the set P we are able to construct upper and lower bounds. Redirecting transitions to more (less) attractive states yields a lower (upper) bound model. Also, the richer P the more flexibility one has to construct bounds. Before turning to the proof of precedence pairs, we first present precedence pairs for the shortest queue model.

Example 4.2 *(Precedence relations for the shortest queue model)*
It may be shown that state $\boldsymbol{m} = (m_1, m_2)$ is more attractive than all states $\boldsymbol{n} = (n_1, n_2)$ satisfying $n_1 + n_2 \leq m_1 + m_2$ and $n_2 \leq m_2$. This means that it is preferable to have less jobs in the system and/or to have more balance in queue lengths. In particular, denoting the unity vectors $(1,0)$ and $(0,1)$ by $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$, resp., it follows that state $\boldsymbol{m}$ is more attractive than its neighboring states $\boldsymbol{m} + \boldsymbol{e}_1$, $\boldsymbol{m} + \boldsymbol{e}_2$ and $\boldsymbol{m} + \boldsymbol{e}_2 - \boldsymbol{e}_1$ (provided these neighbors are in the state space). The latter precedences are illustrated in Figure 4. Note that they imply the precedences mentioned first. This aspect will be exploited in the next section.

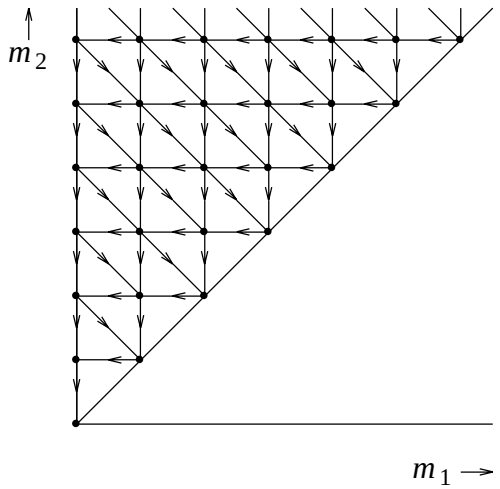

Figure 4: Precedence relations for the shortest queue model with one-period cost given by (1). Each arrow points to a more attractive state

It is easily verified, by using the precedence pairs, that the OIB and TK model give upper bounds for the mean normalized waiting time and the others give lower bounds. In fact, we only need the precedences between neighboring states depicted in Figure 4. And to establish the bounds for the TK and TB model, we even need less. For these models we only use that $\boldsymbol{m}$ is more attractive that $\boldsymbol{m} + \boldsymbol{e}_2$ and $\boldsymbol{m} + \boldsymbol{e}_1$, resp. □

5 ESTABLISHING PRECEDENCE PAIRS

Let us consider the general model again. Suppose that we have a set P of pairs $(\boldsymbol{m}, \boldsymbol{n})$ and that we want to prove for all $n \geq 0$ that (cf. (3))

$$v_n(\boldsymbol{m}) \leq v_n(\boldsymbol{n}) \qquad \text{for all } (\boldsymbol{m}, \boldsymbol{n}) \in P. \tag{4}$$

The proof of (4) is done by induction over n. Taking $n = 1$ in (4) directly leads to the conclusion that the ordering

$$c(\boldsymbol{m}) \leq c(\boldsymbol{n}) \tag{5}$$

should hold for all $(\boldsymbol{m}, \boldsymbol{n}) \in P$. Let us suppose that this is indeed the case and consider the induction step. Assume (4) holds for n. To establish it for $n + 1$ it is convenient to exploit that $\leq$ is transitive: if $(\boldsymbol{m}, \boldsymbol{n})$ and $(\boldsymbol{n}, \boldsymbol{l})$ satisfy (4) for $n + 1$, then so does $(\boldsymbol{m}, \boldsymbol{l})$ for $n + 1$. So, there may be a small subset of P with the property that if inequality (4) holds for the pairs in that subset, then it also holds for all pairs in P by virtue of transitivity of $\leq$. Denote such a subset by $P_{\_}$ (possibly $P_{\_}$ is the same as P). Hence, to

prove inequality (4) for $n+1$ for all pairs in P, it suffices to do so for the pairs in the smaller set P_{-}.

Example 5.1 *(Set P_{-} for the shortest queue problem)*
In the previous section we introduced the set P for the shortest queue problem. This set includes all pairs $(\boldsymbol{m}, \boldsymbol{n})$ satisfying $n_1 + n_2 \leq m_1 + m_2$ and $n_2 \leq m_2$. Let P_{-} be the set of pairs $(\boldsymbol{m}, \boldsymbol{n})$ for which $\boldsymbol{n}$ is equal to $\boldsymbol{m} + \boldsymbol{e}_1$, $\boldsymbol{m} + \boldsymbol{e}_2$ or $\boldsymbol{m} + \boldsymbol{e}_2 - \boldsymbol{e}_1$. Clearly P_{-} is a subset of P and it is easily seen that the inequalities (4) for the pairs in P_{-} generate (by using transitivity) the ones for all pairs in P. □

To establish (4) for $n+1$, we have to show for each $(\boldsymbol{m}, \boldsymbol{n}) \in P_{-}$ that

$$v_{n+1}(\boldsymbol{m}) = c(\boldsymbol{m}) + \sum_{\boldsymbol{i}} p(\boldsymbol{m}, \boldsymbol{i}) v_n(\boldsymbol{i}) \leq c(\boldsymbol{n}) + \sum_{\boldsymbol{j}} p(\boldsymbol{n}, \boldsymbol{j}) v_n(\boldsymbol{j}) = v_{n+1}(\boldsymbol{n}). \qquad (6)$$

By (5), it suffices to show that the sums are also ordered. A common approach is to compare similar terms in the two sums, i.e. terms corresponding to the same event (such as an arrival or departure). Further, it is usually sufficient to distinguish a few cases for $(\boldsymbol{m}, \boldsymbol{n})$ only. It depends on the application on hand which terms are similar and which cases have to be considered. Below it is shown how this works for the shortest queue problem.

Example 5.2 *(Proof of precedence pairs for the shortest queue model)*
Let us illustrate for the shortest queue model how (6) may be proved for the pairs $(\boldsymbol{m}, \boldsymbol{m}+\boldsymbol{e}_1)$. We distinguish four cases, namely $\boldsymbol{m} = (0,1)$, $\boldsymbol{m} = (0, m_2)$ with $m_2 > 1$, $\boldsymbol{m} = (m_1, m_1+1)$ with $m_1 > 0$ and finally $\boldsymbol{m} = (m_1, m_2)$ with $m_1 > 0$ and $m_2 > m_1 + 1$. In the third case we have

$$v_{n+1}(\boldsymbol{m}) = c(\boldsymbol{m}) + \lambda v_n(\boldsymbol{m} + \boldsymbol{e}_1) + \mu v_n(\boldsymbol{m} - \boldsymbol{e}_1) + \mu v_n(\boldsymbol{m} - \boldsymbol{e}_2), \qquad (7)$$

$$v_{n+1}(\boldsymbol{m} + \boldsymbol{e}_1) = c(\boldsymbol{m} + \boldsymbol{e}_1) + \lambda v_n(\boldsymbol{m} + \boldsymbol{e}_1 + \boldsymbol{e}_2) + \mu v_n(\boldsymbol{m}) + \mu v_n(\boldsymbol{m}). \qquad (8)$$

Now compare the right-hand sides of (7) and (8). The direct cost in (7) is less than the one in (8). The second terms, both corresponding to an arrival, are ordered accordingly by the induction assumption. The same holds for the third and fourth terms. So $v_{n+1}(\boldsymbol{m}) \leq v_{n+1}(\boldsymbol{m} + \boldsymbol{e}_1)$. The other cases can be proved similarly. To complete the induction step we also have to prove (6) for the combinations $(\boldsymbol{m}, \boldsymbol{m} + \boldsymbol{e}_2)$ and $(\boldsymbol{m}, \boldsymbol{m} + \boldsymbol{e}_2 - \boldsymbol{e}_1)$, resp. (cf. Section 4 in [3]). □

It is also possible to establish (6) more systematically (which may be useful in complex models). Below we will show that this problem can be translated into a *transportation problem* (see e.g. [5]).

To prove for a pair $(\boldsymbol{m}, \boldsymbol{n}) \in P_{-}$ that the sums in (6) are ordered we may of course restrict the sums to states $\boldsymbol{i}$ and $\boldsymbol{j}$ for which $p(\boldsymbol{m}, \boldsymbol{i})$ and $p(\boldsymbol{n}, \boldsymbol{j})$, resp., are positive. Denote the sets of these states by $V(\boldsymbol{m})$ and $V(\boldsymbol{n})$. If we introduce nonnegative variables $a(\boldsymbol{i}, \boldsymbol{j})$ satisfying

$$p(\boldsymbol{m}, \boldsymbol{i}) = \sum_{\boldsymbol{j} \in V(\boldsymbol{n})} a(\boldsymbol{i}, \boldsymbol{j}), \qquad p(\boldsymbol{n}, \boldsymbol{j}) = \sum_{\boldsymbol{i} \in V(\boldsymbol{m})} a(\boldsymbol{i}, \boldsymbol{j}), \tag{9}$$

then we may write

$$\sum_{\boldsymbol{i} \in V(\boldsymbol{m})} p(\boldsymbol{m}, \boldsymbol{i}) v_n(\boldsymbol{i}) - \sum_{\boldsymbol{j} \in V(\boldsymbol{n})} p(\boldsymbol{n}, \boldsymbol{j}) v_n(\boldsymbol{j}) = \sum_{\boldsymbol{i} \in V(\boldsymbol{m})} \sum_{\boldsymbol{j} \in V(\boldsymbol{n})} a(\boldsymbol{i}, \boldsymbol{j})(v_n(\boldsymbol{i}) - v_n(\boldsymbol{j})). \tag{10}$$

In transportation terminology, the states $\boldsymbol{i}$ are supply nodes with supply $p(\boldsymbol{m}, \boldsymbol{i})$ and the states $\boldsymbol{j}$ are demand nodes with demand $p(\boldsymbol{n}, \boldsymbol{j})$. A solution a satisfying (9) is an allocation (there always exists one since the supply and demand both add up to 1). If there exists an allocation a for which $a(\boldsymbol{i}, \boldsymbol{j}) = 0$ for all pairs $(\boldsymbol{i}, \boldsymbol{j}) \notin P$, then we can conclude from the induction assumption that the right-hand side in (10) is less than or equal to 0. Such an allocation is called *feasible*. Hence the proof of (6) for a pair $(m, n) \in P_{-}$ has now been reduced to that of finding a feasible allocation for a corresponding transportation problem. This transportation problem is denoted by $\boldsymbol{TP(m, n)}$.

Theorem 5.3 *Provided*

(i) $c(\boldsymbol{m}) \leq c(\boldsymbol{n})$ *for all* $(\boldsymbol{m}, \boldsymbol{n}) \in P$*;*

(ii) The transportation problem $\boldsymbol{TP(m, n)}$ *has a feasible allocation for each* $(\boldsymbol{m}, \boldsymbol{n}) \in P_{-}$*, where* P_{-} *is a set of pairs generating* P*,*

it holds for all $n \geq 0$ *that* $v_n(\boldsymbol{m}) \leq v_n(\boldsymbol{n})$ *for all* $(\boldsymbol{m}, \boldsymbol{n}) \in P$*.*

Example 5.4 *(Transportation problem for the shortest queue model)*
We will illustrate for the shortest queue model how (6) can be proved for the pairs $(\boldsymbol{m}, \boldsymbol{m} + \boldsymbol{e}_1)$ by solving the corresponding transportation problem. As before we distinguish four cases. In case $\boldsymbol{m} = (m_1, m_1 + 1)$ with $m_1 > 0$ the supply nodes are $\boldsymbol{m} - \boldsymbol{e}_1$, $\boldsymbol{m} - \boldsymbol{e}_2$ and $\boldsymbol{m} + \boldsymbol{e}_1$ with supply μ, μ and λ, resp. The demand nodes are $\boldsymbol{m}$ and $\boldsymbol{m} + \boldsymbol{e}_1 + \boldsymbol{e}_2$ with

demand 2μ and λ, resp. A feasible allocation may only transship supply between pairs of nodes in P, i.e., between the pairs $(\boldsymbol{m}-\boldsymbol{e}_1, \boldsymbol{m})$, $(\boldsymbol{m}-\boldsymbol{e}_1, \boldsymbol{m}+\boldsymbol{e}_1+\boldsymbol{e}_2)$, $(\boldsymbol{m}-\boldsymbol{e}_2, \boldsymbol{m})$, $(\boldsymbol{m}-\boldsymbol{e}_2, \boldsymbol{m}+\boldsymbol{e}_1+\boldsymbol{e}_2)$ and $(\boldsymbol{m}+\boldsymbol{e}_1, \boldsymbol{m}+\boldsymbol{e}_1+\boldsymbol{e}_2)$. This transportation problem is illustrated in Figure 5. The arrows indicate to which nodes transshipments are allowed.

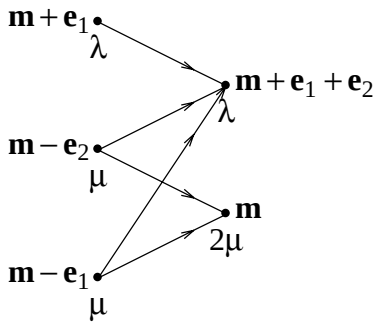

Figure 5: Transportation problem for the pair $(\boldsymbol{m}, \boldsymbol{m}+\boldsymbol{e}_1)$. The arrows indicate to which nodes transshipments are allowed.

It is easily verified that the allocation $a(\boldsymbol{m}-\boldsymbol{e}_1, \boldsymbol{m}) = \mu$, $a(\boldsymbol{m}-\boldsymbol{e}_2, \boldsymbol{m}) = \mu$ and $a(\boldsymbol{m}+\boldsymbol{e}_1, \boldsymbol{m}+\boldsymbol{e}_1+\boldsymbol{e}_2) = \lambda$ is a solution to the transportation problem in Figure 5. □

Remark 5.5 *(Existence of feasible allocations)*
A simple condition for the existence of a feasible allocation is the following one (cf. [25, 5]). There exists a feasible allocation if and only if

$$\sum_{\boldsymbol{i}\in U} p(\boldsymbol{m}, \boldsymbol{i}) \leq \sum_{\boldsymbol{j}\in R(U)} p(\boldsymbol{n}, \boldsymbol{j}) \qquad \text{for all } U \subset V(\boldsymbol{m}),$$

where $R(U)$ are the states which may receive supply from U, i.e. states $\boldsymbol{j} \in V(\boldsymbol{n})$ for which there is an $\boldsymbol{i} \in U$ such that $(\boldsymbol{i}, \boldsymbol{j}) \in P$. □

Remark 5.6 *(Cost functions)*
It will be clear that the precedence relations are valid for any cost function c satisfying ordering (5) for all $(\boldsymbol{m}, \boldsymbol{n}) \in P$. So for appropriate definitions of c we may establish bounds for various performance measures. For instance, for the shortest queue model, if we take $c(\boldsymbol{m}) = 1$ for all states $\boldsymbol{m}$ with $m_1 > M$ and 0 otherwise, then g corresponds to the probability that the shortest queue is longer than M. This cost function, however, only satisfies ordering (5) for the pairs $(\boldsymbol{m}, \boldsymbol{n})$ satisfying $n_1 \geq m_1$ and $n_2 \geq m_2$, but not for pairs like $(\boldsymbol{m}, \boldsymbol{m}-\boldsymbol{e}_1+\boldsymbol{e}_2)$. Fortunately, we do not need the latter for the TK and TB model, and hence we find that these models still produce a lower and upper bound,

resp., for the probability that the shortest queue is longer than M. So, the length of the shortest queue in the TK (TB) model is stochastically smaller (larger) than in the original model. As a consequence, the waiting time is also stochastically smaller (larger). $\square$

6 NUMERICAL RESULTS

In this section, we present for the shortest queue model some numerical results for the bounds on the mean normalized waiting time W. Recall that each of the models introduced in Section 3 can be efficiently solved; see the remarks at the end of that section. The key parameter in each of these models is the threshold T. The larger T the more accurate the bounds will be, but also the greater the effort to compute them.

The effect of T on the accuracy of the bounds is demonstrated in Table 1. For server utilization $\rho = 0.9$, where ρ is defined as $\lambda/2\mu$, we list for increasing values of T and for each model the difference of the bound and the true W, which is 4.475. Note that the service capacity is not efficiently used the OIB and TB model, which explains why these models are not stable for $T = 1$. The results in Table 1 show that the bounds are tight for already small values of T, except for the first model, which performs poorly.

T	FB	CB	TJ	OIB	TK	TB
1	-4.475	-0.212	-0.212	∞	-3.420	∞
2	-4.096	-0.093	-0.093	18.00	-1.844	25.72
3	-3.682	-0.040	-0.037	2.261	-0.761	1.500
4	-3.285	-0.018	-0.013	0.738	-0.263	0.349
5	-2.916	-0.008	-0.005	0.288	-0.083	0.096
6	-2.577	-0.003	-0.002	0.121	-0.025	0.028
7	-2.269	-0.001	-0.001	0.052	-0.008	0.008
8	-1.991	-0.001	-0.000	0.023	-0.002	0.002

Table 1: Differences of the bounds and the true $W = 4.475$ for $\rho = 0.9$

The system with two queues may of course be solved exactly and very efficiently (see [1]). For larger systems, however, no exact analytical results are availaible. The power of the present approach is that also works well for large systems. In [3, 25] extensive numerical material can be found demonstrating that the CB, TJ and TB models produce accurate bounds for the mean waiting time in systems with up to 50 queues.

7 CONCLUSIONS

In this paper we have presented a systematic approach for the construction of bounds for the average cost in an infinite state Markov chain. The essence of the approach is that one should first try to identify precedences between states of the Markov chain. Based on these precedences it appears to be easy to formulate suitable Markov chains producing bounds for the average cost. It is often possible to construct a sequence of bounds converging to the true average cost. But, of course, the more accurate the bound, the more effort it takes to compute it.

Many queueing or inventory systems can be described by Markovian models and relevant performance characteristics in these systems, such as, e.g., mean waiting times or mean lead times, may be represented in the Markovian model by the average cost for some appropriately chosen cost structure. Therefore, the approach presented in this paper may be applied to many systems to generate bounds for the relevant performance characteristics.

References

[1] Adan, I. J. B. F., Wessels, J., Zijm, W. H. M., *Analysis of the symmetric shortest queue problem*, Stochastic Models, 6 (1990), pp. 691–713.

[2] ——, *Matrix-geometric analysis of the symmetric shortest queue problem with threshold jockeying*, Opns. Res. Lett., 13 (1993), pp. 107–112.

[3] Adan, I. J. B. F., Van Houtum, G. J., Van der Wal, J., *Upper and lower bounds for the waiting time in the symmetric shortest queue system*, Annals of Operat. Research, 48 (1994), pp. 197–217.

[4] Adan, I. J. B. F., Hooghiemstra, G., *The $M/M/c$ with critical jobs*, Memorandum COSOR 96-20, Dept. of Math. and Comp. Sc., Eindhoven University of Technology, 1996.

[5] Ahuja, R. K., Magnanti, T. L., Orlin, J. B., *Network flows: Theory, algorithms and applications*, Prentice-Hall, Englewood Cliffs, 1993.

[6] Conolly, B. W. , *The autostrada queueing problem*, J. Appl. Prob., 21 (1984), pp. 394–403.

[7] Flatto, L., McKean, H. P., *Two queues in parallel*, Comm. Pure Appl. Math., 30 (1977), pp. 255–263.

[8] Foschini, G. J., Salz, J., *A basic dynamic routing problem and diffusion*, IEEE Trans. Commun., 26 (1978), pp. 320–327.

[9] Gertsbakh, I., *The shorter queue problem: a numerical study using the matrix-geometric solution*, EJOR, 15 (1984)m pp. 374–381.

[10] Grassmann, W. K., *Transient and steady state results for two parallel queues*, OMEGA Int. J. of Mgmt. Sci., 8 (1980), pp. 105-112.

[11] Halfin, S., *The shortest queue problem*, J. Appl. Prob., 22 (1985), pp. 865–878.

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

[13] Kingman, J. F. C., *Two similar queues in parallel*, Ann. Math. Statist., 32 (1961), pp. 1314–1323.

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

[15] Nelson, R., Philips, T. K., *An approximation to the response time for the shortest queue routing*, Performance Evaluation Review, 7 (1989), pp. 181–189.

[16] ——, *An approximation for the mean response time for the shortest queue routing with general interarrival and service times*, Performance Evaluation, 17 (1993), pp. 123–139.

[17] Neuts, M. F., *Matrix-geometric solutions in stochastic models*, Johns Hopkins University Press, Baltimore, 1981.

[18] Rao, B. M., Posner, M. J. M., *Algorithmic and approximation analysis of the shorter queue model*, Naval Res. Log., 34 (87), pp. 381–398.

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

[20] Van Dijk, N. M., *A formal proof for the insensitivity of simple bounds for finite multi-server non-exponential tandem queues*, Stochastic Processes, 27 (1988), pp. 261–277.

[21] ——, *Simple bounds for queueing systems with breakdowns*, Perf. Evaluation, 8 (1988), pp. 117–128.

[22] ——, *Queueing networks and product forms: a systems approach*, Wiley, Chichester, 1993.

[23] Van Dijk, N. M., Lamond, B. F., *Simple bounds for finite single-server exponential tandem queues*, Opns. Res., 36 (1988), pp. 470–477.

[24] Van Dijk, N. M., Van der Wal, J., *Simple bounds and monotonicity results for finite multi-server exponential tandem queues*, QUESTA, 4 (1989), pp. 1–16.

[25] Van Houtum, G. J., *New approaches for multi-dimensional queueing systems*, Thesis, Eindhoven University of Technology, Eindhoven, 1995.

[26] Van Houtum, G. J., Adan, I. J. B. F., and Van der Wal, J., *The symmetric longest queue system*, Stochastic Models, 13 (1997), pp. 105–120.

[27] Zhao, Y., and Grassmann, W. K., *Queueing analysis of a jockeying model*, Opns. Res., 43 (1995), pp. 520–529.