

On the precedence relation method for deriving flexible bound models for queueing systems

Citation for published version (APA):

Houtum, van, G. J. J. A. N., Adan, I. J. B. F., Wessels, J., & Zijm, W. H. M. (1994). *On the precedence relation method for deriving flexible bound models for queueing systems*. (Memorandum COSOR; Vol. 9427). Technische Universiteit Eindhoven.

Document status and date:

Published: 01/01/1994

Document Version:

Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy

If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

EINDHOVEN UNIVERSITY OF TECHNOLOGY
Department of Mathematics and Computing Science

Memorandum COSOR 94-27

**On the precedence relation method for deriving
flexible bound models for queueing systems**

G.J. van Houtum
I.J.B.F. Adan
J. Wessels
W.H.M. Zijm

Eindhoven, August 1994
The Netherlands

ON THE PRECEDENCE RELATION METHOD FOR DERIVING FLEXIBLE BOUND MODELS FOR QUEUEING SYSTEMS

G.J. van Houtum, I.J.B.F. Adan, J. Wessels and W.H.M. Zijm, Eindhoven

Abstract. In this paper, we describe the so-called *precedence relation method*. This method may be used to derive truncation models which produce bounds for the relevant performance measures of a given Markovian queueing system. The truncation models may be defined such that the size of the state space is flexible in the sense that it depends on the choice of certain truncation parameters. The models which we obtain in this way are called *flexible bound models* and they may lead to efficient procedures for the determination of the performance measures of interest. The precedence relation method will be demonstrated for the symmetric shortest queue system.

1. Introduction

For several queueing systems, the behavior is described by a Markov process on a state space which is discrete and possibly infinite in one or more components. The relevant performance measures for such queueing systems usually may be obtained from the equilibrium distribution of the underlying Markov process. Therefore, in the queueing literature, much attention has been paid to analytical methods for the determination of the equilibrium distribution of Markov processes. This has led to many explicit results for Markov processes with state spaces which are essentially one-dimensional, i.e. which are infinite in at most one direction. However, multi-dimensional Markov processes seem to be much harder to analyze analytically. To our knowledge, only two explicit results have been obtained for the equilibrium distribution of N -dimensional Markov processes with general $N \geq 2$ (see Baskett et al. [3] and [10]); a few more results have been obtained for the case $N = 2$.

Since a lot of multi-dimensional queueing systems cannot be solved analytically, it is desired to have alternative methods to determine the equilibrium distribution, or at least the relevant performance measures. One alternative is constituted by the *power-series algorithm*, which is a numerical technique based on power-series expansions of equilibrium probabilities as function of the traffic load (see Hooghiemstra et al. [5] and Blanc [4]). This method may be used to compute the equilibrium distribution and the relevant performance measures *within a given accuracy*; here, the accuracy that can be reached is restricted by the requirements with respect to the computational effort and the memory space. Another alternative approach is constituted by approximation models which:

- can approximate the exact model as accurate as desired (think of truncation models for which the size of the state space depends on one or more truncation parameters);
- can be solved exactly (or at least within a very high accuracy);
- lead to bounds for the relevant performance measures.

Such models are called *flexible bound models*. Note that bounds/approximations as well as *error bounds* may be obtained by combining a lower and upper bound model; so, the relevant performance measures of a given queueing system may be determined by solving lower and upper bound models for increasing values of the truncation parameters until the desired accuracy has been reached.

In this paper, we explain the main idea of a method that can be used for deriving appropriate flexible bound models. This method is described in Section 2 and it will be called the *precedence relation method*. In Section 3, the method will be applied to the symmetric shortest queue system.

2. The precedence relation method

In general, the relevant performance measures of a given Markovian queueing system may be determined by defining appropriate Markov cost models and computing the average costs. This property is exploited by the so-called precedence relation method, which is based on Markov cost/reward theory and which is similar to the technique used in the papers by Van der Wal [7], Van Dijk and Van der Wal [9], Van Dijk and Lamond [8]. In principle, the precedence relation method is an analytical method which is appropriate for comparing the average costs in two Markov cost models, where the state space of one model is a subset of the state space of the other model. In this section, we shall describe how the method may be used for the comparison between an original system and a truncation model. Here, w.l.o.g. we may restrict ourselves to the discrete-time case, since, (under some mild conditions) continuous-time Markov processes may be transformed to equivalent discrete-time Markov processes by using the uniformization technique.

Consider a discrete-time, irreducible, aperiodic and positive recurrent Markov cost model with a possibly multi-dimensional and/or infinite state space M consisting of N -dimensional vectors $m = (m_1, \dots, m_N)$ and with transition probabilities $q_{m,n}$ and direct costs $c(m)$. Let $\{p_m\}$ be the equilibrium distribution, which is the unique normalized solution of the equilibrium equations. Finally, let g denote the average costs per period:

$$g = \sum_{m \in M} p_m c(m). \quad (1)$$

For the average costs g , we have the following property. Let $v_t(m)$, $t \geq 0$, denote the expected t -period costs, i.e. the expected costs in the next t periods when starting in state m ; so $v_0(m) = 0$, $m \in M$, and for all $t \geq 0$,

$$v_{t+1}(m) = c(m) + \sum_{n \in M} q_{m,n} v_t(n) \quad , m \in M. \quad (2)$$

Then, because of the assumed irreducibility and aperiodicity,

$$g = \lim_{t \rightarrow \infty} \frac{v_t(m)}{t}, \quad (3)$$

where m may be an arbitrary element of the state space M .

Let us now consider a truncation model of the above original model. A truncation model is obtained by defining a truncated state space $M' \subset M$ and modifying all transitions which start in a state $m \in M'$ and end in a state n outside of M' . Suppose that such a transition is *redirected* to a state $n'(m,n)$ inside of M' . Denote the transition probabilities obtained for the truncation model by $q'_{m,n}$. Further, $c'(m)$, $\{p'_{m,n}\}$, g' and $v'_t(m)$ are the direct costs, equilibrium distribution, average costs and t -period cost functions, respectively. Assume that $c'(m) = c(m)$ for all $m \in M'$. Note that for the truncation model, relations similar to (1)-(3) are valid.

Now, suppose that we suspect that the truncation model leads to a lower bound for g , i.e. that $g' \leq g$. Then, by (3) and the corresponding relation for g' , it suffices to show that for some state $m \in M'$,

$$v'_t(m) \leq v_t(m) \quad \text{for all } t \geq 0; \quad (4)$$

however, if (4) holds for some state $m \in M'$, then it probably also holds for all other states of M' .

The main idea of the *precedence relation method* is that (4) may be proved to hold for all $m \in M'$ by performing the following two steps:

1. Derive a set P consisting of pairs (m,n) of states $m,n \in M$ satisfying

$$v_t(m) \leq v_t(n) \quad \text{for all } t \geq 0. \quad (5)$$

This relation is called a *precedence relation*, and it states that in the original model state m is *more attractive* with respect to costs than state n , or equivalently, state n is *less attractive* than state m ;

2. Exploit the precedence relations derived in step 1 to show that (4) holds for all $m \in M'$.

Step 1 usually requires most of the work. This step may be performed by first defining a set P which is expected to consist of precedence pairs, and next proving by induction with respect to t that (5) holds for all (m,n) of this set P . Note that, since $v_1(m) = c(m)$ for all $m \in M$, all pairs $(m,n) \in P$ must satisfy the condition that $c(m) \leq c(n)$. Typical precedence pairs that can be derived in case the states represent queue lengths, are pairs of the type $(m, m+e_i)$, where e_i is the i -th unit vector. Step 2 is further explained in the next paragraph.

In step 2, we must prove that (4) holds for all $m \in M'$. The inequalities $v'_t(m) \leq v_t(m)$, $m \in M'$, hold for $t=0$ by definition. By using induction with respect to t , they can be proved to hold for all $t \geq 0$, if the truncation model has been constructed such that for all $m \in M'$ and $n \in M \setminus M'$ with $q_{m,n} > 0$, the state $n'(m,n)$ to which the transition from m to n has been redirected, is more attractive than n , i.e. if $(n'(m,n), n) \in P$. If this condition is satisfied, then the induction step reads as follows:

$$\begin{aligned} v'_{t+1}(m) &= c(m) + \sum_{\substack{n \in M' \\ q_{m,n} > 0}} q_{m,n} v'_t(n) + \sum_{\substack{n \in M \setminus M' \\ q_{m,n} > 0}} q_{m,n} v'_t(n'(m,n)) \\ &\leq c(m) + \sum_{\substack{n \in M' \\ q_{m,n} > 0}} q_{m,n} v_t(n) + \sum_{\substack{n \in M \setminus M' \\ q_{m,n} > 0}} q_{m,n} v_t(n'(m,n)), \\ &\leq c(m) + \sum_{\substack{n \in M' \\ q_{m,n} > 0}} q_{m,n} v_t(n) + \sum_{\substack{n \in M \setminus M' \\ q_{m,n} > 0}} q_{m,n} v_t(n), \\ &= v_{t+1}(m), \quad m \in M'. \end{aligned}$$

This completes the description of how the precedence relation method may be used to prove that a truncation model leads to a lower bound for the average costs g in the original model. In a similar way, the precedence relation method may be used to prove that a truncation model leads to an upper bound for g ; in that case it is required for all $m \in M'$ and $n \in M \setminus M'$ with $q_{m,n} > 0$, that the transition from m to n is redirected to a state $n'(m,n)$ which is *less* attractive than n , i.e. for which $(n, n'(m,n)) \in P$.

An important property of the method described above is that the introduction of the concept of precedence relations leads to simple sufficient conditions for obtaining lower and upper bound models. Therefore, the precedence relation method may also be used for *deriving* bound models, and especially for deriving *flexible* bound models. The *precedence relation method for deriving flexible bound models* consists of the following two steps:

1. The derivation of a set P consisting of pairs (m,n) , $m,n \in M$, which satisfy (5);
2. The definition of flexible lower and upper bound models: to obtain a flexible lower (upper) bound model, one first must define a flexible truncated state space, and subsequently each transition which starts in a state inside of the truncated state space and ends in a state outside of the truncated state space, must be redirected to a more (less) attractive state than the state in which the transition originally ended.

Note that, once the set of precedence pairs has been derived, a whole set of flexible bound models can be obtained. In the next section, this method will be demonstrated for the symmetric shortest queue system.

3. Application to the symmetric shortest queue system

The symmetric shortest queue system has extensively been studied in the literature. Only for the case with $N=2$ servers, explicit expressions have been found for the equilibrium distribution, and the mean waiting time. For the case with general $N \geq 2$, there are some algorithms available with which the mean waiting time can be determined numerically; see, for example, Blanc [4] and [2]. Up to now, the largest systems, viz. systems with up to $N=50$ servers and workloads up to 0.95, have been solved by the numerical procedure developed in [2]. In this section, we shall derive the two flexible bound models on which this procedure is based; for simplicity, we shall restrict ourselves to the case $N=2$.

The two-dimensional symmetric shortest queue system consists of two parallel servers, which both have their own queue. Jobs arrive according to a Poisson stream with intensity $\lambda > 0$, and an arriving job always joins the shortest queue (ties are broken with equal probabilities). All service times are exponentially distributed with parameter $\mu > 0$. Assume that $\lambda + 2\mu = 1$. In order to have an ergodic system, the workload $\rho = \lambda/(2\mu)$ is assumed to be smaller than 1.

Assume that the servers always work, but that a service completion is only attended by a departure of a job if there is a job present in the corresponding queue. Then the behavior of the system may be described by the discrete-time Markov process on the time instants right after job arrivals and service completions, and with states (m, n) , where m and n represent the lengths of the shortest queue and the longest queue, respectively. So, $M = \{ (m, n) \mid 0 \leq m \leq n \}$. The transition probabilities $q_{m,n}$ are depicted in the first diagram of Figure 1.

We are interested in the mean W of the *normalized* waiting time, which is defined as the waiting time divided by the mean service time. By Little's formula, $W = L_{tw}/(2\rho)$, where L_{tw} denotes the average number of waiting jobs in the system. Define the direct costs $c(m, n)$ by the number of waiting jobs in state (m, n) , so

$$c(m, n) = \max\{m-1, 0\} + \max\{n-1, 0\}, \quad (m, n) \in M. \quad (6)$$

Then L_{tw} is equal to the average costs g in the corresponding Markov cost model.

For the given cost function, we can derive the following set P of precedence pairs:

$$P = \{ ((m, n), (m+1, n)) \mid 0 \leq m < n \} \cup \{ ((m, n), (m, n+1)) \mid 0 \leq m \leq n \} \\ \cup \{ ((m, n), (m-1, n+1)) \mid 0 < m \leq n \}. \quad (7)$$

The pairs in the first two sets state that it is more attractive to be or to start in a state with one job less at one of the two servers. The pairs in the last set state that it is more attractive to be in a state with more balance, i.e. in a state with a smaller difference between the queue lengths.

Since the shortest queue routing causes a strong drift to the diagonal, the original system can be closely approximated by truncation models with state space $M' = \{ (m, n) \mid 0 \leq m \leq n \leq m+T \}$, where T is some positive integer; T is called the *threshold parameter*. For this truncated state space, for all $m \geq 1$, we must redirect the transition from the state $(m, m+T)$ to the state $(m-1, m+T)$. According to the precedence relation method, we obtain a lower bound model by redirecting this transition from $(m-1, m+T)$ to the more attractive state $(m, m+T-1)$, which is equivalent to letting a job jockey from the longest to the shortest queue; therefore, this model is called the *Threshold Jockeying (TJ) model*. An upper bound model is obtained by redirecting the transition to the less attractive state $(m, m+T)$ itself, which is equivalent to blocking the server at the shortest queue in state $(m, m+T)$; therefore, this model is called the *Threshold Blocking (TB) model*. For both truncation models, we have depicted the redirections in Figure 1.

The TJ model leads to a lower bound $L_{tw}^{TJ}(T)$ for L_{tw} , and therefore also to a lower bound $W_{TJ}(T) = L_{tw}^{TJ}(T)/(2\rho)$ for W . The TB model leads to upper bounds $L_{tw}^{TB}(T)$ and $W_{TB}(T) = L_{tw}^{TB}(T)/(2\rho)$.

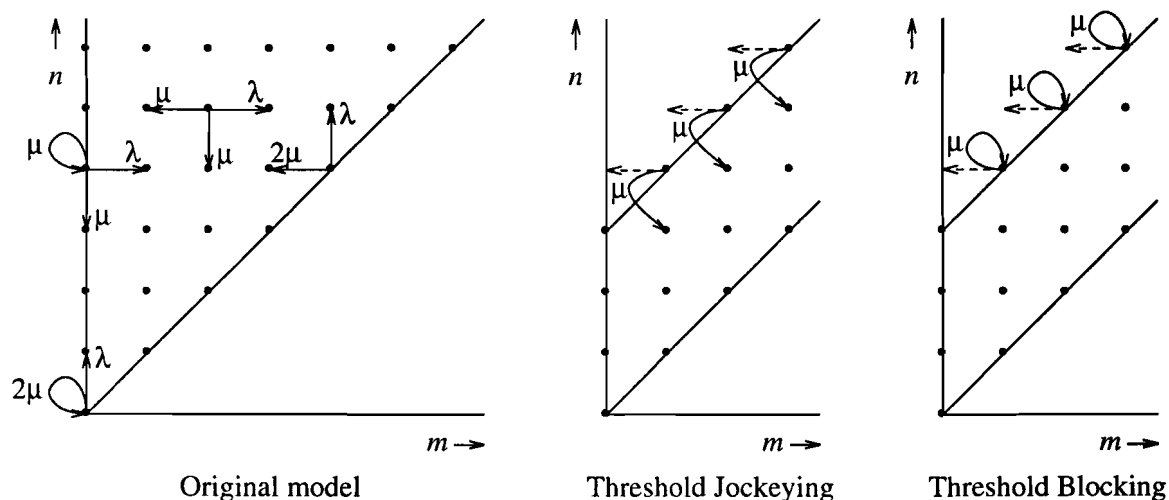


Figure 1. The original model and the two flexible bound models (with $T=3$).

ρ	T	$W_{TJ}(T)$	$W_{TB}(T)$	W	error bound	$W_{M M 2}$	$W - W_{M M 2}$
0.1	2	0.0176	0.0177	0.0177	0.0001	0.0101	0.0076
0.2	2	0.0651	0.0664	0.0657	0.0007	0.0417	0.0240
0.3	2	0.1405	0.1472	0.1439	0.0034	0.0989	0.0450
0.4	3	0.2578	0.2597	0.2587	0.0010	0.1905	0.0683
0.5	3	0.4237	0.4307	0.4272	0.0035	0.3333	0.0939
0.6	4	0.6806	0.6842	0.6824	0.0018	0.5625	0.1200
0.7	5	1.1075	1.1103	1.1089	0.0014	0.9608	0.1481
0.8	6	1.9552	1.9587	1.9570	0.0018	1.7778	0.1792
0.9	7	4.4744	4.4831	4.4787	0.0044	4.2632	0.2156
0.95	9	9.4865	9.4914	9.4890	0.0025	9.2564	0.2326
0.98	11	24.4946	24.4985	24.4965	0.0020	24.2525	0.2440
0.99	12	49.4983	49.5028	49.5006	0.0023	49.2513	0.2470

Table 1. The mean normalized waiting time W for increasing values of ρ ($\epsilon_{abs}=0.005$).

Further, it can be expected that both $W_{TJ}(T)$ and $W_{TB}(T)$ tend to W , as $T \rightarrow \infty$, since for $T=\infty$ both truncation models are identical to the original model. By considering the bound models with threshold parameter T as truncation models of the bound models with threshold parameter $T+1$, it can be proved that the lower bounds $W_{TJ}(T)$ are monotonously increasing and that the upper bounds $W_{TB}(T)$ are monotonously decreasing. So, we find that

$$W_{TJ}(T) \uparrow W \text{ and } W_{TB} \downarrow W, \text{ as } T \rightarrow \infty. \quad (8)$$

It is noted that, since the TJ model with parameter $T=1$ leads to the same mean normalized waiting time as the $M|M|2$ system, a direct consequence of what we have proved is that the mean normalized waiting time W in the symmetric shortest queue system is larger than or equal to the mean normalized waiting time $W_{M|M|2}$ in the related $M|M|2$ system.

The result stated in (8) leads to the following exact method for the determination of W . The mean normalized waiting time W can be determined within an arbitrary, given accuracy by computing $W_{TJ}(T)$ and $W_{TB}(T)$ for increasing values of T ; here, for both truncation models, the equilibrium distribution, and therefore also the bounds for L_{nw} and W , can be determined efficiently by using the matrix-geometric approach, as described by Neuts [6].

We finally present some numerical results. We have developed a numerical procedure which, for a given value of ρ , determines W within a given absolute accuracy ϵ_{abs} by computing $W_{TJ}(T)$ and $W_{TB}(T)$ for $T=1,2,\dots$. For each T , $(W_{TJ}(T)+W_{TB}(T))/2$ is used as an approximation for W and $(W_{TB}(T)-W_{TJ}(T))/2$ is used as an upper bound for the absolute error of this approximation; the computation process is stopped as soon as this error bound is smaller than or equal to ϵ_{abs} . In Table 1, we have listed the numerical results obtained for $\epsilon_{abs}=0.005$ and increasing values of ρ . The values in the second column, denoting the smallest values for T for which the absolute accuracy ϵ_{abs} is reached, show that the truncation models lead to sufficiently accurate approximations for W for already small values of T . It is noted that the numerical results for W can be used, among others, to investigate the difference between the symmetric shortest queue system and the $M|M|2$ system (the last column in Table 1 shows that there is an interesting behavior for the difference between the mean normalized waiting times in both systems).

4. Concluding remarks

In this paper, we have briefly described the precedence relation method; a more extensive treatment is given in [11]. The method has been applied to the symmetric shortest queue system to derive flexible truncation models which lead to bounds for the mean normalized waiting time. These flexible bound models have resulted in an efficient numerical procedure for the computation of the mean normalized waiting time within a given accuracy (see Section 3 and [2]). The method also has led to successful flexible bound models for the symmetric *longest* queue system (see [1]), and it seems to be promising for several other queueing systems.

References

1. ADAN, IVO, VAN HOUTUM, GEERT-JAN, AND VAN DER WAL, JAN, "The symmetric longest queue model," Memorandum COSOR, Eindhoven University of Technology, Dept. of Math. and Comp. Sci., 1994. In preparation.
2. ADAN, IVO, VAN HOUTUM, GEERT-JAN, AND VAN DER WAL, JAN, "Upper and lower bounds for the waiting time in the symmetric shortest queue system," *Ann. Oper. Res.*, vol. 48, pp. 197-217, 1994.
3. BASKETT, F., CHANDY, K.M., MUNTZ, R., AND PALACIOS-GOMEZ, F., "Open, closed and mixed networks of queues with different classes of customers," *Journal of the ACM*, vol. 22, pp. 248-260, 1975.
4. BLANC, J.P.C., "The power-series algorithm applied to the shortest-queue model," *Oper. Res.*, vol. 40, pp. 157-167, 1992.
5. HOOGHMSTRA, 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.
6. NEUTS, MARCEL F., *Matrix-Geometric Solutions in Stochastic Models*, Johns Hopkins University Press, Baltimore, 1981.
7. VAN DER WAL, J., "Monotonicity of the throughput of a closed exponential queueing network in the number of jobs," *OR Spektrum*, vol. 11, pp. 97-100, 1989.
8. VAN DIJK, N. AND LAMOND, B.F., "Simple bounds for finite single-server exponential tandem queues," *Oper. Res.*, vol. 36, pp. 470-477, 1988.
9. VAN DIJK, N. AND VAN DER WAL, J., "Simple bounds and monotonicity results for finite multi-server exponential tandem queues," *Queueing Systems*, vol. 4, pp. 1-16, 1989.
10. VAN HOUTUM, G.J., ADAN, I.J.B.F., WESSELS, J., AND ZIJM, W.H.M., "The equilibrium distribution for a class of multi-dimensional random walks," Memorandum COSOR 94-01, Eindhoven University of Technology, Dept. of Math. and Comp. Sci., 1994. Submitted for publication.
11. VAN HOUTUM, G.J., ADAN, I.J.B.F., WESSELS, J., AND ZIJM, W.H.M., "The precedence relation method for deriving flexible bound models for queueing systems," Memorandum COSOR, Eindhoven University of Technology, Dept. of Math. and Comp. Sci., 1994. In preparation.