On the Precedence Relation Method for Deriving Flexible Bound Models for Queueing Systems

G.J. van Houtum [1),*)]
I.J.B.F. Adan [2)]
J. Wessels [2),3)]
W.H.M. Zijm [1)]

University of Twente, The Netherlands

Abstract

In this paper, we present 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 obtained 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.

Keywords: Queueing systems, Markov cost models, performance analysis, truncation models, bounds.

---

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

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

3) International Institute for Applied Systems Analysis, A-2361 Laxenburg, Austria

*) *Corresponding author.* Mailing address: see 1), Phone: +31 53 893192, Fax: +31 53 356490, E-mail: g.j.j.a.n.vanhoutum@wb.utwente.nl

1. Introduction

For several queueing systems, the behavior is described by a Markov process on a multi-dimensional 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, explicit results for the equilibrium distribution have only been obtained for two special classes of N-dimensional Markov processes with general $N \geq 2$ (see Baskett et al. [3] for the so-called product-form networks, and see [14] or the Chapters 2-4 of [13] for the second class); a few more results have been obtained for the case $N = 2$.

Since many 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 a function of the traffic load (see e.g. Hooghiemstra et al. [6] and Blanc [4, 5]). 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 using 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 an 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 is applied to the symmetric shortest queue system, and after that some concluding remarks are given in Section 4.

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 [10], Van Dijk and Van der Wal [12], and Van Dijk and Lamond [11]. 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, without loss of generality 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 and positive recurrent Markov cost model with a possibly multi-dimensional and/or infinite state space M consisting of N-dimensional vectors $m=(m_1, \ldots, m_N)$ 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:

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

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 $t \geq 0$,

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

Then, because of the assumed irreducibility,

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

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 first defining a truncated state space $M' \subset M$ (M' is usually defined such that it contains the states where most of the probability mass is expected to be present) and next modifying the transitions of the original model such that the states outside M' become transient (initially, all transition probabilities $q'_{m,n}$ of the truncation model are taken equal to the transition probabilities $q_{m,n}$ of the original model). This means that each transition starting in a state $m \in M'$ and ending in a state n outside of M', must be *redirected* to a state $n'(m,n)$ inside M' (the probability $q'_{m,n}$ is set equal to 0 and $q'_{m,n'(m,n)}$ is increased by $q_{m,n}$).

Let $c'(m)$, $\{p'_m\}$, g' and $v'_t(m)$ denote the direct costs, equilibrium distribution, average costs and t-period cost functions, respectively. Assume that $c'(m)=c(m)$ for all $m \in M'$. Further, assume that the constructed truncation model is irreducible. Note that for the truncation model, relations similar to (2) and (3) are valid; a relation similar to (1) holds if the truncation model is also positive recurrent.

Now, suppose that the truncation model is expected to lead to a lower bound for g (since it seems that the transitions ending in states outside the truncated state space have been redirected to 'more favourable' states), i.e. that it is expected that $g' \leq g$. Then, by (3) and the corresponding relation for g', it suffices to prove that for some $m \in M$ and $m' \in M'$,

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

Because of the resemblance between both models, it seems reasonable to try to prove that this relation holds for some states $m \in M$ and $m' \in M'$ with $m=m'$; further, if it holds for some state $m \in M'$ that $v'_t(m) \leq v_t(m)$ for all $t \geq 0$, then probably this also holds for all other states of M'. Therefore, we shall focus on trying to prove that $g' \leq g$ by showing that

$$v'_t(m) \leq v_t(m) \quad \text{for all } m \in M' \text{ and } t \geq 0. \tag{4}$$

The inequalities stated in (4) may be proved by using the *precedence relation method*. The main idea of this method is that the comparison of the t-period costs $v'_t(m)$ in the truncation model to the corresponding t-period costs $v_t(m)$ in the original model may be simplified by first performing a preliminary step, in which on the basis of a *precedence relation* for the t-period costs $v_t(m)$ an ordering for the states of the original model is derived. The precedence relation method consists of the following two steps:

1. Derive a set P consisting of *precedence pairs* (m,n) of states $m,n \in M$, which satisfy the *precedence relation*

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

   This relation states that in the original model, state m *has precedence over* state n with respect to the t-period costs, or equivalently, state m is *more attractive* than n, or n is *less attractive* than m;

2. Exploit the precedence pairs derived in step 1 to show that (4) holds.

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 if the components of the states represent queue lengths and if $c(m)$ is non-decreasing in each component, 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. The inequalities $v'_t(m) \leq v_t(m)$, $m \in M'$, hold for $t=0$ by definition. It appears that, by using induction with respect to t, they can be proved to

hold for all $t \geq 0$, if the following *condition* is satisfied:

for all $m \in M'$ and $n \in M \backslash M'$ with $q_{m,n} > 0$, it holds that the state $n'(m,n)$ to which the transition from m to n has been redirected, is more attractive than the state n, i.e. if $(n'(m,n),n) \in P$.

If this condition is satisfied, then the induction step reads as follows:

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

This completes the description 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 $n \in M \backslash M'$ 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 precedence relation 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 of precedence pairs for the original model, i.e. the derivation of a set P consisting of pairs (m,n) of states $m,n \in M$ which satisfy (5);
2. The definition of flexible lower and upper bound models: to obtain a flexible lower (upper) bound model, first a flexible truncated state space M' must be defined, and next each transition from a state $m \in M'$ to a state $n \in M \backslash M'$ must be redirected to a state $n'(m,n) \in M'$ which, according to the precedence pairs derived in step 1, is more (less) attractive than the state n.

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], Lui and Muntz [7] (see also [8]), and Adan et al. [1] (see also [13]). 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 [1]. 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_1,m_2), where m_1 and m_2 represent the lengths of the shortest queue and the longest queue, respectively. So, $M = \{ m=(m_1,m_2) \mid 0 \leq m_1 \leq m_2 \}$. 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, $W = L_w/(2\rho)$, where L_w denotes the average number of waiting jobs in the system. Define the direct costs $c(m_1,m_2)$ by the number of waiting jobs in state (m_1,m_2), so

$$c(m_1,m_2) = \max\{ m_1-1,0 \} + \max\{ m_2-1,0 \}, \quad (m_1,m_2) \in M. \tag{6}$$

Then L_w is equal to the average costs g in the corresponding Markov cost model.

For the given cost function, the following set P can be proved to consist of precedence pairs (prove that (5) holds for all these pairs by using induction with respect to t):

$$\begin{aligned} P \;=\; & \{ ((m_1,m_2),(m_1+1,m_2)) \mid 0 \leq m_1 < m_2 \} \\ & \cup \{ ((m_1,m_2),(m_1,m_2+1)) \mid 0 \leq m_1 \leq m_2 \} \\ & \cup \{ ((m_1,m_2),(m_1-1,m_2+1)) \mid 0 < m_1 \leq m_2 \}. \end{aligned} \tag{7}$$

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 states on the diagonal, the original system can be closely approximated by truncation models with state space $M' = \{ (m_1,m_2) \mid 0 \le m_1 \le m_2 \le m_1+T \}$, where T is some positive integer; T is called the *threshold parameter*. For this truncated state space, for all $m_1 \ge 1$, we must redirect the transition from the state (m_1,m_1+T) to the state (m_1-1,m_1+T). According to the precedence relation method, we obtain a lower bound model by redirecting this transition from (m_1-1,m_1+T) to the more attractive state (m_1,m_1+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_1,m_1+T) itself, which means that in this state a service completion at the shortest queue is not accompanied by a departure and the job in service has to be served once more. It is easily seen that (because of the memory-less property of the exponential service times) this is equivalent to letting the server at the shortest queue be blocked in state (m_1,m_1+T), and 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 $L_w^{TJ}(T)$ for L_w, and therefore also to a lower bound $W_{TJ}(T) = L_w^{TJ}(T)/(2\rho)$ for W. The TB model leads to upper bounds $L_w^{TB}(T)$ and $W_{TB}(T) = L_w^{TB}(T)/(2\rho)$. Further, it may be expected that both $W_{TJ}(T)$ and $W_{TB}(T)$ tend to W, as $T \to \infty$, 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 monotonically increasing and that the upper bounds $W_{TB}(T)$ are monotonically decreasing. So, we find that

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

It is noted that, although the TJ model is not identical to the $M \mid M \mid 2$ queueing system, it does lead to an equivalent Markov process and therefore to the same values/behavior for several performance measures, among which the average number of waiting jobs in the system and the mean normalized waiting time. So, 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 $W_{M \mid M \mid 2}$ in the related $M \mid M \mid 2$ 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_w and W, can be determined efficiently by using the matrix-geometric approach, as described by Neuts [9].

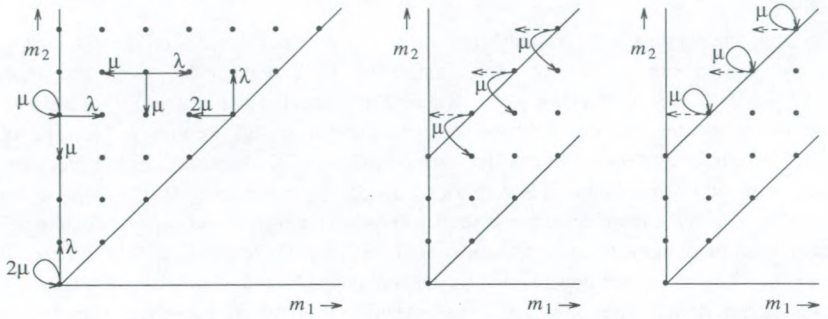

Figure 1. The original model and the two flexible bound models (with $T=3$).

ρ	T	$W_{TJ}(T)$	$W_{TB}(T)$	W	$\Delta(T)$	$W_{M\|M\|2}$	$W-W_{M\|M\|2}$
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	0.2493

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

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 $T=1,2,\cdots$. For each T, $(W_{TJ}(T)+W_{TB}(T))/2$ is used as an approximation for W and $\Delta(T)=W_{TB}(T)-W_{TJ}(T))/2$ is used as an upper bound for the absolute error of this approximation; the computation process is stopped as soon as $\Delta(T)\leq\varepsilon_{abs}$. In Table 1, we have listed the numerical results obtained for $\varepsilon_{abs}=0.005$ 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 other things, 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; for a more extensive treatment, see Chapter 5 of [13] or [15]. The method has been applied to the two-dimensional symmetric shortest queue system, for which we have derived two flexible truncation models leading to lower and upper 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. The method also has led to successful flexible bound models for the N-dimensional symmetric shortest queue system with general $N \geq 2$ and the symmetric *longest* queue system and it seems to be promising for several other queueing systems (see [1, 2, 13]).

References

1. 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.
2. ADAN, IVO, VAN HOUTUM, GEERT-JAN, AND VAN DER WAL, JAN, "The symmetric longest queue system," *Stochastic Models*, 1995. To appear.
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. BLANC, J.P.C., "Performance analysis and optimization with the power-series algorithm," in *Performance Evaluation of Computer and Communication Systems*, ed. R.D. Nelson, pp. 53-90, North-Holland, Amsterdam, 1993.
6. HOOGHIEMSTRA, G., KEANE, M., AND REE, S. VAN DE, "Power series for stationary distributions of coupled processor models," *SIAM J. Appl. Math.*, vol. 48, pp. 1159-1166, 1988.
7. LUI, JOHN C.S. AND MUNTZ, RICHARD R., "Algorithmic approach to bounding the mean response time of a minimum expected delay routing system," *Performance Evaluation Review*, vol. 20, pp. 140-151, 1992.

8. LUI, J.C.S., MUNTZ, R.R., AND TOWSLEY, D., "Bounding the mean response time of a minimum expected delay routing system: An algoritmic approach," CMPSCI Technical report 93-68, University of Massachusetts, 1993. Submitted for publication.

9. NEUTS, MARCEL F., *Matrix-Geometric Solutions in Stochastic Models,* Johns Hopkins University Press, Baltimore, 1981.

10. 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.

11. VAN DIJK, N. AND LAMOND, B.F., "Simple bounds for finite single-server exponential tandem queues," *Oper. Res.*, vol. 36, pp. 470-477, 1988.

12. 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.

13. VAN HOUTUM, GEERT-JAN, *New Approaches for Multi-Dimensional Queueing Systems,* Thesis, Eindhoven University of Technology, Eindhoven, 1995.

14. 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.

15. 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., 1995. In preparation.

Ontvangen: 18-4-1995
Geaccepteerd: 30-11-1995