Teletraffic Analysis and Computer Performance Evaluation
O.J. Boxma, J.W. Cohen, H.C. Tijms (Editors)

STOCHASTIC INEQUALITIES FOR QUEUEING NETWORKS

Arie Hordijk and Ad Ridder
University of Leiden

Abstract. This paper summarizes some techniques for comparing stochastic processes. We focus on (multi-dimensional) birth-death processes because in that case conditions to establish monotonicity and stochastic inequalities become very simple. The conditions concern only the transition rates. A queueing system in computer communication is worked out to illustrate the ideas.

1. INTRODUCTION

This paper exploits comparison techniques of stochastic processes to a queueing system we encountered in a computer communication network. In this system C different types of customers arrive at a service facility consisting of M numbered servers and an infinite waiting room. Server m is able to give service only to a subset A_m of the customer types. The waiting jobs are placed in one queue and as soon server m becomes free he picks out of the queue the first customer whose type belongs to A_m. We assume that the C arrival streams are independent Poisson processes with intensities λ_c, $c=1,\ldots,C$ and that all the service times are exponentially distributed with the same mean μ^{-1}.

Putting the system in a realistic situation the servers are communication lines which link a source with M nodes. In the source calls are generated with different destinations. A connected node is the destination of some calls and transmits other calls further to their destination. In the latter case we deal with a network of nodes which function as sources, destinations and transmitters of calls. Nodes are not (directly) connected with all the others and calls route via shortest paths. Since the analysis of the whole network is too complicated we will study one node in isolation.

The states of the system must describe all the types in service and in the queue as well as their order in the queue. Only with this detailed information the transitions between states are determined and the stochastic process describing

the system is a Markov chain. The service discipline of the queueing system does not fit in the "BCMP-framework" (Baskett et al. [1]). It can be shown that the stationary distribution (we assume it exists) does not fulfill "(type) local balance" (Chandy et al. [2]) nor "job local balance" (Hordijk & van Dijk [3]). So it may not to expect to have a closed form expression for the stationary distribution.

When the queue is part of a network of queues the network does not have a "product form solution". Also the detailed description of the states of this particular queue makes it difficult to handle with. A way to attack this problem is to make an approximate analysis of the network, e.g. decomposition or mean value analysis (cf. chapter 4 in Lavenberg [7]).

The approach in this paper is to try to compare the system with models which are "nicer" or "easier". Especially we hope that these models yield bounds for the interesting quantities. With this approach we have not the intention to get tight approximations. In stead of stressing the numerical analysis we want to develop a general method to produce models which are easier to analyse and give good insight in the performance quantities. Sometimes we obtain lower and upperbounds and then can construct an approximation by an average value dependent of the parameters. Furthermore, we should realize that in real life the system has only a finite buffer. An upperbound for the distribution of the total number of customers in the "infinite case" might give an appropriate size for the buffer.

Before we return to the system we summarize in the following sections some results in the study of comparing stochastic processes. We focus our attention to the cases of comparison between two birth-death processes and comparison between a (more general) Markov process and a birth-death process. The results follow from Stoyan [10] and Whitt [12]. In section 2 we treat monotonicity and in section 3 comparability of the one-step transition probabilities. Together they make the comparison possible of the finite-dimensional marginal distributions and the limiting distributions. In section 4 we sketch the comparison via the sample paths of the processes (cf. Whitt [11], chapter 8 in Ross [9]). The last two sections are devoted to the queueing system. In section 5 we introduce the approximating models and in section 6 we briefly indicate the extension where the queue is part of a network of queues.

2. MONOTONICITY

Let $X = \{X(t): t \geq 0\}$ be a (multi-dimensional) birth-death process on state space $S \subset \mathbb{N}^m$ for some m. The only possible transitions out of state $x = (x_1,\ldots,x_m)$ are to its neighbours $x \pm e_i$ with rates $q(x,x \pm e_i)$ (e_i is the i-th unitvector on S). We assume that X is uniformizable, i.e. the rates of transitions out of any state are uniformly bounded, and set $K = \gamma^{-1}Q + I$ with γ sufficiently large. K is the matrix of transition probabilities of a discrete time Markov chain. Due to the uniformizability assumption monotonicity and comparison of continuous time Markov chains can be expressed through such matrices K (Keilson & Kester [6],Stoyan [10]).

We endow the statespace S with the vector ordering. This ordering induces the collection $\mathcal{I}$ of increasing sets of S by $I \in \mathcal{I}$ iff $x \in I$ and $x \leq y$ imply $y \in I$. The collection $\mathcal{I}$ induces a stochastic ordering (an ordering of probability measures on S) by

(1) $\quad p \leq q$ iff $p(I) \leq q(I)$ for all $I \in \mathcal{I}$.

We call the process X monotone if

(2) $\quad \mathbb{P}(X(t) \in I | X(0) = x)$ is nondecreasing in x for all $I \in \mathcal{I}$ and $t \geq 0$.

Two equivalent characterizations of this monotonicity are

(3) $\quad K(x,I)$ is nondecreasing in x for all $I \in \mathcal{I}$,

(4) $\quad$ K is monotone, i.e. $p \leq q$ implies $pK \leq qK$

(cf. Keilson & Kester [6], Stoyan [10]). In our case of birth-death processes a simple condition on the transition rates establish monotonicity.

LEMMA 1. If for all components i

(5) $\quad q(x,x + e_i)$ is nondecreasing in any argument x_j, $j \neq i$, and

(6) $\quad q(x,x - e_i)$ is nonincreasing in any argument x_j, $j \neq i$,

then X is monotone.

PROOF. It is easily checked that $K(x,I)$ is nondecreasing, due to the single structure of the transition rates. □

3. COMPARISON

Having the monotonicity of the birth-death process X it is possible to compare X with a Markov process $Y = \{Y(t) : t \geq 0\}$. We assume that Y is uniformizable also and that there is a function $f : S_Y \to S_X$ (we use the notation of section 2 with a subindex referring to the processes). The following lemma contains the comparison result (cf. Stoyan [10], Whitt [12]).

LEMMA 2. If

(i) K_X is monotone

(ii) $f(pK_Y) \leq (fp)K_X$ for all probability measures p on S_Y, then

(7) $\mathbb{P}(Y(t) \in f^{-1}(I) \mid Y(0) = y) \leq \mathbb{P}(X(t) \in I \mid X(0) = f(y))$ for all $I \in \mathcal{I}_X$, $y \in S_Y$, $t \geq 0$.

(We have written fp for the induced probability measure on S_X).

REMARKS

(a) Suppose X and Y have stationary distributions π_X resp. π_Y. Then under the conditions of Lemma 2

(8) $f\pi_Y \leq \pi_X$.

(b) Condition (ii) in Lemma 2 is equivalent to

(9) $K_Y(y,f^{-1}(I)) \leq K_X(f(y),I)$ for all $y \in S_Y$, $I \in \mathcal{I}_X$.

(c) If also Y is a birth-death process condition (9) can be relaxed in the same manner as Lemma 1 simplifies the monotonicity condition.

LEMMA 3. (In case Y is a birth-death process).
If for all $y \in S_Y$ and components i

(10) $q_Y(y,f^{-1}(f(y) + e_i)) \leq q_X(f(y),f(y) + e_i)$, and

(11) $q_Y(y,f^{-1}(f(y) + e_i)) \leq q_X(f(y),f(y) - e_i)$,

then (9) holds.

PROOF. Verification. □

(d) In Hordijk & Ridder [4] a comparison result for a queueing model with overflow is shown through lemma's 1 and 3.

(e) In a more general setting the collection $\mathcal{I}_X$ of subsets of S_X which determine a stochastic ordering via (1), is arbitrary. Monotonicity is defined by (4) and is not necessarily equivalent to (2) and (3). Still Lemma 2 holds (see Whitt [12]) but the relaxing Lemma's 1 and 3 not. Massey [8] uses the collection of all subsets of the form $I = \{x' \in S_X : x \leq x'\}$ for all $x \in S_X$ to derive a comparison result for an open Jackson network of queues.

4. COMPARISON OF SAMPLE PATHS

A consequence of defining the stochastic ordering through the increasing sets is the following strong comparison result (cf. Kamae et al. [5]). If we assume $\mathbb{P}(f(Y(0)) \in I) \leq \mathbb{P}(X(0) \in I)$ for all $I \in \mathcal{I}_X$ and (i) and (ii) of Lemma 2 hold, then we

can construct processes $\tilde{X}$ and $\tilde{Y}$ defined on the same probability space and with probability laws equal to X resp. Y such that

$$(12) \qquad \mathbb{P}(f(\tilde{Y}(t)) \leq \tilde{X}(t),\ t \geq 0) = 1 \ .$$

This means that we can compare the sample paths of (version of) the processes.

It is sometimes possible to show a result like (12) directly, particularly for some queueing systems which can be viewed as counting processes (cf. Whitt [11], chapter 8 in Ross [9]). In contrast of the analytical method of section 2 and 3 the comparison of the sample paths is shown with heuristic arguments. We give an example to which we shall refer in the next section.

EXAMPLE 1. Consider two queueing systems Y_1 and Y_2 both with the same Poisson arrivals (possibly more than one type) and with all the service times exponentially distributed with the same mean. If y_1 and y_2 are states of the two systems we denote with $f_1(y_1)$ and $f_2(y_2)$ the respective number of customers present in these states. The service disciplines of the systems are arbitrary but satisfy the following condition.

$$(13) \qquad f_2(y_2) \leq f_1(y_1) \Rightarrow q_{Y_2}(y_2, f_2^{-1}(f_2(y_2)-1)) \leq q_{Y_1}(y_1, f_1^{-1}(f_1(y_1)-1)) \ .$$

The condition says that if state y_2 contains equal or less customers than y_1 the rate to the states with one customer less due to a departure is in the Y_2 system equal or less than in Y_1. The proposition then is that we can construct processes $\tilde{Y}_1$ and $\tilde{Y}_2$ with the same probability laws as Y_1 and Y_2 and such that

$$(14) \qquad \mathbb{P}(f_1(\tilde{Y}_1(t)) \leq f_2(\tilde{Y}_2(t)),\ t \geq 0) = 1.$$

The arguments to establish this result are similar to those in Whitt [11]. We shall briefly mention them.

The construction is from arrival epoch to the next arrival epoch. We can construct the sequences of the arrival epochs of the systems such that they are equal. Let $\tilde{T}_n$ and $\tilde{T}_{n+1}$ be two successive arrival epochs and assume the construction up to $\tilde{T}_n$ is succesfully done. Then from $\tilde{T}_n$ on as long as $f_1(\tilde{Y}_1(t)) < f_2(\tilde{Y}_2(t))$ use any construction of departures (consistent with the properties of the systems). As soon as $f_1(\tilde{Y}_1(T)) = f_2(\tilde{Y}_2(T))$ (assume $T \in [\tilde{T}_n, \tilde{T}_{n+1})$) the departures of $\tilde{Y}_2$ are constructed by "thinning out" the departures of $\tilde{Y}_1$. This means that if a departure occurs in $\tilde{Y}_1$ at a time T' then a departure occurs in $\tilde{Y}_2$ at time T' with a probability

equal to $\{q_{Y_2}(y_2, f_2^{-1}(f_2(y_2)-1))\} \cdot \{q_{Y_1}(y_1, f_1^{-1}(f_1(y_1)-1))\}^{-1}$. The thinning procedure is used until again $f_1(\tilde{Y}_1(T')) < f_2(\tilde{Y}_2(T'))$ for some $T' \in (T, \tilde{T}_{n+1})$ (eventually we repeat the construction) or $\tilde{T}_{n+1}$ is attained.

5. APPLICATION TO THE COMMUNICATION MODEL

In this section we want to apply some of the techniques from the previous sections to the model described in the introduction. To be more specific we consider the case of

(i) 3 arriving customer types,

(ii) 2 servers,

(iii) $A_1 = \{1,3\}, A_2 = \{2,3\}$.

This case is of course after the trivial ones the simplest, but it might give ideas how to handle the general case. We refer to it as the model Y whose states contain detailed information, as mentioned in the introduction.

When we are interested in the distribution of the total number of customers - in the sequel we omit for convenience "the distribution of" - or in the throughput c.q. utilization per server, immediately the following model appears to be a candidate for comparison. Suppose each of the two servers has his own queue and that each arriving type 3 customer chooses proportionally to λ_1 and λ_2 to which server he will go. This model is called X and intuitively will yield an upper bound for the total number of customers, because type 3 customers can choose "the wrong server". This means that a type 3 customer who has chosen server 1, must wait when this server is busy, while it is possible that server 2 is free. In model Y server 2 will then start service of this customer. However, the following situation gives a reversed observation. Suppose server 1 is busy with a type 3 customer, a type 1 customer is waiting and server 2 is free, then in the X model this type 3 customer could have chosen server 2 so that both servers are busy. These observations suggest that comparison of the sample paths will fail. Also we did not succeed in proving analytically that X provides an upper bound for the total number of customers.

We shall give several models which provide lower and upper bounds. The proofs are either analytic (i.e. based on section 2 and 3) or heuristic (i.e. based on section 4) but will be omitted because the first ones are tedious verifications of (3) and (9) and the latter ones follow the reasoning in example 1 (with slight deviations).

(a) Lower bounds

L_1 is a M/M/2 system with arrival intensity $\lambda_1 + \lambda_2 + \lambda_3$. L_2 is the original model without type 3 customers, in other words two parallel independent M/M/1 queues with arrival intensities λ_1 and λ_2. Both L_1 and L_2 give lower bounds for the total number of customers and for the utilization per server. L_2 yield also lower bounds for the number of customers per type. L_1 and L_2 are also lower bounds of the following model L_3. Suppose in the original model type 1 and 2 customers have priority. Type 3 customers in service are preempted and resume their service. L_3 can be modelled as a threedimensional birth-death process with the components indicating the number of the three types of customers. It turns out that L_3 is monotone and that comparison with L_1 and L_2 is easy. Intuitively the total number of customers in the original model Y and in the model L_3 will "not differ too much", due to the assumption on the service times. Actually L_3 will "work faster", because a preempted type 3 customer due to the arrival of a type 1, while server 2 is idle, will get immediately service of server 2. So there occur situations in which more servers are working in L_3 than in Y. It is not possible to create a reversed situation. With this in mind it can be shown that (13) holds (putting $Y_1 = L_3$, $Y_2 = Y$), not for all $f_2(y_2) \leq f_1(y_1)$ but for those y_1 and y_2 which can be attained simultaneously in the construction of versions of the processes as done in section 4 and for which $f_2(y_2) \leq f_1(y_1)$. Just as in example 1 it follows then that the total number of customers in L_3 is a lower bound for that in the original model Y.

L_3 is derived from Y by giving priority to type 1 and type 2 customers. If we do the same with the model X we obtain the model X_p. Recall that X is a system of two parallel M/M/1 queues in which the arrival intensities to queue 1 are λ_1 for type 1 and $\frac{\lambda_2}{\lambda_1+\lambda_2}\lambda_3$ for type 3 customers and to queue 2 resp. λ_2 (for type 2) and $\frac{\lambda_1}{\lambda_1+\lambda_2}\lambda_3$ (for type 3). Then X_p can be modelled as a four-dimensional birth-death process with states indicating the number of customers of each type in the two queues. It turns out that L_3 and X_p can be compared in the sense that L_3 is a lower bound for X_p in the number of customers of each type, consequently also in the total number of customers. Because of the assumptions of the service times the number of customers in each of the two queues in the models X and X_p are equal.

Summarizing, for the total number of customers the following "inequalities" hold

$$(15) \quad L_1 \leq L_3 \leq Y \ , \quad L_2 \leq L_3 \leq Y \ ,$$
$$L_3 \leq X_p = X \ .$$

(b) Upper bounds

Similarly we can derive upper bounds by introducing models which are "larger" than Y. Analogously to L_2 and L_3 are the models U_2 (two parallel M/M/1 queues with arrival intensities $\lambda_1+\lambda_3$ resp. $\lambda_2+\lambda_3$) and U_3 (priority to type 1 and type 2 customers and the service to a type 3 customer is stopped if the other server starts serving).

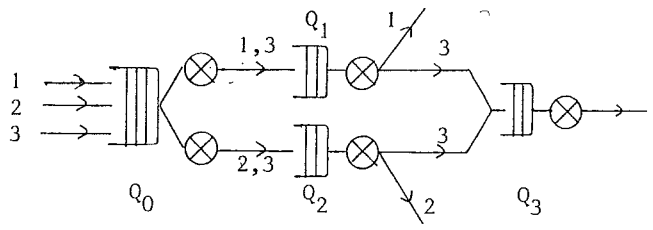
Now we have

$$(16) \qquad Y \le U_3 \le U_2 \quad \text{and} \quad X = X_p \le U_3 \le U_2 \ .$$

We conjectured that X also provides an upper bound for the total number of customers. From (15) and (16) we see that the models Y and X have the same bounds.

6. EXTENSIONS AND CONCLUSIONS

As explained in the introduction the queueing system Y can be a node in a network. Consider the following series of queues.



Q_0 is the queueing system Y described before. Q_i is the destination-node for type i customers, i=1,2,3, and is a "one server FIFO queue".

Replacing Q_0 by the model U_2 of the last section will give upper bounds for the number of customers at Q_0 and for the throughput of each server at Q_0. In this way also the number of customers at each queue will increase. Furthermore, the network becomes a simple product form network and therefore the exact solution is at hand. This sketch roughly indicates that if the original queueing system is part of a network and is replaced by an approximating model, it is possible to derive stochastic inequalities for the network.

Until now we have assumed that all queues have infinite buffer capacity. More complications will arise in queues with finite buffers because the blocking phenomenon appears then. However, we believe that the stochastic inequalities remain valid. Furthermore, in section 5 we considered a special case of the queueing system, namely 3 types of customers and 2 servers, but the results of that section can be extended to more types of customers and servers.
To conclude we want to remark that we presented the system as an illustration of the analytical comparison methods. Furture work will be to give also numerical support for these stochastic inequalities.

REFERENCES

[1] Baskett, F., Chandy, K.M., Muntz, R.R., Palacios, F.G., Open, closed and mixed networks of queues with different classes of customers, J. Ass. Comp. Mach. 22 (1975), 248-260.

[2] Chandy, K.M, Howard, J.H, Towsley,D.F., Product form and local balance in queueing networks, J. Ass. Comp. Mach. 24 (1977), 250-263.

[3] Hordijk, A., Dijk, N.M. Van, Network of queues, Proceedings International Seminar on Modelling and Performance Evaluation Methodology, Lecture Notes in Control and Information Sciences 60(1984), Springer, Berlin, pp.151-205.

[4] Hordijk, A., Ridder, A., Stochastic inequalities for an overflow model, Report University of Leiden, submitted for publication.

[5] Kamae, T, Krengel, U., O'Brien, G.L., Stochastic inequalities on partially ordered spaces, Ann. Prob. 5 (1977), 899-912.

[6] Keilson, J., Kester, A., Monotone matrices and monotone Markov processes, Stoch. Proc. Appl. 5 (1977), 231-241.

[7] Lavenberg, S.S., Computer performance modelling handbook, Academic Press, New York, 1983.

[8] Massey, W.A., An operator-analytic approach to the Jackson network, J. Appl. Prob. 21 (1984), 379-393.

[9] Ross, S.M., Stochastic processes, Wiley, New York, 1983.

[10] Stoyan, D., Comparison methods for queues and other stochastic models, Wiley, New York, 1983.

[11] Whitt, W., Comparing counting processes and queues, Adv.Appl.Prob. 13 (1981), 207-220.

[12] Whitt, W., Stochastic comparisons for non-Markov processes, Submitted for publication.