Performance Evaluation 12 (1991) 5–16
North-Holland

Monotonicity of performance measures in a processor sharing queue

P.R. De Waal
Centre for Mathematics and Computer Science, P.O. Box 4079, 1009 AB Amsterdam, Netherlands

N.M. Van Dijk
Department of Econometrics, Free University, P.O. Box 7161, 1007 MC Amsterdam, Netherlands

Received July 1989
Revised April 1990

Abstract

De Waal, P.R. and N.M. Van Dijk. Monotonicity of performance measures in a processor shearing queue. Performance Evaluation, 12 (1991) 5–16.

In this paper we study the monotonicity of performance measures in a processor sharing queue with two types of customers. The access control law is such that when a new customer arrives he is admitted only if the number of customers of the same type that is already present in the queue does not exceed a predefined threshold. We show that performance measures such as throughput, mean queue length and mean sojourn time are monotonic functions of the threshold for one type if the threshold for the other type is held constant. Monotonicity of throughput and mean queue length is proven by comparing policies for discrete time Markov processes. Monotonicity of the mean sojourn time is proven directly with the closed form formula for this measure using the existence of a product-form equilibrium distribution for the queuing system.

Keywords. Monotonicity, processor sharing.

1. Introduction

During the last few years a number of papers has been published about monotonicity of performance measures in queuing systems. These results are used, for example, to provide structural properties of the queuing systems or bounds on performance measures of analytically untractable queuing systems. The monotonicity results as reported in this paper stem from a study of an optimal control problem for a processor sharing queue. Monotonicity of performance measures here provides necessary and sufficient conditions for the existence of optimal control laws. Furthermore it also suggests an efficient algorithm for finding the optimal control law (cf. [23]).

The techniques that have been used for establishing monotonicity results can roughly be divided into four classes. The first approach is based upon preservation of monotonicity of one-step transition operators (cf. [21]). In analogy with [11, Remark 5.1] one easily finds a counterexample of this preservation under processor sharing disciplines, so this method does not apply.

In the second class of papers monotonicity results are proven for queuing systems that have product-form equilibrium distributions. The closed form formulas of the performance measures that can be derived from these distributions are subsequently used to prove the desired results. In general the proofs are very technical and lack any probabilistic interpretation. Example of papers are [12,15,17,20,22,25].

In the third class of papers monotonicity results are established by stochastic coupling and sample path arguments. In these papers inequalities for the throughput of two related queueing systems are proven by comparing realizations of the arrival or departure processes in these two queuing systems. The inequalities are therefore proven for stochastic variables (i.e. the number of departed customers in a time interval). The drawback of this method is that it relies on the assumption that if the same realization of the arrival process is fed into both queuing systems, the order in which customers are served is the same for both systems. This means that overtaking of customers is not allowed, thus prohibiting the use of this technique in queuing networks with a general routing mechanism or last-come-first-served or processor sharing disciplines. The advantage of the method, when compared to the second technique, is that its use is not restricted to product-form queuing systems, thus allowing blocking for example. Examples of this approach can be found in [1–3,10,18,19].

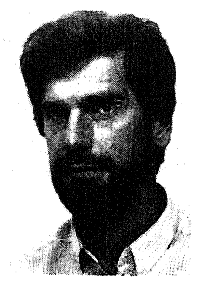
In the fourth category of papers the performance measures are considered as time average rewards for Markov processes. Monotonicity of the measures is then established by proving inequalities for expected rewards over a finite horizon in the discrete time version of these processes. The inequalities thus concern real numbers, viz. expectations of random variables, as opposed to inequalities for random variables as in the second category. Examples are [4,7–9,11,24].

So far, however, no monotonicity results have been reported for systems with processor sharing disciplines. When comparing sample paths for such systems, there are two essential difficulties. Firstly, changing the admission policy for a processor sharing queue may lead to a change in the order in which customers are served. Secondly there is an interaction between the service capacities which are allocated to the different customer types. To this end, the fourth approach will be applied and shown to be successful under natural conditions. The monotonicity results are proven by establishing bounds on the differences of the finite horizon expected rewards. The results are new in the sense that the bounds depend explicitly on the customer's type (cf. Lemma 4.3), and therefore the method itself is of interest.

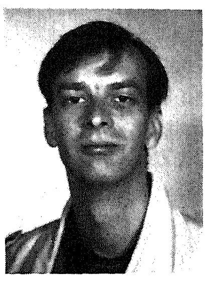
This paper is organized as follows. In Section 2 the queuing system and its performance measures are introduced. The service discipline is defined as a generalization of the standard processor sharing discipline. The queuing process, which is a continuous time Markov process, is transformed into an equivalent discrete time process in Section 3. For this discrete time process monotonicity of the throughput and mean queue length is proven in Sections 4 and 5, respectively. Monotonicity of the mean sojourn time is shown in Section 6.

2. Introduction of the queuing system and its performance measures

Consider the queuing system in Fig. 1, where two types of customers arrive according to two independent Poisson processes with arrival rates λ_1 and λ_2, respectively. If m customers are present of



Nico M. Van Dijk received his M.Sc. and Ph.D. in Applied Mathematics from the University of Leiden, The Netherlands in 1979 and 1983, respectively. Since then he has been with the University of British Columbia, the University of Twente, and the Free University of Amsterdam where he is currently associate professor in the Faculty of Economical Sciences and Econometrics. His current main research interests concern both exact expressions and bounds for queuing networks and their application to various areas such as telecommunication, computer performance evaluation, and flexible manufacturing.



Peter R. de Waal received his M.Sc. in Applied Mathematics from the University of Technology in Eindhoven, the Netherlands in 1985. In 1990 he received his Ph.D. from Tilburg University, the Netherlands. Since 1986 he has been with the Department of Operations Research, Statistics, and System Theory at CWI (Centre for Mathematics and Computer Science) in Amsterdam, the Netherlands. He was Ph.D. research assistant from February 1986 to July 1990 and, since August 1990, he is post-doctoral fellow at CWI.

His current research interests are reliability and availability of networks and optimal stochastic control problems in queuing systems.

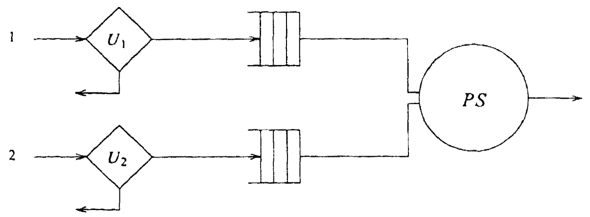


Fig. 1. The model of the processor sharing queue.

type 1 and n customers of type 2, then the population vector of the queue is (m, n), $m, n \in \mathbb{N}$. Admission of new customers is described by a control law $U: \mathbb{N}^2 \to [0, 1]^2$, where $U_i(m, n)$ denotes the probability that a new customer of type i is admitted if the population vector at the moment of arrival is (m, n). Non-admitted customers are assumed to be lost. For the remainder of this paper we will restrict attention to control laws U that use only partial state information, i.e. $U_1(m, n) = U_1(m)$, $U_2(m, n) = U_2(n)$. Furthermore the control law is restricted to be of the *critical level* or *threshold* type, i.e. $U_1(m, n)$ and $U_2(m, n)$ are of the form

$$U_1(m\, n) = 1_{(m<M)}, \tag{2.1}$$

$$U_2(m, n) = 1_{(n<N)}, \tag{2.2}$$

for some $M, N \in \mathbb{N} \cup \{\infty\}$. The control law U as defined in (2.1) and (2.2) will be referred to as $U^{M,N}$. The parameters of the control law, M and N, are referred to as the critical levels or thresholds for type 1 and type 2, respectively. These admission policies make the arrival rates for admitted customers state-dependent, i.e. if the population vector of the queue is (m, n), then the arrival rates of type 1 and 2 are $\lambda_1(m, n) = \lambda_1 1_{(m<M)}$ and $\lambda_2(m, n) = \lambda_2 1_{(n<N)}$, respectively.

The service requirements for customers of type i are assumed to be exponentially distributed with service rate μ_i, $i = 1, 2$. The server is working according to the processor sharing discipline in the sense that at any time each customer of one type receives the same amount of service as any other customer of that type. The speed at which service demands of customers of both types are handled is modeled by two capacity allocation functions $\phi_1, \phi_2: \mathbb{N}^2 \to [0, 1]$. For each $m, n \in \mathbb{N}$, $\phi_i(m, n)$ denotes the speed at which all customers of type i together are served, so $\mu_i\phi_i(m, n)$ is the actual service rate for type i if the population vector is (m, n). The actual service rate for one customer of type 1 and type 2 is then $\mu_1\phi_1(m, n)/m$ and $\mu_2\phi_2(m, n)/n$, respectively.

Let $X: \Omega \times \mathbb{R}_+ \to \mathbb{N}^2$ denote the queuing process, for some appropriately chosen sample space Ω. Performance measures for this queuing system are defined as expected time-average rewards for suitably chosen reward functions. Let $r: \mathbb{N}^2 \to \mathbb{R}_+$ be a reward function: when the population vector is (m, n), a reward $r(m, n)$ is accrued per unit time. The corresponding performance measure is defined as the expected time-average reward if the reward function is r:

$$\overline{\lim_{t\to\infty}}\, E^{M,N}\left\{\frac{1}{t}\int_0^t r(X_s)\, \mathrm{d}s\right\},$$

where the superscript M, N on the expectation operator denotes its dependence on the thresholds of the control law. Most standard performance measures can be expressed in this manner, e.g. the throughput of type i by choosing $r(m, n) = \mu_i\phi_i(m, n)$, and the total mean queue length by choosing $r(m, n) = m + n$.

3. Transformation to discrete time

In this section the queuing system, as introduced in the previous section, will be transformed into a discrete time setting. This formulation will be useful in later sections since it makes times between transitions of the queuing process constant, at the cost of introducing so-called *dummy transitions*. The transformation proceeds as follows (cf. [14,16]).

Assume that the sum of all transitions rates, i.e. $\lambda_1 + \lambda_2 + \mu_1 + \mu_2$, is finite. Furthermore assume that this sum is equal to one. This is no restriction, since it can be established by appropriately scaling the time axis. Introduce $N: \Omega \times \mathbb{R}_+ \to \mathbb{N}$ a Poisson process with stochastic intensity one, and the stopping times $t_n: \Omega \to \mathbb{R}$ as

$$t_n = \inf\{t \mid N_t \geqslant n\}, \quad n \in \mathbb{N}.$$

Let $Y: \Omega \times \mathbb{N} \to \mathbb{N}^2$ be defined as $Y_n := X_{t_n}$, then Y is a discrete time Markov process with transition probabilities (depending on the control law) given by:

$$p^{M,N}((m', n'); (m, n)) = \mathbb{P}\{Y_{k+1} = (m', n') \mid Y_k = (m, n)\}$$
$$= \begin{cases} \lambda_1 1_{(m<M)}, & \text{if } (m', n') = (m+1, n), \\ \lambda_2 1_{(n<N)}, & \text{if } (m', n') = (m, n+1), \\ \mu_1\phi_1(m, n), & \text{if } (m', n') = (m-1, n), \\ \mu_2\phi_2(m, n), & \text{if } (m', n') = (m, n-1), \\ 1 - \lambda_1 1_{(m<M)} - \lambda_2 1_{(n<N)} - \mu_1\phi_1(m, n) - \mu_2\phi_2(m, n), & \\ & \text{if } (m', n') = (m, n). \end{cases} \tag{3.1}$$

From (3.1) we see that by transforming the continuous time process into a discrete time one, we introduce dummy transitions, i.e. transitions that do not change the state.

If we choose the one step reward for process Y equal to $r(m, n)$, then the expected time-average rewards for both Y and X are equal, i.e.

$$\varliminf_{t\to\infty} E^{M,N}\left\{\frac{1}{t}\int_0^t r(X_s)\,\mathrm{d}s\right\} = \varliminf_{k\to\infty} E^{M,N}\left\{\frac{1}{k}\sum_{n=0}^{k-1} r(Y_n)\right\}.$$

4. Monotonicity of the throughput function

In this section we show that the throughput of customers is monotonic in the thresholds: the throughput of one type increases if the threshold for that type is increased, and it decreases if the threshold for the other type is increased.

Let the reward function be $r(m, n) = \mu_1\phi_1(m, n)$, hence the performance measure under consideration is the throughput of type 1 customers. Let $T_1^{M,N}$ and $T_1^{M+1,N}$ denote the throughput of type 1 if the control laws $U^{M,N}$ and $U^{M+1,N}$ are used, respectively. Throughout the remaining sections of this paper we assume the following.

Assumption 4.1. $\phi_1(m, n)$ is non-decreasing in m and non-increasing in n. $\phi_2(m, n)$ is non-increasing in m and non-decreasing in n.

These assumptions are satisfied for example by the normal processor sharing service discipline, i.e. $\phi_1(m, n) = 1_{(m+n>0)} m/(m+n)$.

The main result of this section is the following theorem. It states the intuitively obvious monotonicity for the throughput function.

Theorem 4.2. *If Assumption* 4.1 *holds and if*

$$\phi_1(m+1, n) + \phi_2(m+1, n) \geqslant \phi_1(m, n) + \phi_2(m, n), \tag{4.1a}$$
$$\phi_1(m, n+1) + \phi_2(m, n+1) \geqslant \phi_1(m, n) + \phi_2(m, n), \tag{4.1b}$$

then $T_1^{M,N} \leqslant T_1^{M+1,N}$, $M, N \in \mathbb{N}$.

Before proceeding with the proof of Theorem 4.2 we need the following definitions. Let $V^k_{M,N}(m, n)$ for $m, n \in \mathbb{N}$ denote the total expected reward over k steps when starting in state (m, n), for the policy $U^{M,N}$:

$$V^k_{M,N}(m, n) = E^{M,N}\left[\sum_{i=0}^{k-1} r(Y_i) \mid Y_0 = (m, n)\right], \quad k \geqslant 0, \tag{4.2}$$

where $V^0_{M,N}(m, n) = 0$, $m, n \in \mathbb{N}$.

Since for finite M and N the state (0, 0) is positive recurrent, the Markov chain Y is irreducible. According to the theory of Markov reward processes we thus have

$$T_1^{M,N} = \lim_{k\to\infty} \frac{1}{k} V^k_{M,N}(m, n)$$

and the limit is independent of the initial state (m, n). It is therefore sufficient to prove that for some $m, n \in \mathbb{N}$ and for all $k \in \mathbb{N}$,

$$V^k_{M,N}(m, n) \leqslant V^k_{M+1,N}(m, n).$$

For the proof of Theorem 4.2 we first state the following lemma.

Lemma 4.3. *If Assumption* 4.1 *holds and the conditions* (4.1) *are satisfied, then for all* $k, m, n, M, N \in \mathbb{N}$,

$$0 \leqslant V^k_{M,N}(m+1, n) - V^k_{M,N}(m, n) \leqslant 1, \tag{4.3a}$$

$$0 \leqslant V^k_{M,N}(m, n) - V^k_{M,N}(m, n+1) \leqslant \frac{\mu_1}{\mu_2}. \tag{4.3b}$$

Proof. The proof is by induction in k. Since M and N are constant throughout the proof, we shall drop the subscript M, N from $V^k_{M,N}$.

By definition $V^0(m, n) = 0$, so (4.3) is immediate for $k = 0$. Let $k > 0$. Assume that (4.3) is satisfied for k. From (4.2) we get the following recursion for V^{k+1}:

$$\begin{aligned}
&V^{k+1}(m+1, n) - V^{k+1}(m, n)\\
&\quad = \big\{r(m+1, n)\\
&\qquad +\lambda_1 1_{(m+1<M)} V^k(m+2, n)\\
&\qquad +\lambda_2 1_{(n<N)} V^k(m+1, n+1)\\
&\qquad +\mu_1\phi_1(m+1, n) V^k(m, n)\\
&\qquad +\mu_2\phi_2(m+1, n) V^k(m+1, n-1)\\
&\qquad +\left[1 - \lambda_1 1_{(m+1<M)} - \lambda_2 1_{(n<N)} - \mu_1\phi_1(m+1, n) - \mu_2\phi_2(m+1, n)\right] V^k(m+1, n)\big\}\\
&\quad -\big\{r(m, n)\\
&\qquad +\lambda_1 1_{(m<M)} V^k(m+1, n)\\
&\qquad +\lambda_2 1_{(n<N)} V^k(m, n+1)\\
&\qquad +\mu_1\phi_1(m, n) V^k(m-1, n)\\
&\qquad +\mu_2\phi_2(m, n) V^k(m, n-1)\\
&\qquad +\left[1 - \lambda_1 1_{(m<M)} - \lambda_2 1_{(n<N)} - \mu_1\phi_1(m, n) - \mu_2\phi_2(m, n)\right] V^k(m, n)\big\}.
\end{aligned}$$

Consider the two expressions between braces for $V^k(m+1, n)$ and $V^k(m, n)$, respectively. The second term of $V^k(m, n)$ can be rewritten as

$$\lambda_1\left[1_{(m+1<M)} + 1_{(m+1=M)}\right] V^k(m+1, n),$$

the fourth term of $V^k(m+1, n)$ as

$$\mu_1[\phi_1(m, n) + (\phi_1(m+1, n) - \phi_1(m, n))] V^k(m, n),$$

the fifth term of $V^k(m, n)$ as

$$\mu_2[\phi_2(m+1, n) + (\phi_2(m, n) - \phi_2(m+1, n))] V_k(m, n-1),$$

the last term of $V^k(m+1, n)$ as

$$\begin{aligned}&[1 - \lambda_1 1_{(m<M)} - \lambda_2 1_{(n<N)} - \mu_1\phi_1(m+1, n) - \mu_2\phi_2(m, n)] V^k(m+1, n)\\ &\quad + \lambda_1 1_{(m+1=M)} V^k(m+1, n)\\ &\quad + \mu_2[\phi_2(m, n) - \phi_2(m+1, n)] V^k(m+1, n),\end{aligned}$$

and the last term of $V^k(m, n)$ as

$$\begin{aligned}&[1 - \lambda_1 1_{(m<M)} - \lambda_2 1_{(n<N)} - \mu_1\phi_1(m+1, n) - \mu_2\phi_2(m, n)] V^k(m, n)\\ &\quad \mu_1[\phi_1(m+1, n) - \phi_1(m, n)] V^k(m, n).\end{aligned}$$

By combining the corresponding terms from $V^k(m+1, n)$ and $V^k(m, n)$ we get

$$\begin{aligned}&V^{k+1}(m+1, n) - V^{k+1}(m, n) && (4.4)\\ &\quad = r(m+1, n) - r(m, n) && \circ\\ &\qquad + \lambda_1 1_{(m+1<M)}[V^k(m+2, n) - V^k(m+1, n)]\\ &\qquad + \lambda_2 1_{(n<N)}[V^k(m+1, n+1) - V^k(m, n+1)]\\ &\qquad + \mu_1\phi_1(m, n)[V^k(m, n) - V^k(m-1, n)]\\ &\qquad + \mu_2\phi_2(m+1, n)[V^k(m+1, n-1) - V^k(m, n-1)]\\ &\qquad + [1 - \lambda_1 1_{(m<M)} - \lambda_2 1_{(n<N)} - \mu_1\phi_1(m+1, n) - \mu_2\phi_2(m, n)]\\ &\qquad\quad \times [V^k(m+1, n) - V^k(m, n)]\\ &\qquad + \mu_2[\phi_2(m, n) - \phi_2(m+1, n)][V^k(m+1, n) - V^k(m, n)]\\ &\qquad + \mu_2[\phi_2(m, n) - \phi_2(m+1, n)][V^k(m, n) - V^k(m, n-1)] && \circ\end{aligned}$$

Observe that apart from the terms marked $\circ$ all other terms consist of a difference as in (4.3a) multiplied by a constant between 0 and 1, with all multiplication constants summing up to a number smaller than or equal to one. Note also that by Assumption 4.1 and (4.3b) the second term marked $\circ$ is negative. We will now bring the marked terms in a similar form, distinguishing between two cases.

Case $\phi_1(m+1, n) = \phi_1(m, n)$. If $\phi_1(m+1, n) = \phi_1(m, n)$ then by Assumption 4.1 and condition (4.1a) also $\phi_2(m+1, n) = \phi_2(m, n)$ and by definition $r(m+1, n) = r(m, n)$. Both terms marked $\circ$ thus vanish, and all the remaining terms are of the form (4.3a) multiplied by a non-negative constant. Furthermore all multiplication constants sum op to $1 - \lambda_1 1_{(m+1=M)}$, which is smaller than or equal to one, thus completing the proof.

Case $\phi_1(m+1, n) > \phi_1(m, n)$. Assume that $\phi_1(m+1, n) > \phi_1(m, n)$. The $\circ$ marked terms in (4.4) equal

$$\begin{aligned}&r(m+1, n) - r(m, n) + \mu_2[\phi_2(m, n) - \phi_2(m+1, n)][V^k(m, n) - V^k(m, n-1)]\\ &\quad = \mu_1[\phi_1(m+1, n) - \phi_1(m, n)]\\ &\qquad \times \left\{1 + \frac{\mu_2[\phi_2(m, n) - \phi_2(m+1, n)]}{\mu_1[\phi_1(m+1, n) - \phi_1(m, n)]}[V^k(m, n) - V^k(m, n-1)]\right\}.\end{aligned}$$

With (4.1a) and the induction assumption (4.3b) we have

$$0 \leqslant 1+\frac{\mu_2[\phi_2(m, n)-\phi_2(m+1, n)]}{\mu_1[\phi_1(m+1, n)-\phi_1(m, n)]}[V^k(m, n)-V^k(m, n-1)] \leqslant 1. \tag{4.5}$$

so

$$\begin{aligned} & V^{k+1}(m+1, n)-V^{k+1}(m, n) \\ & \quad =\lambda_1 1_{(m+1<M)}[V^k(m+2, n)-V^k(m+1, n)] \\ & \qquad +\lambda_2 1_{(n<N)}[V^k(m+1, n+1)-V^k(m, n+1)] \\ & \qquad +\mu_1 \phi_1(m, n)[V^k(m, n)-V^k(m-1, n)] \\ & \qquad +\mu_2 \phi_2(m+1, n)[V^k(m+1, n-1)-V^k(m, n-1)] \\ & \qquad +\mu_1[\phi_1(m+1, n)-\phi_1(m, n)][\text{TERM (4.5)}] \\ & \qquad +\mu_2[\phi_2(m, n)-\phi_2(m+1, n)][V^k(m+1, n)-V^k(m, n)] \\ & \qquad +[1-\lambda_1 1_{(m<M)}-\lambda_2 1_{(n<N)}-\mu_1 \phi_1(m+1, n)-\mu_2 \phi_2(m, n)] \\ & \qquad \quad \times[V^k(m+1, n)-V^k(m, n)], \end{aligned}$$

where $0 \leqslant$ TERM (4.5) $\leqslant 1$, thus completing the induction step. The proof of (4.3b) proceeds in an analogous way. □

With Lemma 4.3 we can proceed the proof of Theorem 4.2.

Proof (of Theorem 4.2). Recall that we had to prove $V_{M,N}^k(m, n) \leqslant V_{M+1,N}^k(m, n)$ for some $m, n \in \mathbb{N}$ and for all $k \in \mathbb{N}$. In fact we will prove that this holds for all $m, n \in \mathbb{N}$. Again the proof is by induction in k.

$k=0$. Trivial.

$k>0$. Observe that

$$\begin{aligned} & V_{M+1,N}^{k+1}(m, n)-V_{M,N}^{k+1}(m, n) \\ & \quad =\sum_{(m', n')} p^{M+1,N}((m', n'); (m, n)) V_{M+1,N}^k(m', n') \\ & \qquad -p^{M,N}((m', n'); (m, n)) V_{M,N}^k(m', n') \\ & \quad =\sum_{(m', n')} p^{M,N}((m', n'); (m, n))[V_{M+1,N}^k(m', n')-V_{M,N}^k(m', n')] \\ & \qquad +\sum_{(m', n')}[p^{M+1,N}((m', n'); (m, n))-p^{M,N}((m', n'); (m, n))] V_{M+1,N}^k(m', n'). \end{aligned} \tag{4.6}$$

Since the first term on the right hand side of (4.6) is positive due to the induction assumption, the positivity of the second term remains to be proven. Examination of this term shows that it is equal to $\lambda_1[V_{M+1,N}^k(M+1, n)-V_{M+1,N}^k(M, n)]$. With the result of Lemma 4.3 this completes the proof. □

Observe that by equation (4.6) proving the original inequality for two policies is reduced to proving an inequality for one policy. This is due to the fact that the control law for customers of type 2 is the same for both $U^{M,N}$ and $U^{M+1,N}$.

By choosing the appropriate bounds as in Lemma 4.3 the method should in principle be extendible to more than two customer types. We have, however, not addressed this problem yet.

The introduction of the capacity allocation functions allows more elaborate service disciplines than the usual Processor Sharing mechanism, including those which do not lead to product-form equilibrium probabilities.

Example 4.4 *(Monotonic Generalized Processor Sharing).* If we choose the capacity allocation function as in the Generalized Processor Sharing model (cf. [6,13], i.e.

$$\phi_1(m, n) = f(m+n)\frac{m}{m+n}1_{(m+n>0)}, \qquad \phi_2(m, n) = f(m+n)\frac{n}{m+n}1_{(m+n>0)},$$

for some non-decreasing function $f: \mathbb{N} \to \mathbb{R}_+$, we have $\phi_1(m, n) + \phi_2(m, n)$ equal to $f(m+n)1_{(m+n>0)}$, so Assumption 4.1 and the conditions (4.1) are satisfied. Moreover, observe that for this choice the equilibrium probability distribution has a product-form (cf. [5,6,13]) and the throughput depends on the service time distribution only through its mean value. This leads to the following corollary.

Corollary 4.5. *If the service discipline is Monotonic Generalized Processor Sharing, then the throughput of type* 1 *customer is non-decreasing in the threshold of type* 1 *for general service time distributions.*

The following example shows that Theorem 4.2 holds also for non-standard service disciplines.

Example 4.6 *(Priority Processor Sharing).*
Take

$$\phi_1(m, n) = 1_{(m>0)}, \qquad \phi_2(m, n) = 1_{(m=0)}.$$

Here we have $\phi_1(m, n) + \phi_2(m, n) = 1$, so conditions (4.3) are satisfied. In this example all processor capacity is awarded to type 1 customers, when they are present. This queue does not have a product-form equilibrium distribution.

We conclude this section with a theorem similar to Theorem 4.2 referring to the monotonicity of the throughput of one type of customers if the threshold of the other customer type is increased.

Theorem 4.7. *If the conditions of* (4.1) *are satisfied, then* $T_2^{M,N} \geqslant T_2^{M+1,N}$.

Proof. Take $r(m, n) = \mu_2\phi_2(m, n)$. The proof now proceeds analogously to that of Theorem 4.2. □

5. Monotonicity of the mean queue length

In this section monotonicity of the mean queue length is shown with respect to both thresholds. If either of the thresholds is increased, then the mean queue length of both types (and the total mean queue length of course) increases. The result can be stated for a rather general class of reward functions.

Theorem 5.1. *Let* $V_{M,N}^k(m, n)$ *be the expected total reward over k steps, when starting in state* (m, n) *and using policy* $U^{M,N}$. *If the reward function* $r(m, n)$ *is non-decreasing in both its arguments m and n and Assumption* 4.1 *holds, then* $V_{M,N}^k(m, n) \leqslant V_{M+1,N}^k(m, n)$ *for all* $m, n \in \mathbb{N}$.

Proof. As in the proof of Theorem 4.2 it is plain from (4.6) that it suffices to proves that

$$0 \leqslant V_{M,N}^k(m+1, n) - V_{M,N}^k(m, n), \quad m, n, M, N \in \mathbb{N}. \quad \square$$

Lemma 5.2. *If* $r(m, n)$ *is non-decreasing in both arguments, then*

$$0 \leqslant V_{M,N}^k(m+1, n) - V_{M,N}^k(m, n), \tag{5.1a}$$

$$0 \leqslant V_{M,N}^k(m, n+1) - V_{M,N}^k(m, n), \tag{5.1b}$$

for all $m, n \in \mathbb{N}$.

Proof. Again the proof is by induction in k and the subscript M, N is dropped from $V^k_{M,N}$. The case $k=0$ is trivial. For $k>0$ write $V^{k+1}(m+1, n) - V^{k+1}(m, n)$ as

$$\begin{aligned}
&V^{k+1}(m+1, n) - V^{k+1}(m, n)\\
&\quad = r(m+1, n) - r(m, n)\\
&\qquad + \lambda_1 1_{(m+1<M)}\left[V^k(m+2, n) - V^k(m+1, n)\right]\\
&\qquad + \lambda_2 1_{(n<N)}\left[V^k(m+1, n+1) - V^k(m, n+1)\right]\\
&\qquad + \mu_1\phi_1(m, n)\left[V^k(m, n) - V^k(m-1, n)\right]\\
&\qquad + \mu_2\phi_2(m+1, n)\left[V^k(m+1, n-1) - V^k(m, n-1)\right]\\
&\qquad + \left[1 - \lambda_1 1_{(m<M)} - \lambda_2 1_{(n<N)} - \mu_1\phi_1(m+1, n) - \mu_2\phi_2(m, n)\right]\\
&\qquad\quad \times\left[V^k(m+1, n) - V^k(m, n)\right]\\
&\qquad + \mu_2\left[\phi_2(m, n) - \phi_2(m+1, n)\right]\left[V^k(m+1, n) - V^k(m, n-1)\right].
\end{aligned}\tag{5.2}$$

Since $\phi_2(m, n)$ is non-increasing in m due to Assumption 4.1 and $V^k(m+1, n) - V^k(m, n-1) = V^k(m+1, n) - V^k(m, n) + V^k(m, n) - V^k(m, n-1) \geqslant 0$, all terms in (5.2) are non-negative, thus completing the induction step. The proof of (5.1b) is analogous. □

Note that if the reward function r is positive and non-decreasing in both arguments, then r^i also has these properties for all $i \in \mathbb{N}$. The monotonicity of r thus makes the higher moments also monotonic in the thresholds.

Example 5.3. If we take $r(m, n) = m$ or $r(m, n) = m+n$, we see that both the mean queue length of type 1 and the total mean queue length are non-decreasing if the threshold of type 1 is increased. Due to the symmetry of Theorem 5.1 this also holds if the threshold of type 2 is increased.

6. Monotonicity of the mean sojourn time

With Little's result we can combine Theorems 4.2 and 5.1 to show that the mean sojourn time of type 2 customers increases if the threshold for type 1 customers increases. Unfortunately this technique cannot be used for the monotonicity of the type 1 customers' mean sojourn time. In this section we show that this performance measure is also non-decreasing if either of the thresholds is increased. We have been able to prove this only for standard Processor Sharing, however. The capacity allocation functions ϕ_1 and ϕ_2 thus are chosen as

$$\phi_1(m, n) = \frac{m}{m+n} 1_{(m+n>0)}, \qquad \phi_2(m, n) = \frac{n}{m+n} 1_{(m+n>0)}.$$

The proof of the theorem relies on the product-form of the equilibrium distribution and the closed form formula of the mean sojourn time that can be derived from this distribution. It is well known (cf. [5,13]) that the equilibrium probability of the population vector (m, n) under the policy $U^{M,N}$ is equal to

$$\pi^{M,N}(m, n) = C[M, N]\pi(m, n), \quad 0 \leqslant m \leqslant M, 0 \leqslant n \leqslant N, \tag{6.1}$$

where

$$\pi(m, n) = \binom{m+n}{n}\rho_1^m\rho_2^n, \quad 0 \leqslant m \leqslant M, 0 \leqslant n \leqslant N, \tag{6.2}$$

$$C[M, N] = \left[\sum_{m=0}^{M}\sum_{n=0}^{N}\binom{m+n}{n}\rho_1^m\rho_2^n\right]^{-1}. \tag{6.3}$$

Here $\rho_i := \lambda_i/\mu_i$ denotes the workload of type i, $i = 1, 2$. By Little's formula the mean sojourn time of type 1 customers is equal to

$$S_1^{M,N} = \frac{A(M)}{B(M)} \tag{6.4}$$

where

$$A(M) = \sum_{m=0}^{M} \sum_{n=0}^{N} m\pi(m, n), \tag{6.5}$$

$$B(M) = \sum_{m=0}^{M} \sum_{n=0}^{N} \frac{m}{m+n}\mu_1\pi(m, n). \tag{6.6}$$

The variable N is suppressed in the notation of A and B, since N is held constant throughout this section.

Theorem 6.1. *If* $\rho_2 > 0$, *then*

$$S_1^{M,N} < S_1^{M+1,N}, \tag{6.7}$$

Again we need a preliminary lemma for the proof of the theorem.

Lemma 6.2. *If* $A, B : \mathbb{N} \to \mathbb{R}_+$, *increasing*, $A(0) = B(0) = 0$ *and*

$$\frac{\Delta A(M+1)}{\Delta B(M+1)} > \frac{\Delta A(M)}{\Delta B(M)}, \quad M \geqslant 1, \tag{6.8}$$

where $\Delta A(M) := A(M) - A(M-1)$ *and* $\Delta B(M)$ *analogously, then*

$$\frac{A(M+1)}{B(M+1)} > \frac{A(M)}{B(M)}, \quad M \geqslant 1. \tag{6.9}$$

Proof (by induction).

($M = 1$). First note that for positive a, b, c, d,

$$\frac{a+c}{b+d} > \frac{a}{b} \Leftrightarrow \frac{c}{d} > \frac{a}{b} \Leftrightarrow \frac{c}{d} > \frac{a+c}{b+d}. \tag{6.10}$$

Equation (6.9) reads for $M = 1$

$$\frac{A(2)}{B(2)} > \frac{A(1)}{B(1)}.$$

Note that since $A(0) = B(0) = 0$, by definition $A(1) = \Delta A(1)$, $B(1) = \Delta B(1)$. Since $A(2) = A(1) + \Delta A(2)$ and $B(2) = B(1) + \Delta B(2)$, this holds if and only if

$$\frac{A(1) + \Delta A(2)}{B(1) + \Delta B(2)} > \frac{A(1)}{B(1)} \Leftrightarrow \frac{\Delta A(2)}{\Delta B(2)} > \frac{A(1)}{B(1)}.$$

The latter inequality is true by $A(1) = \Delta A(1)$ and $B(1) = \Delta B(1)$.

($M > 0$). Assume that we have

$$\frac{A(m+1)}{B(m+1)} > \frac{A(m)}{B(m)}, \quad m = 1, \ldots, M-1.$$

For $m = M - 1$ this yields

$$\frac{A(M)}{B(M)} > \frac{A(M-1)}{B(M-1)} \Leftrightarrow \frac{A(M-1)+\Delta A(M)}{B(M-1)+\Delta B(M)} > \frac{A(M-1)}{B(M-1)}$$

$$\Leftrightarrow \frac{\Delta A(M)}{\Delta B(M)} > \frac{A(M-1)+\Delta A(M)}{B(M-1)+\Delta B(M)}$$

$$\Leftrightarrow \frac{\Delta A(M)}{\Delta B(M)} > \frac{A(M)}{B(M)},$$

where the second equivalence is an application of (6.10). With the condition (6.8) this implies

$$\frac{\Delta A(M+1)}{\Delta B(M+1)} > \frac{A(M)}{B(M)} \Leftrightarrow \frac{A(M)+\Delta A(M+1)}{B(M)+\Delta B(M+1)} > \frac{A(M)}{B(M)}$$

$$\Leftrightarrow \frac{A(M+1)}{B(M+1)} > \frac{A(M)}{B(M)}. \qquad \square$$

Proof (of Theorem 6.1). According to Lemma 6.2 it suffices to prove that

$$\frac{\Delta A(M+1)}{\Delta B(M+1)} > \frac{\Delta A(M)}{\Delta B(M)}.$$

This is equivalent to

$$\frac{\sum_{n=0}^{N} \binom{M+1+n}{n} \rho_2^n}{\sum_{n=0}^{N} \frac{1}{M+1+n} \binom{M+1+n}{n} \rho_2^n} > \frac{\sum_{n=0}^{N} \binom{M+n}{n} \rho_2^n}{\sum_{n=0}^{N} \frac{1}{M+n} \binom{M+n}{n} \rho_2^n}. \tag{6.11}$$

If we define $w_n = (M+n)! \rho_2^n / n!$, then (6.11) reads

$$\left(\sum_{n=0}^{N} (M+n+1) w_n \right) \left(\sum_{n=0}^{N} \frac{1}{M+n} w_n \right) > \left(\sum_{n=0}^{N} w_n \right)^2 \tag{6.12}$$

$$\Leftrightarrow \sum_{n=0}^{N} \frac{M+n+1}{M+n} w_n^2 + \sum\sum_{i<j} \left(\frac{M+i+1}{M+j} + \frac{M+j+1}{M+i} \right) w_i w_j > \sum_{n=0}^{N} w_n^2 + 2 \sum\sum_{i<j} w_i w_j. \tag{6.13}$$

The first terms on both sides of (6.13) can be compared immediately and yield the desired inequality. The second terms also satisfy this inequality since

$$(M+i+1)(M+i) + (M+j+1)(M+j) \geqslant 2(M+i)(M+j)$$

$$\Leftrightarrow (M+i)(i-j+1) + (M+j)(j-i+1) > 0$$

$$\Leftrightarrow (i-j)^2 + 2M + i + j > 0. \qquad \square$$

Acknowledgement

The authors are indebted to Onno Boxma for suggesting a much simpler proof of Theorem 6.1.

References

[1] I.J.B.F. Adan and J. van der Wal, Monotonicity of the throughput of a closed queueing network in the number of jobs, Memorandum COSOR 87-03, Dept. Math. Comp. Science, Eindhoven University of Technology, Eindhoven, 1987.

[2] I.J.B.F. Adan and J. van der Wal, Monotonicity of the throughput of an open queueing network in the inter-arrival and service times, Memorandum COSOR 87-05, Dept. Math. Comp. Science, Eindhoven University of Technology, Eindhoven, 1987.

[3] I.J.B.F. Adan and J. van der Wal, Monotonicity properties of the throughput of an open finite capacity queueing network, Memorandum COSOR 87-12, Dept. Math. Comp. Science, Eindhoven University of Technology, Eindhoven, 1987.

[4] I.J.B.F. Adan and J. van der Wal, Monotonicity of the throughput of a closed Erlang queueing network in the number of jobs, Memorandum COSOR 87-01, Dept. Math. Comp. Science, Eindhoven University of Technology, Eindhoven, 1987.

[5] F. Baskett, K.M. Chandy, R.R. Muntz and F.G. Palacios, Open, closed, and mixed networks of queues with different classes of customers, *J. ACM.* **22** (1975) 248–260.

[6] J.W. Cohen, The multiple phase service network with generalized processor sharing, *Acta Inform.* **12** (1979) 245–284.

[7] N.M. van Dijk, A formal proof for the insensitivity of simple bounds for finite multiserver non-exponential tandem queues based on monotonicity results, *Stochast. Process. Appl.* **27** (1988) 261–277.

[8] N.M. van Dijk, Simple bounds for queueing systems with breakdowns, *Perform. Eval.* **8** (1988) 117–128.

[9] N.M. van Dijk and B.F. Lamond, Simple bounds for finite single-server exponential tandem queues, *Oper. Res.* **36** (1988) 470–477.

[10] N.M. van Dijk, P. Tsoucas and J. Walrand, Simple bounds for the call congestion of finite multi-server delay systems, *Probab. Engin. Inform. Sci.* **2** (1988) 129–138.

[11] N.M. van Dijk and J. van der Wal, Simple bounds and monotonicity results for multi-server exponential tandem queues, *Queueing Systems*, to appear.

[12] J.B.M. van Doremalen and P.R. de Waal, An approximation method for closed queueing networks with two-phase servers, Memorandum COSOR 85-15, Dept. Math. Comp. Science, Eindhoven University of Technology, Eindhoven, 1985.

[13] F.P. Kelly, *Reversibility and Stochastic Networks* (Wiley, New York, 1979).

[14] S.A. Lippman, Applying a new device in the optimization of exponential queueing systems, *Oper. Res.* **23** (1975) 687–710.

[15] T.G. Robertazzi and A.A. Lazar, On the modeling and optimal flow control of the Jacksonian network, *Perform. Eval.* **5** (1985) 29–43.

[16] R.F. Serfozo, An equivalence between continuous and discrete time Markov decision processes, *Oper. Res.* **27** (1979) 616–620.

[17] J.G. Shanthikumar and D.D. Yao, The effect of increasing service rates in a closed queueing network, *J. Appl. Probab.* **23** (1986) 474–483.

[18] J.G. Shanthikumar and D.D. Yao, Optimal server allocation in a system of multi-server stations, *Manag. Sci.* **33** (1987) 1173–1180.

[19] J.G. Shanthikumar and D.D. Yao, General queueing networks: Representation and stochastic monotonicity, Research Report, University of California, Berkeley, 1987.

[20] J.G. Shanthikumar and D.D. Yao, Stochastic monotonicity of the queue lengths in closed queueing networks, *Oper. Res.* **35** (1987) 583–588.

[21] D. Stoyan, *Comparison Methods for Queues and Other Stochastic Models* (Wiley, New York, 1983).

[22] R. Suri, A concept of monotonicity and its characterization for closed queueing networks, *Oper. Res.* **33** (1985) 606–624.

[23] P.R. de Waal, Overload control of telephone exchanges, Ph.D. Thesis, Centre for Mathematics and Computer Science, Amsterdam, 1990.

[24] J. van der Wal, Monotonicity of the throughput of a closed exponential single chain queueing network in the number of jobs, Memorandum COSOR 85-21, Dept. Math. Comp. Science, Eindhoven University of Technology, Eindhoven, 1985.

[25] D.D. Yao, Some properties of the throughput function of closed networks of queues, *Oper. Res. Lett.* **3** (1985) 313–317.