Monotonicity properties for multi-class queueing systems*

I.M. Verloop[1], U. Ayesta[2,3], S.C. Borst[1,4,5]

[1]CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
[2]LAAS-CNRS, Université de Toulouse, 7 Avenue Colonel Roche,
31077 Toulouse Cedex, France
[3]BCAM - Basque Center for Applied Mathematics,
Bizkaia Technology Park, 48170 Zamudio, Spain
[4]Bell Laboratories, Alcatel-Lucent, P.O. Box 636, Murray Hill, NJ 07974, USA
[5]Department of Mathematics & Computer Science, Eindhoven University of Technology
P.O. Box 513, 5600 MB Eindhoven, The Netherlands

Abstract

We study multi-dimensional stochastic processes that arise in queueing models used in the performance evaluation of wired and wireless networks. The evolution of the stochastic process is determined by the scheduling policy used in the associated queueing network. For general arrival and service processes, we give sufficient conditions in order to compare sample-path wise the workload and the number of users under different policies. This allows us to evaluate the performance of the system under various policies in terms of stability, the mean overall delay and the mean holding cost.

We apply the general framework to linear networks, where users of one class require service from several shared resources simultaneously. For the important family of weighted α-fair policies, stability results are derived and monotonicity of the mean holding cost with respect to the fairness parameter α and the relative weights is established. In order to broaden the comparison results, we investigate a heavy-traffic regime and perform numerical experiments. In addition, we study a single-server queue with two user classes, and show that under Discriminatory Processor Sharing (DPS) or Generalized Processor Sharing (GPS) the mean overall sojourn time is monotone with respect to the ratio of the weights. Finally we extend the framework to obtain comparison results that cover the single-server queue with an arbitrary number of classes as well.

1 Introduction

In recent years a lot of attention has been devoted to multi-class stochastic networks where the capacity allocated to the various classes depends on the number of users present in all classes. Analyzing multi-class stochastic systems tends to be very challenging. Metrics like the joint (marginal) distribution of the number of users of the various classes, or even the mean number of users of the various classes, can only be determined in some special cases. In order to gain insight into the performance of the system, researchers have therefore resorted to deriving various broader related properties of the underlying stochastic processes, such as stability conditions, comparison results and performance bounds.

Stability of stochastic systems is a well-founded theory [28, 9]. Recently new results have been derived for systems with state-dependent (and time-varying) capacities. For example, in [22] the stability conditions for utility-based allocation policies in a time-varying scenario are characterized. In [6] necessary and sufficient stability conditions for parallel-server queues with state-dependent capacities are derived.

There is a wide range of literature on the ordering of random processes, see for example [35, 30]. In particular, stochastic comparison is often used. In the seminal paper [25] (see also [24]) necessary and sufficient conditions on the transition rates are given for the existence of a stochastic ordering between two Markov processes defined on ordered state spaces, starting from any two ordered initial states. It turns out that these conditions are often too strong in a queueing context. In particular, the conditions

*A shorter version with preliminary results appeared in the proceedings of ValueTools [38].

are not satisfied in the examples we study in this paper. Here we consider a special case of stochastic ordering: We use a sample-path approach to compare two stochastic networks, that is, for both networks we assume the same realizations of the arrival processes and service requirements (see [11, 23] for more details).

A related research direction is to obtain bounds for the stochastic process of interest [5, 41, 8]. In a recent paper [5] the authors consider a network of processor sharing queues with independent Poisson arrival processes. The capacity of the various queues is variable and depends on the number of users present in all the queues. Stochastic bounds for the number of users present in each queue are obtained for so-called monotone policies (removing a user from any queue increases the capacity allocated to every other user).

Our main interest is in stochastic processes that arise in so-called bandwidth-sharing networks introduced in [27] to model the dynamic interaction among competing elastic data flows that traverse several links in the Internet. An important family of rate allocation policies originally introduced in [29] are the so-called weighted α-fair bandwidth-sharing policies, where as a function of the parameter α one obtains popular disciplines such as maximum throughput ($\alpha \to 0$), Proportional Fairness (PF, $\alpha = 1$) and max-min fairness ($\alpha \to \infty$). It has been argued that the bandwidth sharing realized by TCP (Transmission Control Protocol) in the Internet can be well approximated by an α-fair policy with parameter $\alpha = 2$ [16]. In [4] it is shown that any α-fair policy ($\alpha > 0$) achieves maximum stability assuming Poisson arrival processes and exponentially distributed flow sizes. Obtaining closed-form expressions for the performance metrics of α-fair policies has proved to be rather difficult. Therefore, researchers have studied the performance under various probabilistic limiting regimes. For example, in [13, 14, 17] the authors study the number of users of the various classes under a fluid and a diffusion scaling when at least one node is in heavy traffic, and investigate diffusion approximations for the queue lengths.

In this paper we start off by considering a general multi-class queueing system setting with general arrival and service processes. The allocation to the various classes is feasible when it belongs to a rate region, which may vary in time. We give sufficient conditions on two allocation policies in order to compare sample-path wise the workload and the number of users of the various classes. We obtain weaker sufficient conditions on the transition rates than [25, 24]. Since our result is a pure sample-path comparison, it holds for arbitrary arrival processes, service time processes and rate region variations. Our sample-path comparison yields stability results and monotonicity of the mean holding cost. Then we apply our framework to linear networks. This is the canonical model to study the bandwidth sharing of data traffic that traverses multiple links and the cross-traffic it meets on its route. Linear networks can also model mutual interference in wireless networks or write permission in a shared database. For the family of weighted α-fair policies in the linear network, we obtain stability results and, under certain restrictions on the service requirements, show monotonicity of the mean holding cost with respect to the fairness parameter α and the relative weights. To cover all service requirement parameters, we consider a two-node linear network in a heavy-traffic regime and obtain further monotonicity results based on a conjecture in [13, 14]. For a normally-loaded system we perform numerical experiments that provide further insight into the performance of the α-fair policies. Finally, we consider a multi-class single-server queue for which we are especially interested in weighted time-sharing policies such as Discriminatory Processor Sharing (DPS) [20, 12, 1] and Generalized Processor Sharing (GPS) [10, 32]. For a single server with two classes we obtain that the mean holding cost is monotone for DPS and GPS with respect to the ratio of the weights. Then we extend the framework to cover the single-server queue with an arbitrary number of classes.

The remainder of the paper is organized as follows. In Section 2 the model is introduced and Section 3 describes the results for the general framework. We apply this framework to a linear network in Section 4 and we focus on weighted α-fair policies in Section 5. In Section 6 we consider the multi-class single-server queue.

2 Model description

We consider a multi-class queueing system with $L+1$ classes of users. Class-i users arrive according to a renewal process with mean inter-arrival time $1/\lambda_i$, and have service requirements B_i with mean $1/\mu_i$, $i = 0, \ldots, L$. Let $\rho_i = \frac{\lambda_i}{\mu_i}$ represent the offered work of class i per time unit. The inter-arrival times and service requirements are mutually independent random variables.

For a given scheduling policy π, denote by $N_i^\pi(t)$ the number of class-i users in the system at time t and let $\vec{N}^\pi(t) = (N_0^\pi(t), N_1^\pi(t), \ldots, N_L^\pi(t))$. Let $W_i^\pi(t)$ denote the total residual amount of work in class i

(i.e. the workload in class i) at time t. We assume the processes $N_i^\pi(t)$ and $W_i^\pi(t)$ to be right continuous with left limits. We further define N_i^π and W_i^π as random variables with the corresponding steady-state distributions (when they exist).
For a given policy π, denote by $s_i^\pi(t,\vec{n})$ the instantaneous service rate received by class i at time t when the system is in state $\vec{n}=(n_0,n_1,\dots,n_L)$. Hence the allocation given to class i can only depend on the time and on the number of users present in the system. We assume that $s_i^\pi(t,\vec{n})=0$ when $n_i=0$. In addition, the allocation vector $\vec{s}^\pi(t,\vec{n})=(s_0^\pi(t,\vec{n}),\dots,s_L^\pi(t,\vec{n}))$ has to lie in a certain rate region $R(t)\subset\mathbf{R}_+^{L+1}$ which may depend on the time t but not on the state $\vec{n}$ itself, that is $\vec{s}^\pi(t,\vec{n})\in R(t)$. In the remainder of the paper we suppress the dependence on t and write $\vec{s}^\pi(\vec{n})$ instead of $\vec{s}^\pi(t,\vec{n})$. The service discipline within a particular class, the intra-class policy, is the First Come First Served discipline (FCFS).
Denote by

$$S_i^\pi(t):=\int\limits_{u=0}^{t} s_i^\pi(\vec{N}^\pi(u))\mathrm{d}u$$

the cumulative amount of service received by class i during the time interval $[0,t]$. Let $A_i(0,t)$ be the amount of class-i work that arrived in the time interval $(0,t]$. Then the workload in class i at time t can be written as

$$W_i^\pi(t)=W_i^\pi(0)+A_i(0,t)-S_i^\pi(t). \tag{1}$$

Remark 2.1 *When the service requirements are exponentially distributed, for any non-anticipating intra-class policy, the stochastic behavior of the system (for example the distribution of the number of users of the various classes) is determined completely by the allocation vector $\vec{s}^\pi(\vec{n})$ and does not depend on the intra-class policy used. A policy is called non-anticipating when the discipline is not based on any knowledge of the actual realizations of the remaining service requirements. This implies that when the service requirements are exponentially distributed, the results we obtain (by assuming FCFS) are also valid for non-anticipating policies like the Processor Sharing discipline (PS), the Last Come First Served discipline and the Foreground Background discipline.*

Remark 2.2 *When the service requirements are exponentially distributed, the arrival processes are Poisson and the rate region $R(t)=R$ does not vary in time, the process $\{N_0^\pi(t),N_1^\pi(t),\dots,N_L^\pi(t)\}_{t\geq 0}$ is a continuous-time Markov process. The transition rates are given by*

$$(n_0,\dots,n_i,\dots,n_L)\rightarrow(n_0,\dots,n_i+1,\dots,n_L)\quad \text{at rate}\quad \lambda_i,$$

and

$$(n_0,\dots,n_i,\dots,n_L)\rightarrow(n_0,\dots,n_i-1,\dots,n_L)\quad \text{at rate}\quad \mu_i s_i^\pi(\vec{n}).$$

As indicated in Remark 2.1, the transition rates are independent of the non-anticipating intra-class policy used.

Our goal in this paper is to compare the performance of a multi-class queueing system under different policies. First of all, we will be interested in whether a policy can achieve stability. Another important performance measure we consider is the holding cost, $\sum_{i=0}^L c_i N_i^\pi(t)$, where c_i is an arbitrary nonnegative cost associated with class i, $i=0,\dots,L$. Because of Little's law, a policy that minimizes the total mean (weighted) number of users present in the system, minimizes the mean overall (weighted) sojourn time as well.
In Sections 4 and 5 we focus on a particular example of a multi-class queueing system: the linear network, see Figure 1. It might be convenient for the reader to bear this network in mind when reading Section 3. A linear network consists of L nodes. The capacity of node i at time t is equal to $C_i(t)$, $i=1,\dots,L$. Class-i users require service at node i only, $i=1,\dots,L$, while class-0 users require service at all nodes simultaneously. Hence the rate region corresponding to the linear network is equal to

$$R(t)=\{\vec{s}\in\mathbf{R}^{L+1}: s_0+s_i\leq C_i(t),\ \forall i=1,\dots,L\}.$$

When $C_i(t)=C$ for all i and all t, we refer to it as a symmetric linear network. The linear network can model situations such as bandwidth sharing in wired networks, mutual interference in wireless networks, and write permission in a global database. This will be discussed in more detail in Section 4.

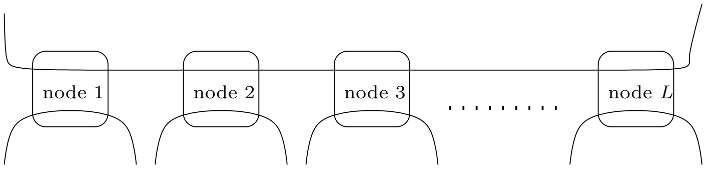

Figure 1: Linear network.

3 Comparison of policies

In this section we consider the behavior of a linear network under two different policies π and $\tilde{\pi}$ for the same realizations of the arrival processes and service requirements. The following property states conditions that will allow us to compare the two policies π and $\tilde{\pi}$.

Property 3.1 *Let π and $\tilde{\pi}$ be two policies such that*

(i) $s_0^{\pi}(\vec{n}^{\pi}) \leq s_0^{\tilde{\pi}}(\vec{n}^{\tilde{\pi}})$, *when* $n_0^{\pi} = n_0^{\tilde{\pi}}$ *and* $n_j^{\pi} \geq n_j^{\tilde{\pi}}, \forall j = 1, \ldots, L$.

(ii) $s_0^{\pi}(\vec{n}^{\pi}) + s_i^{\pi}(\vec{n}^{\pi}) \leq s_0^{\tilde{\pi}}(\vec{n}^{\tilde{\pi}}) + s_i^{\tilde{\pi}}(\vec{n}^{\tilde{\pi}})$, $i = 1, \ldots, L$, *for all states* $\vec{n}^{\pi}$ *and* $\vec{n}^{\tilde{\pi}}$ *that satisfy one of the following conditions:*

- $n_0^{\pi} > 0$, $n_0^{\pi} \geq n_0^{\tilde{\pi}}$, $0 < n_i^{\tilde{\pi}}$ *and* $n_i^{\pi} \leq n_i^{\tilde{\pi}}$.
- $n_0^{\pi} = n_0^{\tilde{\pi}} = 0$, $0 < n_i^{\pi} = n_i^{\tilde{\pi}}$ *and* $n_j^{\pi} \geq n_j^{\tilde{\pi}}$ *for all* $j \neq 0, i$.

In Section 4 we show how this property allows us to compare policies in a linear network.
We now establish a sample-path comparison result for the number of class-0 users and for the workload in the system. This result will play a key role in the remainder of the paper.

Proposition 3.2 *Let π and $\tilde{\pi}$ be two policies that satisfy Property 3.1 and consider the same realizations of the arrival processes and service requirements. Assume* $W_0^{\pi}(0) \geq W_0^{\tilde{\pi}}(0)$ *and* $W_0^{\pi}(0) + W_i^{\pi}(0) \geq W_0^{\tilde{\pi}}(0) + W_i^{\tilde{\pi}}(0)$ *for all* $i = 1, \ldots, L$. *It holds that for all* $t \geq 0$,

(i) $S_0^{\pi}(t) - W_0^{\pi}(0) \leq S_0^{\tilde{\pi}}(t) - W_0^{\tilde{\pi}}(0)$,

(ii) $S_0^{\pi}(t) - W_0^{\pi}(0) + S_i^{\pi}(t) - W_i^{\pi}(0) \leq S_0^{\tilde{\pi}}(t) - W_0^{\tilde{\pi}}(0) + S_i^{\tilde{\pi}}(t) - W_i^{\tilde{\pi}}(0), \quad i = 1, \ldots, L$,

and hence

(iii) $N_0^{\pi}(t) \geq N_0^{\tilde{\pi}}(t)$, $W_0^{\pi}(t) \geq W_0^{\tilde{\pi}}(t)$,

(iv) $W_0^{\pi}(t) + W_i^{\pi}(t) \geq W_0^{\tilde{\pi}}(t) + W_i^{\tilde{\pi}}(t), \quad i = 1, \ldots, L$.

We like to emphasize that because of the FCFS assumption and the same realizations of the arrival processes and service requirements, we implicitly assume that at time 0 the k-th most recently arrived class-i user has the same service requirement under both policies, $i = 0, 1, \ldots, L$, $k = 1, \ldots, \min(N_i^{\pi}(0), N_i^{\tilde{\pi}}(0)) - 1$. Hence, the condition in Proposition 3.2 always holds when both processes start in the same state $\vec{N}^{\pi}(0) = \vec{N}^{\tilde{\pi}}(0)$, where at time $t = 0$ each user has the same (remaining) service requirement under both policies.
In the proof of Proposition 3.2 we use $f(t^+) > g(t^+)$ to denote that there exists a sufficiently small $\delta > 0$ such that $f(u) > g(u)$ for all $u \in (t, t + \delta]$. Since $(N_i(t))_{t \geq 0}$ is a piece-wise constant right-continuous process, this ensures that an inequality on $N_i^{\pi}(t)$ and $N_i^{\tilde{\pi}}(t)$ at time t, immediately translates to the same inequality on $N_i^{\pi}(t^+)$ and $N_i^{\tilde{\pi}}(t^+)$ at time t^+. This property is used throughout the proof.

Proof of Proposition 3.2: From (1) we obtain that inequality (i) implies $W_0^{\pi}(t) \geq W_0^{\tilde{\pi}}(t)$ and inequality (ii) implies inequality (iv). Also note that $W_0^{\pi}(t) \geq W_0^{\tilde{\pi}}(t)$ implies $N_0^{\pi}(t) \geq N_0^{\tilde{\pi}}(t)$, since the intra-class policy is FCFS and the k-th most recently arrived class-0 user before the current time t has the same (original) service requirement under both policies. Therefore, it suffices to prove that inequalities (i) and (ii) hold.

We prove (i) and (ii) by contradiction. Suppose they do not hold sample-path wise. Let t be the first time epoch at which one of the two inequalities is violated.
First assume that inequality (i) is the first one to be violated, i.e., $S_0^{\pi}(t) - W_0^{\pi}(0) = S_0^{\tilde{\pi}}(t) - W_0^{\tilde{\pi}}(0)$ and $s_0^{\pi}(\vec{N}^{\pi}(t^+)) > s_0^{\tilde{\pi}}(\vec{N}^{\tilde{\pi}}(t^+))$ (with strict inequality), but $S_0^{\pi}(t) - W_0^{\pi}(0) + S_i^{\pi}(t) - W_i^{\pi}(0) \leq S_0^{\tilde{\pi}}(t) - W_0^{\tilde{\pi}}(0) + S_i^{\tilde{\pi}}(t) - W_i^{\tilde{\pi}}(0)$ for all $i = 1, \ldots, L$. Hence, from (1) we obtain $W_0^{\pi}(t) = W_0^{\tilde{\pi}}(t)$ and $W_i^{\pi}(t) \geq W_i^{\tilde{\pi}}(t)$ for all $i = 1, \ldots, L$. Since the k-th most recently arrived class-j user before the current time t has the same (original) service requirement under both policies and the intra-class policy is FCFS, we have as well

$$N_0^{\pi}(t) = N_0^{\tilde{\pi}}(t) \quad \text{and} \quad N_i^{\pi}(t) \geq N_i^{\tilde{\pi}}(t) \quad \text{for all} \quad i = 1, \ldots, L. \tag{2}$$

The process $\{N_i(t)\}_{t \geq 0}$ is a piece-wise constant process and is right continuous, hence (2) remains true at time t^+. Together with Property 3.1 this gives $s_0^{\pi}(\vec{N}^{\pi}(t^+)) \leq s_0^{\tilde{\pi}}(\vec{N}^{\tilde{\pi}}(t^+))$, which contradicts the initial assumption.
Next, assume that inequality (ii) is violated at time t, i.e., $S_0^{\pi}(t) - W_0^{\pi}(0) + S_i^{\pi}(t) - W_i^{\pi}(0) = S_0^{\tilde{\pi}}(t) - W_0^{\tilde{\pi}}(0) + S_i^{\tilde{\pi}}(t) - W_i^{\tilde{\pi}}(0)$ and $s_0^{\pi}(\vec{N}^{\pi}(t^+)) + s_i^{\pi}(\vec{N}^{\pi}(t^+)) > s_0^{\tilde{\pi}}(\vec{N}^{\tilde{\pi}}(t^+)) + s_i^{\tilde{\pi}}(\vec{N}^{\tilde{\pi}}(t^+))$ (with strict inequality), but $S_0^{\pi}(t) - W_0^{\pi}(0) \leq S_0^{\tilde{\pi}}(t) - W_0^{\tilde{\pi}}(0)$ and $S_0^{\pi}(t) - W_0^{\pi}(0) + S_j^{\pi}(t) - W_j^{\pi}(0) \leq S_0^{\tilde{\pi}}(t) - W_0^{\tilde{\pi}}(0) + S_j^{\tilde{\pi}}(t) - W_j^{\tilde{\pi}}(0)$ for all $j \neq 0, i$. Hence $W_0^{\pi}(t) \geq W_0^{\tilde{\pi}}(t)$ and $W_i^{\pi}(t) \leq W_i^{\tilde{\pi}}(t)$, from which (as before) we can conclude that $N_0^{\pi}(t^+) \geq N_0^{\tilde{\pi}}(t^+)$ and $N_i^{\pi}(t^+) \leq N_i^{\tilde{\pi}}(t^+)$. We now distinguish between the following possibilities:

- If $N_i^{\tilde{\pi}}(t^+) > 0$.
  - If $N_0^{\pi}(t^+) > 0$, then by Property 3.1 (ii) it follows that $s_0^{\pi}(\vec{N}^{\pi}(t^+)) + s_i^{\pi}(\vec{N}^{\pi}(t^+)) \leq s_0^{\tilde{\pi}}(\vec{N}^{\tilde{\pi}}(t^+)) + s_i^{\tilde{\pi}}(\vec{N}^{\tilde{\pi}}(t^+))$ which contradicts the initial assumption.
  - If $N_0^{\pi}(t^+) = 0$, then $N_0^{\tilde{\pi}}(t^+) = 0$ and hence $S_0^{\pi}(t) - W_0^{\pi}(0) = S_0^{\tilde{\pi}}(t) - W_0^{\tilde{\pi}}(0)$ which implies $S_i^{\pi}(t) - W_i^{\pi}(0) = S_i^{\tilde{\pi}}(t) - W_i^{\tilde{\pi}}(0)$ and $S_j^{\pi}(t) - W_j^{\pi}(0) \leq S_j^{\tilde{\pi}}(t) - W_j^{\tilde{\pi}}(0)$ for $j \neq 0, i$. So $0 = N_0^{\pi}(t^+) = N_0^{\tilde{\pi}}(t^+)$, $N_i^{\pi}(t^+) = N_i^{\tilde{\pi}}(t^+) > 0$, and $N_j^{\pi}(t^+) \geq N_j^{\tilde{\pi}}(t^+)$ for all $j \neq 0, i$. By Property 3.1 (ii) it follows that $s_0^{\pi}(\vec{N}^{\pi}(t^+)) + s_i^{\pi}(\vec{N}^{\pi}(t^+)) \leq s_0^{\tilde{\pi}}(\vec{N}^{\tilde{\pi}}(t^+)) + s_i^{\tilde{\pi}}(\vec{N}^{\tilde{\pi}}(t^+))$ which contradicts the initial assumption.
- If $N_i^{\tilde{\pi}}(t^+) = 0$, then $N_i^{\pi}(t^+) = 0$ as well, and hence $S_i^{\pi}(t) - W_i^{\pi}(0) = S_i^{\tilde{\pi}}(t) - W_i^{\tilde{\pi}}(0)$. This implies $S_0^{\pi}(t) - W_0^{\pi}(0) = S_0^{\tilde{\pi}}(t) - W_0^{\tilde{\pi}}(0)$ and $S_j^{\pi}(t) - W_j^{\pi}(0) \leq S_j^{\tilde{\pi}}(t) - W_j^{\tilde{\pi}}(0)$ for all j, implying $W_0^{\pi}(t) = W_0^{\tilde{\pi}}(t)$ and $W_j^{\pi}(t) \geq W_j^{\tilde{\pi}}(t)$. As before, we obtain that $N_0^{\pi}(t^+) = N_0^{\tilde{\pi}}(t^+)$ and $N_j^{\pi}(t^+) \geq N_j^{\tilde{\pi}}(t^+)$ for all $j \neq 0$. By virtue of Property 3.1 this means that $s_0^{\pi}(\vec{N}^{\pi}(t^+)) \leq s_0^{\tilde{\pi}}(\vec{N}^{\tilde{\pi}}(t^+))$. Since $N_i^{\tilde{\pi}}(t^+) = N_i^{\pi}(t^+) = 0$, we also have that $s_i^{\pi}(\vec{N}^{\pi}(t^+)) = s_i^{\tilde{\pi}}(\vec{N}^{\tilde{\pi}}(t^+)) = 0$, and hence $s_0^{\pi}(\vec{N}^{\pi}(t^+)) + s_i^{\pi}(\vec{N}^{\pi}(t^+)) \leq s_0^{\tilde{\pi}}(\vec{N}^{\tilde{\pi}}(t^+)) + s_i^{\tilde{\pi}}(\vec{N}^{\tilde{\pi}}(t^+))$, which contradicts the initial assumption.

□

Remark 3.3 *Proposition 3.2 is a sample-path result and does not require any distributional or independence assumptions with respect to the inter-arrival times and service requirements. The only assumption required is that the arrival characteristics are independent of the state of the system, since in Proposition 3.2 we use the same realizations of the arrival processes and service requirements when comparing the policies.*

Proposition 3.2 (iii) states in fact a sample-path wise pre-ordering on two continuous-time processes $\{\vec{N}^{\pi}(t)\}_{t \geq 0}$ and $\{\vec{N}^{\tilde{\pi}}(t)\}_{t \geq 0}$ starting from ordered initial states. There is a broad range of literature on the existence of orderings of stochastic processes. An important ordering is the stochastic ordering $\leq_{st}$ ([30, 35]). The sample-path ordering is a special case of this. Let $X(t)$ and $Y(t)$ be two continuous-time processes. We say that $\{X(t)\}_{t \geq 0} \leq_{st} \{Y(t)\}_{t \geq 0}$ if and only if there exists a coupling $(X'(t), Y'(t))$, i.e. $X(t) \overset{d}{=} X'(t)$ and $Y(t) \overset{d}{=} Y'(t)$, which is order-preserving, i.e. $\mathbb{P}(X'(t) \leq Y'(t), \forall t \geq 0) = 1$ (here $\leq$ is an ordering on the state space). So if the processes X and Y are initially ordered, then the order is kept at all times.
When $X(t)$ and $Y(t)$ are two continuous-time Markov processes, in [25, Theorem 5.3] and [24, Theorem 2] necessary and sufficient conditions on the transition rates are given in order for an order-preserving coupling to exist ($\{X(t)\}_{t \geq 0} \leq_{st} \{Y(t)\}_{t \geq 0}$) for any ordered initial states ($X(0) \leq Y(0)$). Here $\leq$ denotes a pre-order relation. In particular, in a Markovian setting (Poisson arrivals, exponentially distributed

service requirements and a fixed rate region, see Remark 2.2) the necessary and sufficient conditions on the policies π and $\tilde{\pi}$ to obtain

$$\{N_0^\pi(t)\}_{t\geq 0} \geq_{st} \{N_0^{\tilde{\pi}}(t)\}_{t\geq 0}, \quad \text{for any two ordered initial states} \quad N_0^\pi(0) \geq N_0^{\tilde{\pi}}(0), \tag{3}$$

are

$$s_0^\pi(\vec{n}^\pi) \leq s_0^{\tilde{\pi}}(\vec{n}^{\tilde{\pi}}) \;\; \text{when} \;\; n_0^\pi = n_0^{\tilde{\pi}}. \tag{4}$$

(The pre-ordering relation used here for the $L+1$-dimensional process $\vec{N}(t)$ is defined by the number of class-0 users.) The sufficient condition in Property 3.1 for the sample-path comparison of Proposition 3.2 to hold, and the necessary and sufficient condition in (4) for the stochastic comparison in (3) to hold, are not directly comparable. Given two policies, it is possible that either only Property 3.1 is satisfied, or only (4) is satisfied. Note that the stochastic ordering result in (3) holds for any two initial states that are ordered, $N_0^\pi(0) \geq N_0^{\tilde{\pi}}(0)$. In Proposition 3.2 the initial states are ordered as well, but we assume that at time $t = 0$ we have additional knowledge on the service requirements of the users present under policy π and $\tilde{\pi}$. So in this respect we would expect Property 3.1 to be weaker than (4). On the other hand, in Proposition 3.2 the coupling is specified in advance, namely the two processes are coupled by their arrival processes and service requirements, while in (3) any coupling is allowed to obtain the desired order-preserving result. So in this respect we would expect (4) to be weaker than Property 3.1.
In a queueing context, condition (4) is rather strong. One often encounters examples where $s_0(\vec{n}) \to 0$ as $n_i \to \infty$, $i \neq 0$. If this is the case for policy $\tilde{\pi}$, then (4) will not be satisfied. In Sections 5 and 6 we will consider settings for which Property 3.1 is satisfied, while (4) does not hold. In addition, Proposition 3.2 is not restricted to Markov processes, hence it applies as well for general arrival processes, service requirements and time-varying rate regions.
The results in [25] and [24] provide a notion of ordering that holds for any ordered initial states. In this paper we use a weaker notion, that is, we use additional information on the service requirements at time $t = 0$. This allows us to prove the auxiliary inequalities in Proposition 3.2 (i) and (ii) for policies π and $\tilde{\pi}$ that satisfy Property 3.1, which are crucial in proving the final ordering result. Since we are interested in performance metrics like stability and mean number of users, the chosen initial states are not relevant.
In the next two subsections, Proposition 3.2 is used to derive results for the stability and mean holding cost.

3.1 Stability

Recall that the stability conditions depend on the policy being used. The sample-path comparison in Proposition 3.2 does not require the system to be stable. In particular, Proposition 3.2 (iv) implies the following result.

Corollary 3.4 *Assume policies π and $\tilde{\pi}$ satisfy Property 3.1. If the system is stable under policy π, then it is stable under policy $\tilde{\pi}$ as well, in the sense that the system is empty under policy $\tilde{\pi}$ whenever it is empty under policy π.*
In particular, if the empty state is positive recurrent under policy π in the case of Poisson arrivals, then it is positive recurrent under policy $\tilde{\pi}$ as well.

Proof: The first statement follows by noting that if $\sum_{i=0}^L W_i^\pi(t) = 0$, then we obtain from Proposition 3.2 (iv) that $\sum_{i=0}^L W_i^{\tilde{\pi}}(t) = 0$. The second assertion is a direct implication of the first one. □

3.2 Mean number of users

In case the service requirements are exponentially distributed with $\sum_{i=1}^L c_i\mu_i \leq c_0\mu_0$, the sample-path comparison established in Proposition 3.2 allows us to compare the mean holding cost.

Proposition 3.5 *Assume the service requirements are exponentially distributed. Let π and $\tilde{\pi}$ be two policies that satisfy Property 3.1 and assume policy π gives a stable system. If $\sum_{i=1}^L c_i\mu_i \leq c_0\mu_0$, then*

$$\sum_{i=0}^{L} c_i \mathbb{E}(N_i^\pi(t)) \geq \sum_{i=0}^{L} c_i \mathbb{E}(N_i^{\tilde{\pi}}(t)), \quad \forall\; t \geq 0.$$

Proof: Assume at time $t = 0$ the conditions as stated in Proposition 3.2 are satisfied (for example, assume both policies π and $\tilde{\pi}$ start with an empty system). From Proposition 3.2 (iii) we have that $N_0^\pi(t) \geq N_0^{\tilde{\pi}}(t)$ for all $t \geq 0$. Taking expectations we get

$$\mathbb{E}(N_0^\pi(t)) \geq \mathbb{E}(N_0^{\tilde{\pi}}(t)). \tag{5}$$

From Proposition 3.2 (iv) we have that $W_0^\pi(t)+W_i^\pi(t) \geq W_0^{\tilde{\pi}}(t)+W_i^{\tilde{\pi}}(t)$ for all $t \geq 0$. Taking expectations we get $\mathbb{E}(W_0^\pi(t)) + \mathbb{E}(W_i^\pi(t)) \geq \mathbb{E}(W_0^{\tilde{\pi}}(t)) + \mathbb{E}(W_i^{\tilde{\pi}}(t))$ for all $i = 1, \ldots, L$. Since the policy is non-anticipating and the service requirements are exponentially distributed, and thus memoryless, we obtain $\mathbb{E}(W_i^\pi(t)) = \frac{1}{\mu_i}\mathbb{E}(N_i^\pi(t))$ and hence for all $i = 1, \ldots, L$,

$$\frac{1}{\mu_0}\mathbb{E}(N_0^\pi(t)) + \frac{1}{\mu_i}\mathbb{E}(N_i^\pi(t)) \geq \frac{1}{\mu_0}\mathbb{E}(N_0^{\tilde{\pi}}(t)) + \frac{1}{\mu_i}\mathbb{E}(N_i^{\tilde{\pi}}(t)). \tag{6}$$

Inequalities (5) and (6) together with $\sum_{i=1}^L c_i\mu_i \leq c_0\mu_0$ give

$$\begin{aligned}
\sum_{i=0}^L c_i\mathbb{E}(N_i^\pi(t)) &= \frac{c_0\mu_0 - \sum_{i=1}^L c_i\mu_i}{\mu_0}\mathbb{E}(N_0^\pi(t)) + \sum_{i=1}^L c_i\mu_i\left(\frac{1}{\mu_0}\mathbb{E}(N_0^\pi(t)) + \frac{1}{\mu_i}\mathbb{E}(N_i^\pi(t))\right)\\
&\geq \frac{c_0\mu_0 - \sum_{i=1}^L c_i\mu_i}{\mu_0}\mathbb{E}(N_0^{\tilde{\pi}}(t)) + \sum_{i=1}^L c_i\mu_i\left(\frac{1}{\mu_0}\mathbb{E}(N_0^{\tilde{\pi}}(t)) + \frac{1}{\mu_i}\mathbb{E}(N_i^{\tilde{\pi}}(t))\right)\\
&= \sum_{i=0}^L c_i\mathbb{E}(N_i^{\tilde{\pi}}(t)).
\end{aligned}$$

□

Note that by Remark 2.1, Proposition 3.5 holds for any non-anticipating intra-class policy, so not only for FCFS.

Remark 3.6 *We only obtain a comparison result in terms of the* mean *holding cost, while we start from a sample-path comparison as stated in Proposition 3.2. The derivation of stochastic ordering results remains as a challenging topic for further research.*
When $\vec{N}^\pi(t)$ and $\vec{N}^{\tilde{\pi}}(t)$ are two Markov processes, the necessary and sufficient conditions in order to obtain $\sum_{i=0}^L N_i^\pi(t) \geq_{st} \sum_{i=0}^L N_i^{\tilde{\pi}}(t)$ for any ordered initial states $\sum_{i=0}^L N_i^\pi(0) \geq \sum_{i=0}^L N_i^{\tilde{\pi}}(0)$, are $\sum_{i=0}^L \mu_i s_i^\pi(\vec{n}^\pi) \leq \sum_{i=0}^L \mu_i s_i^{\tilde{\pi}}(\vec{n}^{\tilde{\pi}})$ for all states with $\sum_{i=0}^L n_i^\pi = \sum_{i=0}^L n_i^{\tilde{\pi}}$, [24, 25]. In a queueing context this condition is rather strong. In Sections 4 and 6 we will see settings for which this condition is not satisfied.

4 Linear network

In this section we apply the results obtained in Section 3 to a linear network as depicted in Figure 1. As mentioned in the Introduction, the linear network provides a useful model for the interaction of data flows that traverse several links in a wired network, and experience bandwidth contention from independent cross traffic. A linear network also arises in simple models for the mutual interference in wireless networks. Consider the following setting of a wireless cellular network. Users can be either in cell 0, cell 1 or cell 2, see Figure 2. Users in cells 1 and 2 can be served in parallel by their own base station. Because of interference, a user in cell 0 can only be served when exactly one base station is on and transmits the requested file to the user in cell 0. Hence, class 0 can only be served when both classes 1 and 2 are not served, which can be modeled by a linear network consisting of two nodes. The results for the linear network that we obtain later in this section can be applied to a wireless network if coordination between base stations is possible. Coordination has recently been proposed in [3, 40].
As a further motivating example we could think of write permission in a shared database. Consider L servers that each perform tasks involving read/write operations in some shared database. Read operations can occur in parallel. However, if a server needs to perform a task involving write operations, then the database needs to be locked, and no tasks whatsoever can be performed by any of the other

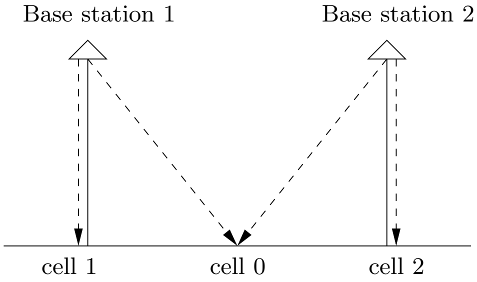

Figure 2: Two base stations.

servers. This may be modeled as a linear network with L nodes, where class-0 tasks corresponds to the write operations.

From now on we focus on *efficient* policies. A policy π is said to be efficient if it does not leave any capacity unnecessarily unused. So for the linear network this implies

$$s_i^{\pi}(\vec{n}) = C_i(t) - s_0^{\pi}(\vec{n}) \quad \text{when} \quad n_i > 0,\ i = 1\ldots, L \ \text{ for all } t.$$

Thus, the remaining capacity in node i is fully allocated to class-i users whenever possible. It can be shown that any policy that leaves capacity unused, can be improved sample-path wise (in terms of the workload and the number of users of the various classes) by an efficient policy. However, an efficient policy is *not* sufficient to ensure a stable system under the necessary stability conditions. Consider for example a symmetric linear network with unit capacities. It is clear that the necessary stability conditions are $\rho_0 + \rho_i < 1$ for all i. In fact, for the policy that gives preemptive priority to class 0 these conditions are sufficient for stability as well. However, the policy that gives preemptive priority to classes $1, \ldots, L$ (this is an efficient policy) is stable if and only if $\rho_0 < \Pi_{i=1}^{L}(1 - \rho_i)$ which is more stringent than the necessary stability conditions. The instability can arise here since the latter policy can leave a substantial portion of the capacity unused, regardless of how large the number of class-0 users is.

Condition (ii) in Property 3.1 is always satisfied for an efficient policy $\tilde{\pi}$, since $s_0^{\tilde{\pi}}(\vec{n}^{\tilde{\pi}}) + s_i^{\tilde{\pi}}(\vec{n}^{\tilde{\pi}}) = C_i(t)$ whenever $n_i^{\tilde{\pi}} > 0$. Hence, in the specific case of a linear network, Property 3.1 simplifies as follows.

Property 4.1 *Let π and $\tilde{\pi}$ be two efficient policies such that $s_0^{\pi}(\vec{n}^{\pi}) \leq s_0^{\tilde{\pi}}(\vec{n}^{\tilde{\pi}})$, when $n_0^{\pi} = n_0^{\tilde{\pi}}$ and $n_i^{\pi} \geq n_i^{\tilde{\pi}}$ for all $i = 1, \ldots, L$.*

In particular, Property 4.1 is implied by the following property.

Property 4.1' *Let π and $\tilde{\pi}$ be two efficient policies such that $s_0^{\pi}(\vec{n}) \leq s_0^{\tilde{\pi}}(\vec{n})$, and either $s_0^{\pi}(\vec{n})$ or $s_0^{\tilde{\pi}}(\vec{n})$ is non-increasing with respect to n_i for all $i \neq 0$.*

In order to see this, assume that Property 4.1' is satisfied with (for example) $s_0^{\tilde{\pi}}(\vec{n})$ non-increasing with respect to n_i for all $i \neq 0$. Then we have

$$s_0^{\pi}(\vec{n}^{\pi}) \leq s_0^{\tilde{\pi}}(\vec{n}^{\pi}) \leq s_0^{\tilde{\pi}}(\vec{n}^{\tilde{\pi}}),$$

with $n_i^{\pi} \geq n_i^{\tilde{\pi}}$ for all $i \neq 0$ and $n_0^{\pi} = n_0^{\tilde{\pi}}$. This is exactly Property 4.1. So for the linear network, Property 3.1 can be replaced by Property 4.1 or 4.1'.

Assume policies π and $\tilde{\pi}$ satisfy either Property 4.1 or 4.1'. This basically means that higher priority is given to class 0 under policy $\tilde{\pi}$ compared to π. From Section 3 we then obtain the following results. Under policy $\tilde{\pi}$ the number of class-0 users is less than under policy π (Proposition 3.2 (iii)) and the stability conditions are less strict for policy $\tilde{\pi}$ (Corollary 3.4). These results arise from the fact that when class 0 is served, it simultaneously uses capacity in all nodes. Hence, giving more preference to class 0 makes better use of the available capacity and hence makes the workload in each node smaller, i.e. $W_0^{\pi}(t)+W_i^{\pi}(t) \geq W_0^{\tilde{\pi}}(t)+W_i^{\tilde{\pi}}(t)$, $i = 1, \ldots, L$ (Proposition 3.2 (iv)). When in addition $c_0\mu_0 \geq \sum_{i=1}^{L} c_i\mu_i$, that is the maximum weighted departure rate is obtained when class 0 is served, giving higher priority to class 0 decreases the mean holding cost ($\sum_{i=0}^{L} c_i\mathbb{E}(N_i(t))$) as well (Proposition 3.5). More intuition on this will be given later. One natural choice for the weights c_i could be to relate them to the number of links each class uses. For example, take $c_0 = L$ and $c_i = 1$, $i = 1, \ldots, L$. In this case the result of

Proposition 3.5 will be valid under the intuitively appealing condition $\frac{1}{L}\sum_{i=1}^{L}\mu_i \leq \mu_0$, i.e. the departure rate of class 0 is larger than or equal to the average departure rate for classes $1,\ldots,L$.

Remark 4.2 *Assume $\vec{N}^{\pi}(t)$ and $\vec{N}^{\tilde{\pi}}(t)$ are two Markov processes for any two policies π and $\tilde{\pi}$. When Property 4.1 is satisfied, a sample-path comparison for the number of class-0 users in a linear network holds. The condition (4) is a necessary and sufficient condition for a stochastic ordering relation for the number of class-0 users to exist as in the framework of [24, 25]. It can be immediately seen that Property 4.1 is a weaker condition than (4). Interestingly, for applications as will be given later in the paper, the policies do satisfy Property 4.1, but not (4).*
When $c_0\mu_0 \geq \sum_{i=1}^{L} c_i\mu_i$ and Property 4.1 is satisfied, it is possible to compare the total (weighted) mean number of users in a linear network under the two policies. As mentioned in Remark 3.6, in a queueing context the sufficient and necessary conditions to stochastically order the total number of users for any ordered initial states, are rather strong. For the special case of a linear network it is even never satisfied. When choosing the states such that $\vec{n}^{\pi} = (0,1,\ldots,1)$ and $\vec{n}^{\tilde{\pi}} = (L,0,\ldots,0)$, it is needed that $\sum_{i=1}^{L}\mu_i \leq \mu_0$, but when choosing the states such that $\vec{n}^{\pi} = (1,0,\ldots,0)$ and $\vec{n}^{\tilde{\pi}} = (0,\ldots,0,1,0,\ldots,0)$, it is needed that $\mu_0 \leq \mu_i$, $i = 1,\ldots,L$, see Remark 3.6. Hence, we see that there does not exist any combination of the variables $\mu_0,\ldots,\mu_L$, for which these conditions are satisfied, and a stochastic ordering relation for the total number of users as in the framework of [24, 25] does not hold.

A natural objective in queueing networks is to minimize the total number of users in the system or the holding cost. Classical results for a single-server system indicate that giving preference to "small" users is beneficial in terms of the number of users present in the system [34, 36, 33, 31]. For exponentially distributed service requirements, the $c\mu$-rule, i.e. giving priority to the class with the highest weighted departure rate $c_i\mu_i$, minimizes the mean holding cost, $\sum_{i=1}^{K} c_i\mathbb{E}(N_i)$, among all non-anticipating policies. The problem of how to allocate the capacity of the nodes among the various users in a linear network is more complex. Besides trying to maximize the weighted departure rate, we must take into account that giving more preference to class 0 makes better use of the available capacity.
When $\sum_{i=1}^{L} c_i\mu_i > c_0\mu_0$, it can be the case that the maximum total instantaneous weighted departure rate is obtained when class 0 is not served. However, this does not necessarily make full use of the available resources. Some care has to be taken in allocating the available capacity. More information on the structure of the optimal policy for this case can be found in [39].
When $\sum_{i=1}^{L} c_i\mu_i \leq c_0\mu_0$, there is no conflict between these two objectives. The maximum total instantaneous weighted departure rate is obtained when class 0 is served at its maximum possible rate, i.e. $\min_i C_i(t)$, and the other classes obtain what is left. At the same time, this makes maximum use of the available capacity. Intuitively it is clear that the policy that gives preference to class 0 minimizes the mean holding cost. Using Proposition 3.5 it can be proved that this is indeed the case.

Corollary 4.3 *Consider a linear network with time-varying capacities. Assume the service requirements are exponentially distributed. Let policy π^* be the policy that serves class 0 at maximum rate, i.e., $s_0^{\pi^*}(\vec{n}) = \min_i C_i(t)$ if $n_0 > 0$ and $s_0^{\pi^*}(\vec{n}) = 0$ otherwise. Classes $1,\ldots,L$ obtain what is left, i.e., $s_i^{\pi^*}(\vec{n}) = C_i(t) - s_0^{\pi^*}(\vec{N})$ if $n_i > 0$ and $s_i^{\pi^*}(\vec{n}) = 0$ otherwise. If $\sum_{i=1}^{L} c_i\mu_i \leq c_0\mu_0$, then policy π^* minimizes the mean holding cost $\sum_{i=0}^{L} c_i\mathbb{E}(N_i(t))$, for all $t \geq 0$, among all non-anticipating policies.*

Proof: Note that $s_0^{\pi^*}(\vec{n})$ is constant with respect to n_i, $i \neq 0$. In addition, $s_0^{\pi^*}(\vec{n}) \geq s_0^{\pi}(\vec{n})$ for any policy π. Hence, Property 4.1' is satisfied and from Proposition 3.5 we obtain $\sum_{i=0}^{L} c_i\mathbb{E}(N_i^{\pi}(t)) \geq \sum_{i=0}^{L} c_i\mathbb{E}(N_i^{\pi^*}(t))$ for all $t \geq 0$ and any policy π. □

In [37] it was proved that for a symmetric linear network, policy π^*, as defined in Corollary 4.3, is in fact stochastically optimal in terms of the total number of users. That is, for every $t \geq 0$ and for any non-anticipating policy π we have $\sum_{i=0}^{L} N_i^{\pi}(t) \geq_{st} \sum_{i=0}^{L} N_i^{\pi^*}(t)$ given that $\vec{N}^{\pi}(0) = \vec{N}^{\pi^*}(0)$.
Proposition 3.2 and Property 4.1 are stated in order to compare two different *policies*. However, they also allow us to evaluate the impact of removing a node from the linear network on the performance of class 0, i.e., compare two different *networks* under the same policy. In the following corollary we show that the number of class-0 users is reduced when a node (and hence the corresponding cross traffic) is removed.

Corollary 4.4 *Let π be a policy in a linear network with L nodes that satisfies the following property:*

$$s_0^\pi(n_0, n_1, \dots, n_L) \leq s_0^\pi(n_0, m_1, \dots, m_{L-1}, 0)$$

for all $n_i \geq m_i, i = 1, \dots, L-1$.
Also consider the linear network where node L is removed (and hence has $L-1$ nodes) and apply the same policy π in the following way: $s_0^\pi(n_0, \dots, n_{L-1}) := s_0^\pi(n_0, \dots, n_{L-1}, 0)$.
If $W_0^{\pi,L}(0) \geq W_0^{\pi,L-1}(0)$ and $W_0^{\pi,L}(0) + W_i^{\pi,L}(0) \geq W_0^{\pi,L-1}(0) + W_i^{\pi,L-1}(0)$, then

$$N_0^{\pi,L}(t) \geq N_0^{\pi,L-1}(t)$$

and for $i = 1, \dots, L-1$

$$W_0^{\pi,L}(t) + W_i^{\pi,L}(t) \geq W_0^{\pi,L-1}(t) + W_i^{\pi,L-1}(t),$$

with $N_i^{\pi,l}(t)$ and $W_i^{\pi,l}(t)$ the number of class-i users and the class-i workload, respectively, at time t under policy π in a linear network with l nodes.

Proof: Policy π in a linear network with $L-1$ nodes can be seen as a policy in a linear network with L nodes by ignoring the class-L users. Denote this policy by $\tilde{\pi}$. So for all $x \geq 0$, $s_0^{\tilde{\pi}}(n_0, n_1, \dots, n_{L-1}, x) := s_0^\pi(n_0, n_1, \dots, n_{L-1})$. Hence

$$\begin{aligned} s_0^\pi(n_0, n_1, \dots, n_{L-1}, n_L) &\leq s_0^\pi(n_0, m_1, \dots, m_{L-1}, 0) \\ &= s_0^\pi(n_0, m_1, \dots, m_{L-1}) \\ &= s_0^{\tilde{\pi}}(n_0, m_1, \dots, m_{L-1}, x) \end{aligned}$$

for all x and all $n_i \geq m_i$, $i = 1, \dots, L-1$. This implies that policies π and $\tilde{\pi}$ satisfy Property 4.1 and from Proposition 3.2 the result follows. □

5 Weighted α-fair policies in a linear network

Weighted α-fair policies are an important family of policies that have received a lot of attention in recent years [4, 16, 17, 29]. For a given population $\vec{n}$, the weighted-α fair allocation is the solution to the following optimization problem:

$$\begin{cases} \max_{\vec{s} \in R(t)} \sum_{i=0}^L w_i n_i \left(\frac{s_i}{n_i}\right)^{1-\alpha} / (1-\alpha) & \text{if } \alpha > 0,\ \alpha \neq 1, \\ \max_{\vec{s} \in R(t)} \sum_{i=0}^L w_i n_i \log(\frac{s_i}{n_i}) & \text{if } \alpha = 1. \end{cases} \tag{7}$$

As mentioned in the Introduction, for different values of α, one obtains common bandwidth allocation principles, like maximum total throughput, proportional fairness, and max-min fairness. Denote the weighted α-fair discipline with weights $w = (w_0, w_1, \dots, w_L)$ and parameter α by $\pi(\alpha, w)$ and the corresponding allocation vector by $\vec{s}^{\,\pi(\alpha,w)}(\vec{N})$. The allocated capacity to class i is shared equally among all class-i users, hence the intra-class policy is PS. Recall that in the model description we assumed that the intra-class policy is FCFS. In all the results of this section we assume exponentially distributed service requirements. Thus, the results we obtain will also be valid if the intra-class policy is PS, see Remark 2.1. In order to compare two α-fair policies we only need to check whether Property 4.1' holds. In [4] it was shown that for a symmetric linear network with unit capacity for all nodes the weighted α-fair allocation is given by

$$s_0^{\pi(\alpha,w)}(\vec{n}) = \frac{(w_0 n_0^\alpha)^{1/\alpha}}{(w_0 n_0^\alpha)^{1/\alpha} + (\sum_{i=1}^L w_i n_i^\alpha)^{1/\alpha}} \tag{8}$$

and $s_i^{\pi(\alpha,w)}(\vec{n}) = 1 - s_0^{\pi(\alpha,w)}(\vec{n})$ for all i with $n_i > 0$. Using (8), it can be checked that Property 4.1' is satisfied for a symmetric linear network when comparing policies $\pi(\beta, w)$ and $\pi(\gamma, \tilde{w})$ with $\beta \leq \gamma$ and $\frac{w_0}{w_i} \leq \frac{\tilde{w}_0}{\tilde{w}_i}$, $i = 1, \dots, L$ (see also [21, Proposition 6.1]). For an asymmetric network we have no expression for the weighted α-fair allocation available. However, the optimization problem (7) allows us to prove that Property 4.1' is satisfied then as well. The proof may be found in Appendix A.

Lemma 5.1 *The following results hold in a linear network:*

(i) $s_0^{\pi(\alpha,w)}(\vec{n})$ *is non-increasing in* n_i, $i = 1, \dots, L$.

(ii) If $\beta \leq \gamma$, *then* $s_0^{\pi(\beta,w)}(\vec{n}) \leq s_0^{\pi(\gamma,w)}(\vec{n})$ *for all* $\vec{n}$.

(iii) If $\frac{w_0}{w_i} \leq \frac{\tilde{w}_0}{\tilde{w}_i}$, $i = 1, \dots, L$, *then* $s_0^{\pi(\alpha,w)}(\vec{n}) \leq s_0^{\pi(\alpha,\tilde{w})}(\vec{n})$ *for all* $\vec{n}$.

Since Property 4.1' holds for weighted α-fair policies, the comparison results in Proposition 3.2 apply. This allows us to gain insights into the performance of such policies in linear networks, see Subsections 5.1 and 5.2.
The stochastic comparison results in [24, Theorem 2] and [25, Theorem 5.3] are not applicable to the weighted α-fair policies. As we already mentioned in Remark 4.2, such an ordering is not possible for the total number of users present in the system. Also, an ordering for the number of class-0 users for any ordered intitial states is not possible, since equation (4) is not satisfied for the class of weighted α-fair policies in linear networks. Consider for example the simple symmetric linear network and choose states such that $n_0^\pi = n_0^{\tilde{\pi}}$, $n_1^\pi = 1$ and $n_1^{\tilde{\pi}} = m$ with π and $\tilde{\pi}$ two α-fair policies. From (8) we see that if m tends to ∞ then $s_0^{\pi(\alpha,w)}(\vec{n}^{\tilde{\pi}})$ tends to 0. Hence (4) cannot hold for any pair of α-fair policies.
In [5] the authors obtain stochastic bounds for the number of users present in any queue for policies that satisfy the monotonicity property (removing a user from any queue, increases the capacity allocated to every other user). This property fails to hold for a linear network under α-fair policies, as also indicated in [5]. For example, removing a class-1 user implies that class 1 gets less capacity and class 0 gets more. This however implies that classes $i = 2, \dots, L$ obtain less capacity as well and hence a class-i user gets less capacity, $i = 2, \dots, L$. The only requirement in Property 4.1' is that removing a class-i user, $i \neq 0$, increases the capacity allocated to the class-0 users. As shown in Lemma 5.1, this holds under natural conditions on the parameters of weighted α-fair policies.

Remark 5.2 *From Lemma 5.1 and Corollary 4.4 we obtain that under a weighted α-fair policy, the number of class-0 users in a linear network with L nodes is larger than in a linear network with $L-1$ nodes.*

In Section 5.1 the stability results are presented and in Section 5.2 monotonicity of the mean holding cost with respect to the fairness parameter and the relative weights is established. In order to broaden the comparison result, in Section 5.3 we investigate a heavy-traffic regime and in Section 5.4 we perform numerical experiments. In Section 5.5 we describe a time-scale separation (the dynamics of class-0 users are infinitely faster than those of classes $1, \dots, L$) and derive approximations for the mean number of users.

5.1 Stability

In [4] it is proved that for Poisson arrivals and exponentially distributed service requirements, any weighted α-fair allocation in a bandwidth-sharing network with *fixed* capacity gives a stable system, in the sense that the queue length process is positive-recurrent, under the necessary stability conditions that the load in each node is smaller than the available capacity. For example, in the case of a linear network the necessary stability conditions are $\rho_0 + \rho_i < C_i$, for all $i = 1, \dots, L$. Corollary 3.4 and Lemma 5.1 allow us to derive stability results for a linear network with *time-varying* capacities.

Corollary 5.3 *Consider a linear network with time-varying capacities. Let the service requirements be exponentially distributed. Assume $\beta \leq \gamma$ and $\frac{w_0}{w_i} \leq \frac{\tilde{w}_0}{\tilde{w}_i}$, $i = 1, \dots, L$. If policy $\pi(\beta, w)$ gives a stable system, then policy $\pi(\gamma, \tilde{w})$ gives a stable system as well.*

Proof: The α-fair policies have PS as intra-class policy. However, since we assume that the service requirements are exponentially distributed, the stochastic behavior of the network does not depend on which non-anticipating intra-class policy is being used. Therefore we can assume that we have a FCFS intra-class policy. From Lemma 5.1 we obtain that Property 4.1 is satisfied, hence the result in Corollary 3.4 applies. □

In [22] the authors consider the stability conditions for systems with a time-varying general rate region under an α-fair policy with unit weights. They assume that the rate region can be in a finite number of states according to a stationary and ergodic process. The authors characterize the stability conditions and show that the stability region is non-increasing in the value of α. Interestingly, Corollary 5.3 indicates that the stability region is in fact also non-decreasing in the value of α in the setting of a linear network. We obtain the following result.

Corollary 5.4 *Assume Poisson arrivals and exponentially distributed service requirements. Consider a linear network and assume the set of all the possible capacity vectors $(C_1(t), \ldots, C_L(t))$ can be in a finite number of states and evolves as a stationary and ergodic process. Let $\overline{C}_i$ be the average of the process $C_i(t)$.*
Policy $\pi(\alpha, w)$ with $w_i \leq w_0, i = 1, \ldots, L$, gives a stable system under the necessary stability conditions $\rho_0 + \rho_i < \overline{C}_i$, $i = 1, \ldots, L$.

Proof: In [22] it is shown that for α-fair policies with unit weights $(w_j = 1, j = 0, \ldots, L)$ the necessary stability conditions are given by $\rho_0 + \rho_i < \overline{C}_i$, $i = 1, \ldots, L$. Moreover, it is established that these conditions are sufficient as well for the policy $\pi(\alpha, \vec{1})$ when $\alpha \downarrow 0$. On the other hand, Corollary 5.3 states that the stability conditions become less strict when α increases. This proves that $\pi(\alpha, \vec{1})$ is stable under the necessary stability conditions, for all $\alpha > 0$. From Corollary 5.3 we can then conclude that the same holds for policy $\pi(\alpha, w)$ with $w_i \leq w_0, i = 1, \ldots, L$. □

5.2 Mean number of users

We are now ready to derive a monotonicity result for the mean number of users for weighted α-fair policies in a time-varying linear network. When $\sum_{i=1}^{L} c_i\mu_i \leq c_0\mu_0$, the instantaneous weighted departure rate of class 0 is relatively large, hence, it will be attractive to give preference to class-0 users, either by increasing the relative weight given to class 0, w_0/w_i, or by increasing the parameter α, see Lemma 5.1. At the same time this makes better use of the available capacity of the nodes, see Proposition 3.2 (iv). In the next corollary we prove that the mean holding cost indeed decreases when more preference is given to class 0. More precisely, the mean holding cost is non-increasing in α and in $\frac{w_0}{w_i}$, $i = 1, \ldots, L$.

Corollary 5.5 *Consider a linear network with time-varying capacities. Assume exponentially distributed service requirements with $\sum_{i=1}^{L} c_i\mu_i \leq c_0\mu_0$. If $\beta \leq \gamma$ and $\frac{w_0}{w_i} \leq \frac{\tilde{w}_0}{\tilde{w}_i}$, $i = 1, \ldots, L$, then*

$$\sum_{i=0}^{L} c_i \mathbb{E}(N_i^{\pi(\beta,w)}(t)) \geq \sum_{i=0}^{L} c_i \mathbb{E}(N_i^{\pi(\gamma,\tilde{w})}(t)), \quad \forall\ t \geq 0.$$

Proof: From Lemma 5.1 we obtain that $\pi(\beta, w)$ and $\pi(\gamma, \tilde{w})$ satisfy Property 4.1'. The result then follows from Proposition 3.5. □

When $\sum_{i=1}^{L} c_i\mu_i > c_0\mu_0$ the analysis is more difficult. For example, in a two-node linear network ($L = 2$) with $c_1\mu_1 + c_2\mu_2 > c_0\mu_0$, it is beneficial to give more preference to classes 1 and 2 (and hence less preference to class 0) since that will maximize the total instantaneous weighted departure rate. From Lemma 5.1 we see that this can be done by choosing α small. In the case of exponentially distributed service requirements and a *heavily loaded* system, the mean holding cost is indeed strictly increasing in α, as we will see in Section 5.3. For a *normally loaded* system this is however not the case (see the simulations in Section 5.4). Then the effect that a smaller α uses the available capacity in each node less efficiently becomes more apparent.

5.3 Heavy-traffic regime

In this section we compare α-fair policies in a heavy-traffic scenario for a two-node linear network with fixed capacities C_1 and C_2. Throughout this section we consider α-fair policies with unit weights $w_j = 1$, $j = 0, \ldots, L$. We consider the setting of [13, 14, 17], where a general bandwidth-sharing network under weighted α-fair allocations is considered with Poisson arrivals and exponentially distributed service requirements. Below we briefly state the results specialized to the two-node linear network under α-fair policies with unit weights. We refer to [13, 14] for the full details.
Assume the heavy-traffic setting $\rho_i + \rho_0 = C_i$, $i = 1, 2$. Define the diffusion scaled processes as follows:

$$\hat{N}_i^{k,\pi(\alpha)}(t) := \frac{N_i^{\pi(\alpha,\vec{1})}(kt)}{\sqrt{k}}, \ i = 0, 1, 2$$

and

$$\hat{V}_i^{k,\pi(\alpha)}(t) := \frac{N_0^{\pi(\alpha,\vec{1})}(kt)/\mu_0 + N_i^{\pi(\alpha,\vec{1})}(kt)/\mu_i}{\sqrt{k}} = \hat{N}_0^{k,\pi(\alpha)}(t)/\mu_0 + \hat{N}_i^{k,\pi(\alpha)}(t)/\mu_i,\ i = 1, 2.$$

Here $\hat{V}_i^{k,\pi(\alpha)}(t)$ can be seen as the total workload in node i under the diffusion scaling. In [14, Conjecture 5.1] it is conjectured that for an arbitrary bandwidth-sharing network, the diffusion scaled workload process $\hat{\vec{V}}^{k,\pi(\alpha)}(t)$ converges in distribution as $k \to \infty$ to $\hat{\vec{V}}^{\pi(\alpha)}(t)$, where $\hat{\vec{V}}^{\pi(\alpha)}(t)$ is a semimartingale reflecting Brownian motion (with a covariance matrix independent of α) living in a workload cone. For α equal to 1 this conjecture is proved in [13, 14] for an arbitrary bandwidth-sharing network. In addition, it is mentioned that for the case of a two-node linear network, this result can be extended to $\alpha \neq 1$. Throughout this section we will assume that the conjecture holds for the two-node linear network for general α.

The workload cone for a two-node linear network under an α-fair policy with unit weights is given by

$$\{\vec{v} : v_i = \frac{\rho_0}{\mu_0}(q_1 + q_2)^{\frac{1}{\alpha}} + \frac{\rho_i}{\mu_i} q_i^{\frac{1}{\alpha}},\ q_1, q_2 \geq 0,\ i = 1, 2\} \tag{9}$$

$$= \{\vec{v} : v_1 \geq 0,\ \ v_1 \frac{\rho_0/\mu_0}{(C_1 - \rho_0)/\mu_1 + \rho_0/\mu_0} \leq v_2 \leq v_1 \frac{(C_2 - \rho_0)/\mu_2 + \rho_0/\mu_0}{\rho_0/\mu_0}\}, \tag{10}$$

which is independent of the parameter α. Hence, the workload process $\hat{\vec{V}}^{\pi(\alpha)}(t)$ is independent of α as well. The diffusion scaled number of users, $\hat{\vec{N}}^{k,\pi(\alpha)}(t)$, converges in distribution as $k \to \infty$ to some process $\hat{\vec{N}}^{\pi(\alpha)}(t)$, which does depend on α (this process is specified in Appendix B).

Since the process of the total workload in a node does not depend on α, we are able to derive monotonicity results for the mean holding cost over the whole range of the parameter μ_0. We can express the scaled holding cost as follows:

$$\begin{aligned}\sum_{i=0}^{2} c_i \hat{N}_i^{\pi(\alpha)}(t) &= \frac{c_0\mu_0 - c_1\mu_1 - c_2\mu_2}{\mu_0} \cdot \hat{N}_0^{\pi(\alpha)}(t) + \sum_{i=1}^{2} c_i\mu_i \cdot (\frac{1}{\mu_0}\hat{N}_0^{\pi(\alpha)}(t) + \frac{1}{\mu_i}\hat{N}_i^{\pi(\alpha)}(t)) \\ &\overset{d}{=} \frac{c_0\mu_0 - c_1\mu_1 - c_2\mu_2}{\mu_0} \cdot \hat{N}_0^{\pi(\alpha)}(t) + \sum_{i=1}^{2} c_i\mu_i \hat{V}_i^{\pi(\alpha)}(t). \end{aligned} \tag{11}$$

From Proposition 3.2 we know that $N_0^{\pi(\alpha,\vec{1})}(t)$ is decreasing in α, and hence $\hat{N}_0^{\pi(\alpha)}(t)$ is decreasing in α as well. Since $\hat{V}_i^{\pi(\alpha)}(t)$ is independent of α, and by taking expectations in (11), we obtain that if $c_1\mu_1 + c_2\mu_2 \leq c_0\mu_0$ or $c_1\mu_1 + c_2\mu_2 \geq c_0\mu_0$, then $\mathbb{E}(\sum_{i=0}^{2} c_i \hat{N}_i^{\pi(\alpha)}(t))$ is non-increasing or non-decreasing in α, respectively.

When in addition we use the characterization of $\hat{\vec{N}}^{\pi(\alpha)}(t)$, we are able to derive a stronger monotonicity result. The proof may be found in Appendix B.

Proposition 5.6 *Consider a linear network with fixed capacities C_1 and C_2. Assume that the inter-arrival times and service requirements are exponentially distributed, and $\rho_i + \rho_0 = C_i$ for $i = 1, 2$. If the conjecture in [14] is valid, then*

- *If $c_1\mu_1 + c_2\mu_2 < c_0\mu_0$, then $\mathbb{E}(\sum_{i=0}^{2} c_i \hat{N}_i^{\pi(\alpha)}(t))$ is strictly decreasing in α.*
- *If $c_1\mu_1 + c_2\mu_2 = c_0\mu_0$, then $\mathbb{E}(\sum_{i=0}^{2} c_i \hat{N}_i^{\pi(\alpha)}(t))$ is constant in α.*
- *If $c_1\mu_1 + c_2\mu_2 > c_0\mu_0$, then $\mathbb{E}(\sum_{i=0}^{2} c_i \hat{N}_i^{\pi(\alpha)}(t))$ is strictly increasing in α.*

5.4 Numerical results

In this section we present numerical experiments to provide further insight into the performance of α-fair policies. We consider a two-node linear network where both nodes have unit capacity. We assume Poisson arrivals and exponentially distributed service requirements and fix $\mu_1 = 1, \mu_2 = 0.5$, $\rho_1 = \rho_2$ and

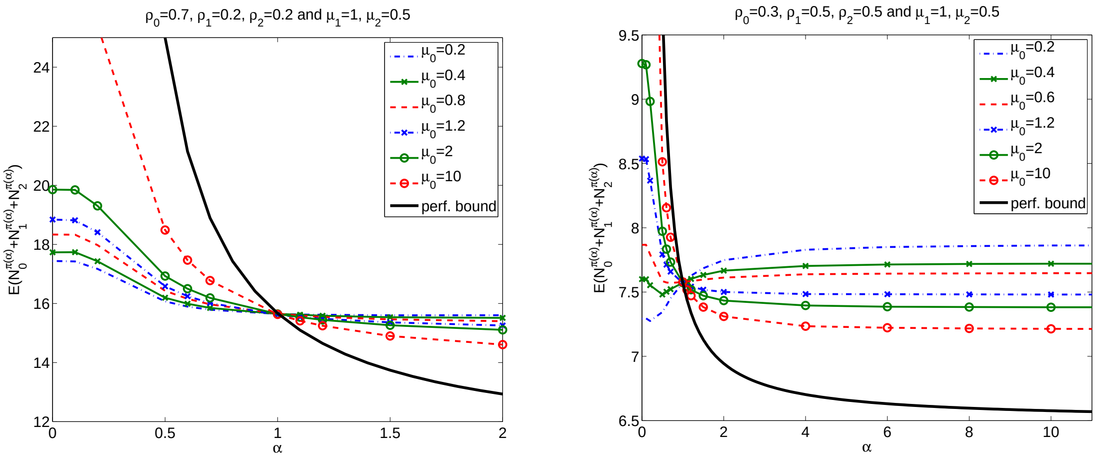

Figure 3: Total mean number of users under α-fair policies in a two-node linear network with a) $\rho_0 = 0.7, \rho_1 = 0.2$ and $\rho_2 = 0.2$, and b) $\rho_0 = 0.3, \rho_1 = 0.5$ and $\rho_2 = 0.5$.

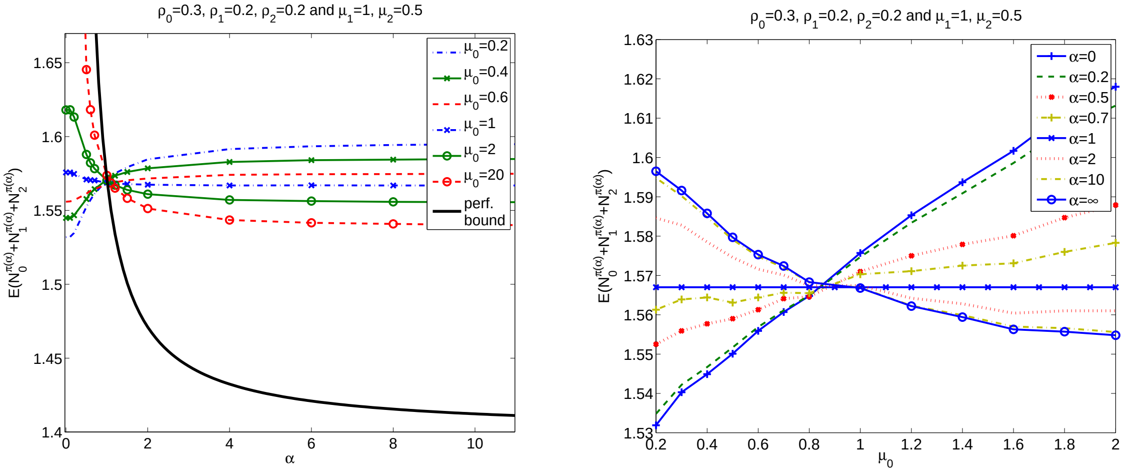

Figure 4: Total mean number of users under α-fair policies in a two-node linear network with $\rho_0 = 0.3, \rho_1 = 0.2$ and $\rho_2 = 0.2$.

$w_j = c_j = 1,\ j = 0, 1, 2$. The numerical experiments are performed using Matlab©, and in the order of 10^7 busy periods are simulated.

In Figures 3 a) and b) and Figure 4 a) we let α vary on the horizontal axis and plot the corresponding total mean number of users for various values of μ_0. As expected from Corollary 5.5, we observe that for $\mu_0 \geq \mu_1 + \mu_2 = 1.5$ the total mean number of users is decreasing with respect to the value of α. When $\mu_0 < \mu_1 + \mu_2 = 1.5$, we observe that the total mean number of users is monotone (either decreasing or increasing) in α as well in the range $\alpha \in [1, \infty)$. However, when $\alpha \in (0, 1)$ and $\mu_0 < \mu_1 + \mu_2 = 1.5$, it is possible that the total mean number of users is not monotone in α. This fact may be explained as follows. Since $\mu_0 < \mu_1 + \mu_2 = 1.5$, it is attractive to give more preference to classes 1 and 2 when they are both present (hence less preference to class 0). This corresponds to a small value for α. However, an α-fair policy with a small α uses the available capacity less efficiently, see Proposition 3.2 (iv) and Lemma 5.1 (ii). These two opposite effects might cause the total mean number of users to not be monotone in α. Note that for the heavy-traffic regime as considered in Section 5.3, the workload in a node was independent of the parameter α and hence every value for α had the same efficiency. Therefore, there was no trade-off and we were able to prove the monotonicity results for $\mu_0 < \mu_1 + \mu_2$ as well.

In Figure 4 b) we let μ_0 vary on the horizontal axis and plot the corresponding total mean number of users for various values of α. We observe that the total mean number of users is mostly increasing in μ_0 when $\alpha < 1$ and decreasing in μ_0 when $\alpha > 1$, respectively. This can be explained as follows. First of all, if $\alpha = 1$, the policy reduces to PF. For PF with unit weights, the mean total number of users is exactly known and equals

$$\mathbb{E}\left(\sum_{i=0}^{L} N_i^{\pi(1,\vec{1})}\right) = \frac{\rho_1}{1-\rho_0-\rho_1} + \frac{\rho_2}{1-\rho_0-\rho_2} + \frac{\rho_0}{1-\rho_0}\Big(1 + \frac{\rho_1}{1-\rho_0-\rho_1} + \frac{\rho_2}{1-\rho_0-\rho_2}\Big), \tag{12}$$

see [26]. In fact, PF is insensitive to the service requirement distributions apart from their respective means (see [26]) and hence (12) holds for generally distributed service requirements. In particular, the total mean number of users is independent of the parameters μ_0, μ_1 and μ_2 for given values of ρ_0, ρ_1 and ρ_2. When $\alpha > 1$, from Lemma 5.1 (ii) we observe that class 0 is treated preferentially over classes 1 and 2 (compared to PF). Under an α-fair policy that gives preference to class 0, it is likely that the total mean number of users decreases when the class-0 users become smaller, i.e., when μ_0 increases, while $\mu_1, \mu_2, \rho_0, \rho_1$ and ρ_2 are kept fixed. Similarly, when $\alpha < 1$, classes 1 and 2 are treated preferentially over class 0 (compared to PF). When μ_0 becomes larger (while $\mu_1, \mu_2, \rho_0, \rho_1$ and ρ_2 are kept fixed), class-1 and 2 users become relatively larger. Under an α-fair policy that gives preference to classes 1 and 2, it is likely that the total mean number of users increases when μ_0 increases.

5.5 Time-scale separation

In [5] the authors introduce the so-called quasi-stationary and fluid-limit regimes (see also [18]). In these regimes, the flow dynamics of the various classes occur on separate time scales, which can greatly simplify the analysis. It was conjectured in [5] that these limiting regimes provide performance bounds. For the symmetric linear network with unit weights, Poisson arrivals and generally distributed service requirements, we refer to the quasi-stationary and fluid regimes when $\mu_0 \to \infty$ and $\mu_0 \to 0$, respectively, and keeping $\mu_1, \ldots, \mu_L$ and $\rho_0, \rho_1, \ldots, \rho_L$ fixed. From our simulation results for a linear network it seems that these limiting regimes can indeed be performance bounds, see Figure 4 b). When $\alpha > 1$, the quasi-stationary regime ($\mu_0 \to \infty$) is a lower bound on the total mean number of users and the fluid regime ($\mu_0 \to 0$) an upper bound on the total mean number of users, and when $\alpha < 1$ vice versa. A similar observation was made in [18] for a DPS queue.

We develop here an approximate analysis of the quasi-stationary regime. The approximate formulae might be useful in assessing the performance of α-fair policies, since exact closed-form formulae are not available. In the quasi-stationary regime, $\mu_0 \to \infty$, the dynamics of class 0 will "average out" on the relevant time scale for class i, $i = 1, \ldots, L$. Hence, we can say that class 0 takes away a constant service rate ρ_0 and class i sees capacity $1 - \rho_0$. Class i behaves as in a PS system with capacity $1 - \rho_0$, which implies that the number of class-i users in the system is geometrically distributed with mean $\frac{\rho_i}{1-\rho_0-\rho_i}$ [15]. Hence, $\lim_{\mu_0\to\infty} \mathbb{E}(N_i^{\pi(\alpha,w)}) = \frac{\rho_i}{1-\rho_0-\rho_i}$, which is independent of α and $\frac{w_0}{w_i}$.

The time scale of class 0 is infinitely faster than that of classes $1, \ldots, L$. Thus on the time scale of class 0, the dynamics of classes $1, \ldots, L$ almost vanish. It can be assumed that for a given number of class-i users, $i = 1, \ldots, L$, class 0 will reach some sort of statistical equilibrium. We recall from (8) that $s_0^{\pi(\alpha,w)}(\vec{n}) = \frac{n_0}{n_0+c}$, with $c = c(n_1, \ldots, n_L) = (\sum_{i=1}^{L} \frac{w_i}{w_0} n_i^\alpha)^{1/\alpha}$. Thus, given a population $\vec{n}$, class 0 behaves like a PS system with c permanent users. The mean number of users in such a system is $\frac{\rho_0}{1-\rho_0}(1+c)$. Unconditioning and noting that $N_i^{\pi(\alpha,w)}$ is in the limit geometrically distributed with mean $\frac{\rho_i}{1-\rho_0-\rho_i}$, $i = 1, \ldots, L$, we get that approximately

$$\begin{aligned}
\lim_{\mu_0\to\infty} \mathbb{E}(N_0^{\pi(\alpha,w)}) &= \lim_{\mu_0\to\infty} \sum_{n_1,\ldots,n_L} \mathbb{E}(N_0^{\pi(\alpha,w)} | N_i^{\pi(\alpha,w)} = n_i, i = 1, \ldots, L) \cdot \mathbb{P}(N_i^{\pi(\alpha,w)} = n_i, i = 1, \ldots, L) \\
&= \lim_{\mu_0\to\infty} \sum_{n_1,\ldots,n_L} \frac{\rho_0}{1-\rho_0} \cdot \left(1 + (\sum_{i=1}^{L} \frac{w_i}{w_0} n_i^\alpha)^{1/\alpha}\right) \cdot \mathbb{P}(N_i^{\pi(\alpha,w)} = n_i, i = 1, \ldots, L) \\
&\approx \frac{\rho_0}{1-\rho_0} \cdot \left(1 + \left(\sum_{i=1}^{L} \frac{w_i}{w_0} (\frac{\rho_i}{1-\rho_0-\rho_i})^\alpha\right)^{1/\alpha}\right). \qquad (13)
\end{aligned}$$

We ignored here the non-linearity induced by the parameter α. We see that the performance of class 0 does depend on α and the weights w_i, and using similar arguments as in the proof of Lemma 5.1, it

can be checked that the mean number of class-0 users as given in (13) indeed decreases when α or $\frac{w_0}{w_i}$ increases (as was proved already in Proposition 3.2).
As an approximation for the total mean number of users we then obtain

$$\lim_{\mu_0\to\infty} \mathbb{E}\left(\sum_{i=0}^{L} N_i^{\pi(\alpha,w)}\right) \approx \frac{\rho_0}{1-\rho_0}\cdot\left(1+\left(\sum_{i=1}^{L}\frac{w_i}{w_0}(\frac{\rho_i}{1-\rho_0-\rho_1})^{\alpha}\right)^{1/\alpha}\right)+\sum_{i=1}^{L}\frac{\rho_i}{1-\rho_0-\rho_i}. \tag{14}$$

The approximation (14) gives the correct expression for $\alpha = 1$ and unit weights, see (12). In Figures 3 and Figure 4 a) we plotted (14) against α (denoted in the figures by "perf. bound"). We observe that (14) provides indeed an upper bound on the performance when $\alpha < 1$, and a lower bound when $\alpha > 1$. Even for moderate values of μ_0, the bound is quite tight and not off by more than 10% as long as the value of α is not too small or too large.
Unfortunately, it does not seem possible to derive an approximation for the fluid regime. When $\mu_0 \to 0$, the dynamics of classes $1,\dots,L$ "average out" on the relevant time scale of class 0. Thus, class 0 sees a system with capacity $1-\max(\rho_1,\dots,\rho_L)$. The time scale of classes $1,\dots,L$ are infinitely faster than that of class 0, hence on the relevant time scale of classes $1,\dots,L$, the dynamics of class 0 nearly vanish. Thus, given a certain number of class-0 users, class i obtains capacity $s_i^{\pi(\alpha,w)}(\vec{n}) = (\sum_{i=1}^{L}\frac{w_i}{w_0}n_i^{\alpha})^{1/\alpha}/(n_0 + (\sum_{i=1}^{L}\frac{w_i}{w_0}n_i^{\alpha})^{1/\alpha})$, where n_0 can be considered fixed. From this equation we cannot approximate the behavior of classes $1,\dots,L$ by any known queueing system unless $\alpha = 1$.

6 Multi-class single-server queue

In Section 4 and Section 5 we have focused on a linear network. In this section we turn our attention to the multi-class single-server queue with time-varying capacity $C(t)$. There are K classes of users, where class-i users arrive according to a general arrival process with rate λ_i, and have generally distributed service requirements with mean $1/\mu_i$, $i = 1,\dots,K$. Let $\rho = \sum_{i=1}^{K}\rho_i$. The inter-arrival times and the service requirements are mutually independent random variables. We consider allocation policies that are work-conserving, i.e. if $\sum_{i=1}^{K} n_i > 0$ then $\sum_{i=1}^{K} s_i(\vec{n}) = C(t)$, and if $n_i = 0$ then $s_i(\vec{n}) = 0$. The intra-class policy is FCFS.
In Section 6.1 we consider two popular weighted time-sharing policies and, using the general results from Section 3, we obtain monotonicity properties in the case of two classes of users. In Section 6.2 we derive a framework (similar to the one derived in Section 3) for a multi-class single-server system (with an arbitrary number of classes) under work-conserving disciplines.

6.1 GPS and DPS policies

The policies we are particularly interested in are GPS [10, 32] and DPS [20, 12, 1], two popular non-anticipating policies in multi-class single-server queues. Let $GPS(\phi)$ ($DPS(\phi)$) denote a GPS (DPS) discipline that assigns weight ϕ_j to class j, with $\sum_{j=1}^{K}\phi_j = 1$.
The GPS allocation is given by

$$s_i^{GPS(\phi)}(\vec{n}) = C(t)\frac{\phi_i}{\sum_{j=1}^{K}\phi_j \mathbf{1}_{(n_j>0)}}, \quad i = 1,\dots,K,$$

for $\sum_{j=1}^{K} n_j > 0$. We take as intra-class policy in GPS the FCFS discipline.
The DPS allocation is given by

$$s_i^{DPS(\phi)}(\vec{n}) = C(t)\frac{\phi_i n_i}{\sum_{j=1}^{K}\phi_j n_j}, \quad i = 1,\dots,K, \tag{15}$$

for $\sum_{j=1}^{K} n_j > 0$. With DPS, the allocated capacity to class i is shared equally among all class-i users, hence the intra-class policy in DPS is PS.
Assume the service requirements are exponentially distributed with $c_1\mu_1 \geq c_2\mu_2 \geq \cdots \geq c_K\mu_K$. The $c\mu$-rule (give priority to the class with highest $c_i\mu_i$) minimizes the mean holding cost among all non-anticipating policies (see for example [31]). For both GPS and DPS, a class is given more preference

when its weight is increased. Hence it seems plausible that giving relatively more weight to classes with a high $c_i\mu_i$, will decrease the mean holding cost. For a single-server system with only two classes of users ($K = 2$) we can indeed prove this. Such a system is equivalent to a linear network with one node ($L = 1$). When $\phi_1 < \tilde{\phi}_1$, the policies $GPS(\phi)$ and $GPS(\tilde{\phi})$ ($DPS(\phi)$ and $DPS(\tilde{\phi})$) satisfy Property 4.1'. Hence, we can use the results of Section 3 to obtain monotonicity results for GPS and DPS in a single-server system with two classes of users and time-varying capacity.

Proposition 6.1 *Consider a single-server queue with two classes of users and time-varying capacity. Let $\phi_1 < \tilde{\phi}_1$. Assume $W_1^{\pi}(0) \geq W_1^{\tilde{\pi}}(0)$, $W_2^{\pi}(0) \leq W_2^{\tilde{\pi}}(0)$ and $W_1^{\pi}(0) + W_2^{\pi}(0) = W_1^{\tilde{\pi}}(0) + W_2^{\tilde{\pi}}(0)$, where either $\pi = GPS(\phi)$ and $\tilde{\pi} = GPS(\tilde{\phi})$, or $\pi = DPS(\phi)$ and $\tilde{\pi} = DPS(\tilde{\phi})$. We consider the same realizations of the arrival processes and service requirements for both processes.*
For generally distributed service requirements it holds that

$$W_1^{GPS(\phi)}(t) \geq W_1^{GPS(\tilde{\phi})}(t) \quad \text{and} \quad N_1^{GPS(\phi)}(t) \geq N_1^{GPS(\tilde{\phi})}(t). \tag{16}$$

The opposite inequalities hold for class 2.
For exponentially distributed service requirements it holds that

$$\{W_1^{DPS(\phi)}(t)\}_t \geq_{st} \{W_1^{DPS(\tilde{\phi})}(t)\}_t \quad \text{and} \quad \{N_1^{DPS(\phi)}(t)\}_t \geq_{st} \{N_1^{DPS(\tilde{\phi})}(t)\}_t. \tag{17}$$

The opposite inequalities hold for class 2.
If the service requirements are exponentially distributed with $c_1\mu_1 \geq c_2\mu_2$ and the system can be made stable, then

$$\sum_{i=1}^{2} c_i \mathbb{E}(N_i^{GPS(\phi)}(t)) \geq \sum_{i=1}^{2} c_i \mathbb{E}(N_i^{GPS(\tilde{\phi})}(t)), \quad \forall\ t \geq 0, \tag{18}$$

and

$$\sum_{i=1}^{2} c_i \mathbb{E}(N_i^{DPS(\phi)}(t)) \geq \sum_{i=1}^{2} c_i \mathbb{E}(N_i^{DPS(\tilde{\phi})}(t)), \quad \forall\ t \geq 0. \tag{19}$$

Proof: Since the respective pair of policies satisfy Property 4.1', equations (16) and (17) follow directly from Propositions 3.2, and equations (18) and (19) follow directly from Proposition 3.5. For exponentially distributed service requirements, the stochastic behavior is independent of the used intra-class policy, see Remark 2.1. For DPS we consider exponentially distributed service requirements. Hence, the sample-path comparison in Proposition 3.2 obtained for FCFS, allows us to obtain the stochastic comparison result in (17) for DPS when the intra-class policy is PS. □

Inequalities (16) and (17) are rather natural, but to the best of our knowledge have not been obtained previously. In particular, the comparison results from [24] and [25] do not allow for such a comparison, as explained later in Remark 6.10. The result for GPS is particularly interesting. The GPS discipline is used to model the queueing delay experienced by packets in packet networks. An important body of research on GPS is devoted to the characterization of the workload when there are two classes of users, see for example [32, 7].

Inequalities (18) and (19) show that for two classes, the mean holding cost under DPS or GPS is monotone in the whole range $\phi_1 \in [0, \infty)$, where one extreme corresponds to giving preemptive priority to class 2 ($\phi_1 = 0$) and the other extreme to preemptive priority to class 1 ($\phi_1 = 1$). To the best of our knowledge, this kind of monotonicity result is new for GPS. In the case of a two-class single-server DPS-system *with* fixed capacity and Poisson arrivals, this result could also be obtained from the analysis in [12] (see [2] for more details).

For an arbitrary number of classes little is known on monotonicity results for GPS and DPS. As mentioned before, motivated by the optimality of the $c\mu$-rule, one would expect that giving relatively more weight to classes with a high $c_i\mu_i$, will decrease the mean holding cost. One of the most relevant results is obtained in [19]. The authors consider a single server with fixed capacity, Poisson arrivals and exponentially distributed service requirements with $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_K$. Using the results of [12] they prove that if $\tilde{\phi}_1 \geq \tilde{\phi}_2 \geq \cdots \geq \tilde{\phi}_K$, then $\mathbb{E}(\sum_{i=1}^K N_i^{PS}) \geq \mathbb{E}(\sum_{i=1}^K N_i^{DPS(\tilde{\phi})})$. Note that PS is equivalent to a DPS policy with weights $\phi_i = \phi_j$, $\forall i, j$, see equation (15). In general, we expect the following results to hold (a similar conjecture for the steady-state of DPS has been made in [19]).

Conjecture 6.2 *Consider a single-server queue with K classes of users. Assume the service requirements are exponentially distributed with $c_1\mu_1 \geq \ldots \geq c_K\mu_K$. If the weights ϕ and $\tilde{\phi}$ are such that class i obtains a relatively larger weight (compared to class $i+1$) under $\tilde{\phi}$ than under ϕ, that is*

$$\phi_i/\phi_{i+1} < \tilde{\phi}_i/\tilde{\phi}_{i+1}, \quad i = 1, \ldots, K-1,$$

then

$$\sum_{j=1}^{K} c_j \mathbb{E}(N_j^{GPS(\phi)}(t)) \geq \sum_{j=1}^{K} c_j \mathbb{E}(N_j^{GPS(\tilde{\phi})}(t)),$$

and

$$\sum_{j=1}^{K} c_j \mathbb{E}(N_j^{DPS(\phi)}(t)) \geq \sum_{j=1}^{K} c_j \mathbb{E}(N_j^{DPS(\tilde{\phi})}(t)), \quad \forall\, t \geq 0.$$

In the next example we perform numerical experiments that support Conjecture 6.2 for a single-server system with three classes.

Example 6.3 (Numerical experiments for GPS and DPS) *We consider a single server with fixed unit capacity and three classes of users with exponentially distributed service requirements and Poisson arrivals. We consider both GPS and DPS with weights $\phi_i(r) = \Omega(r)\cdot r^{K-i}$, $r \geq 1$, and $\Omega(r) = 1/(\sum_{i=0}^{K-1} r^i)$ a normalization constant. Note that $\phi_i/\phi_{i+1} = r$, $i = 1, \ldots, K$. Hence, as the parameter r increases, class i obtains relatively a larger weight compared to class $i+1$. We choose $\mu_1 = 2, \mu_2 = 1$ and $\mu_3 = 0.5$, hence we expect that the functions $\mathbb{E}(\sum_{i=1}^{K} N_i^{GPS(\phi(r))})$ and $\mathbb{E}(\sum_{i=1}^{K} N_i^{DPS(\phi(r))})$ are decreasing in r. When $r \to \infty$, both $GPS(r)$ and $DPS(r)$ become a priority rule that gives preemptive priority to class 1, and if class 1 is empty, it serves class 2. Since $\mu_1 > \mu_2 > \mu_3$, this policy minimizes the total mean number of users present in the system (follows from the optimality of the $c\mu$-rule).*
For GPS with weights $\phi_i(r)$ we simulated the system and Figure 5 a) plots the total mean number of users as a function of the parameter r. We observe that the total mean number of users indeed reduces as r increases.
In Figure 5 b) we consider a single server under $DPS(r)$ and plot the mean total number of users as a function of the parameter r. The total mean number of users was obtained by solving a system of linear equations as given in [12]. When $r = 1$, the policy reduces to PS, hence $\mathbb{E}(\sum_{i=1}^{K} N_i^{DPS(\phi(1))}) = \mathbb{E}(\sum_{i=1}^{K} N_i^{PS}) = \frac{\rho_1+\rho_2+\rho_3}{1-\rho_1+\rho_2+\rho_3}$. We observe that the mean total number of users is again decreasing in r.

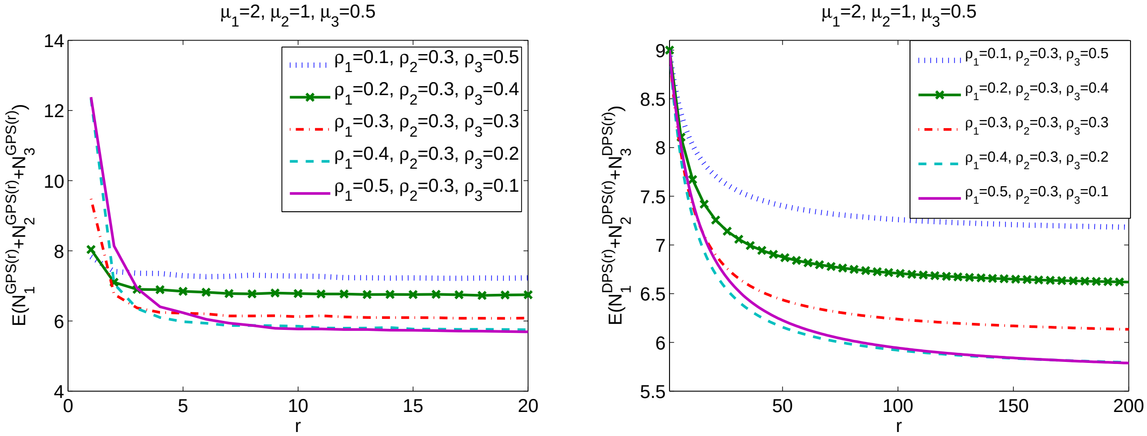

Figure 5: a) Total mean number of users under GPS policies, b) Total mean number of users under DPS policies.

For a single server with more than two classes, the framework and results as developed in Section 3 and in particular Property 3.1 are not applicable. Therefore, in the next subsection we develop a similar analysis as in Section 3, but now for a single-server system with an arbitrary number of classes. Unfortunately, this sample-path framework does not allow a full comparison of either two DPS or two GPS policies for more than two classes. This will be explained as well in the next subsection.

6.2 Comparison results for the multi-class single-server queue

We now focus on a single-server system with K classes of users, $K \geq 2$. We only consider efficient policies, which in a single-server scenario is equivalent to the work-conserving property. In this section we develop sample-path results similar to the ones obtained in Section 3. We start by giving sufficient conditions on two policies in order to compare sample-path wise two policies.

Property 6.4 *Let π and $\tilde{\pi}$ be two work-conserving policies such that for any $k = 1, \ldots, K-1$ we have*

$$\sum_{i=1}^{k} s_i^{\pi}(\vec{n}^{\pi}) \leq \sum_{i=1}^{k} s_i^{\tilde{\pi}}(\vec{n}^{\tilde{\pi}}), \tag{20}$$

for all states $\vec{n}^{\tilde{\pi}}$ and $\vec{n}^{\pi}$ that satisfy the following:

- *$n_1^{\pi} \geq n_1^{\tilde{\pi}}$, $n_k^{\pi} \leq n_k^{\tilde{\pi}}$, $n_{k+1}^{\pi} \geq n_{k+1}^{\tilde{\pi}}$ and $n_K^{\pi} \leq n_K^{\tilde{\pi}}$.*
- *If $n_k^{\tilde{\pi}} = 0$, then in addition $n_{k-1}^{\pi} \leq n_{k-1}^{\tilde{\pi}}$. If $n_k^{\tilde{\pi}} = 0$ and $n_{k-1}^{\tilde{\pi}} = 0$, then in addition $n_{k-2}^{\pi} \leq n_{k-2}^{\tilde{\pi}}$. Etc.*
- *If $n_{k+1}^{\pi} = 0$, then in addition $n_{k+2}^{\pi} \geq n_{k+2}^{\tilde{\pi}}$. If $n_{k+1}^{\pi} = 0$ and $n_{k+2}^{\pi} = 0$, then in addition $n_{k+3}^{\pi} \geq n_{k+3}^{\tilde{\pi}}$. Etc.*

Property 6.4 represents a weak notion of priority, with strict priority as a special case. When two policies satisfy Property 6.4, we can derive the following sample-path comparison result.

Proposition 6.5 *Let π and $\tilde{\pi}$ be two work-conserving policies that satisfy Property 6.4 and consider the same realizations of the arrival processes and service requirements. If $\sum_{i=1}^{m} W_i^{\pi}(0) \geq \sum_{i=1}^{m} W_i^{\tilde{\pi}}(0)$, $m = 1, \ldots, K-1$, and $\sum_{i=1}^{K} W_i^{\pi}(0) = \sum_{i=1}^{K} W_i^{\tilde{\pi}}(0)$, then for all $t \geq 0$*

$$\sum_{i=1}^{m} \left(S_i^{\pi}(t) - W_i^{\pi}(0)\right) \leq \sum_{i=1}^{m} \left(S_i^{\tilde{\pi}}(t) - W_i^{\tilde{\pi}}(0)\right), \quad m = 1, \ldots, K. \tag{21}$$

For $m = K$, (21) holds with equality.
In particular we have

$$N_1^{\pi}(t) \geq N_1^{\tilde{\pi}}(t), \quad N_K^{\pi}(t) \leq N_K^{\tilde{\pi}}(t), \tag{22}$$

and

$$\sum_{i=1}^{m} W_i^{\pi}(t) \geq \sum_{i=1}^{m} W_i^{\tilde{\pi}}(t), \quad m = 1, \ldots, K. \tag{23}$$

For $m = K$, (23) holds with equality.

Proof: Equation (23) follows from (1) and equation (21). The first relation in equation (22) follows from equation (23) with $m = 1$, since the intra-class policy is FCFS and the k-th most recently arrived class-1 user before time t has the same (original) service requirement under both policies. Similarly, the second relation in equation (22) follows from equation (23) with $m = K-1$ and $\sum_{i=1}^{K} W_i^{\pi}(t) = \sum_{i=1}^{K} W_i^{\tilde{\pi}}(t)$. Therefore, it suffices to prove equation (21).
The policies are work-conserving, so $\sum_{i=1}^{K} W_i^{\pi}(0) = \sum_{i=1}^{K} W_i^{\tilde{\pi}}(0)$ gives that $\sum_{i=1}^{K} S_i^{\pi}(t) = \sum_{i=1}^{K} S_i^{\tilde{\pi}}(t)$, and hence (21) holds with equality for $m = K$.
Equation (21) for $m < K$ is proved by contradiction. Let t be the first time epoch at which (21) is violated for some k, $1 \leq k \leq K-1$. So we have $\sum_{i=1}^{k}(S_i^{\pi}(t) - W_i^{\pi}(0)) = \sum_{i=1}^{k}(S_i^{\tilde{\pi}}(t) - W_i^{\tilde{\pi}}(0))$ and $\sum_{i=1}^{k} s_i^{\pi}(\vec{N}^{\pi}(t^+)) > \sum_{i=1}^{k} s_i^{\tilde{\pi}}(\vec{N}^{\tilde{\pi}}(t^+))$ (with strict inequality), but $\sum_{i=1}^{m}(S_i^{\pi}(t) - W_i^{\pi}(0)) \leq \sum_{i=1}^{m}(S_i^{\tilde{\pi}}(t) - W_i^{\tilde{\pi}}(0))$, for $m \neq k$. Hence,

$$S_1^{\pi}(t) - W_1^{\pi}(0) \leq S_1^{\tilde{\pi}}(t) - W_1^{\tilde{\pi}}(0), \qquad S_k^{\pi}(t) - W_k^{\pi}(0) \geq S_k^{\tilde{\pi}}(t) - W_k^{\tilde{\pi}}(0),$$
$$S_{k+1}^{\pi}(t) - W_{k+1}^{\pi}(0) \leq S_{k+1}^{\tilde{\pi}}(t) - W_{k+1}^{\tilde{\pi}}(0), \qquad S_K^{\pi}(t) - W_K^{\pi}(0) \geq S_K^{\tilde{\pi}}(t) - W_K^{\tilde{\pi}}(0)$$

Together with (1), we obtain $W_1^{\pi}(t) \geq W_1^{\tilde{\pi}}(t), W_k^{\pi}(t) \leq W_k^{\tilde{\pi}}(t), W_{k+1}^{\pi}(t) \geq W_{k+1}^{\tilde{\pi}}(t)$ and $W_K^{\pi}(t) \leq W_K^{\tilde{\pi}}(t)$. Since the k-th class-j user under both policies has the same (original) service requirement and the intra-class policy is FCFS, we have as well

$$N_1^{\pi}(t) \geq N_1^{\tilde{\pi}}(t), \quad N_k^{\pi}(t) \leq N_k^{\tilde{\pi}}(t), \quad N_{k+1}^{\pi}(t) \geq N_{k+1}^{\tilde{\pi}}(t) \quad \text{and} \quad N_K^{\pi}(t) \leq N_K^{\tilde{\pi}}(t).$$

Since $\{N_i(t)\}_{t \geq 0}$ is a piece-wise constant process and is right continuous, the same holds at time t^+: $N_1^{\pi}(t^+) \geq N_1^{\tilde{\pi}}(t^+), N_k^{\pi}(t^+) \leq N_k^{\tilde{\pi}}(t^+), N_{k+1}^{\pi}(t^+) \geq N_{k+1}^{\tilde{\pi}}(t^+)$ and $N_K^{\pi}(t^+) \leq N_K^{\tilde{\pi}}(t^+)$.
Note that if $N_k^{\tilde{\pi}}(t^+) = 0$, then $S_k^{\pi}(t) - W_k^{\pi}(0) = S_k^{\tilde{\pi}}(t) - W_k^{\tilde{\pi}}(0)$ and hence $\sum_{i=1}^{k-1}(S_i^{\pi}(t) - W_i^{\pi}(0)) = \sum_{i=1}^{k-1}(S_i^{\tilde{\pi}}(t) - W_i^{\tilde{\pi}}(0))$. So $S_{k-1}^{\pi}(t) - W_{k-1}^{\pi}(0) \geq S_{k-1}^{\tilde{\pi}}(t) - W_{k-1}^{\tilde{\pi}}(0)$ and by (1) we obtain $N_{k-1}^{\pi}(t^+) \leq N_{k-1}^{\tilde{\pi}}(t^+)$. Now if also $N_{k-1}^{\tilde{\pi}}(t^+) = 0$, then we obtain in the same way that $N_{k-2}^{\pi}(t^+) \leq N_{k-2}^{\tilde{\pi}}(t^+)$, etc.
Also note that if $N_{k+1}^{\pi}(t^+) = 0$, then $S_{k+1}^{\pi}(t) - W_{k+1}^{\pi}(0) = S_{k+1}^{\tilde{\pi}}(t) - W_{k+1}^{\tilde{\pi}}(0)$ and hence $\sum_{i=1}^{k+1}(S_i^{\pi}(t) - W_i^{\pi}(0)) = \sum_{i=1}^{k+1}(S_i^{\tilde{\pi}}(t) - W_i^{\tilde{\pi}}(0))$. So $S_{k+2}^{\pi}(t) - W_{k+2}^{\pi}(0) \leq S_{k+2}^{\tilde{\pi}}(t) - W_{k+2}^{\tilde{\pi}}(0)$ and by (1) we obtain $N_{k+2}^{\pi}(t^+) \geq N_{k+2}^{\tilde{\pi}}(t^+)$. Now if also $N_{k+2}^{\pi}(t^+) = 0$, then we obtain in the same way that $N_{k+3}^{\pi}(t^+) \geq N_{k+3}^{\tilde{\pi}}(t^+)$, etc.
So at time t^+ we are in states $\vec{N}^{\pi}(t^+)$ and $\vec{N}^{\tilde{\pi}}(t^+)$ that satisfy Property 6.4 and hence $\sum_{i=1}^{k} s_i^{\pi}(\vec{N}^{\pi}(t^+)) \leq \sum_{i=1}^{k} s_i^{\tilde{\pi}}(\vec{N}^{\tilde{\pi}}(t^+))$. This contradicts the initial assumption. □

Every work-conserving policy gives a stable system whenever possible. However, for a subset of the classes, the stability conditions can still depend on the policy being employed. We have the following result:

Corollary 6.6 *Assume policies π and $\tilde{\pi}$ satisfy Property 6.4. If classes $1, 2, \ldots, m$ are stable under policy π, then these classes are stable under policy $\tilde{\pi}$ as well, in the sense that the system is empty under policy $\tilde{\pi}$ whenever it is empty under policy π.*
In particular, if the empty state is positive-recurrent under policy π in the case of Poisson arrivals, then it is positive-recurrent under policy $\tilde{\pi}$ as well.

Proof: If $\sum_{i=1}^{m} W_i^{\pi}(t) = 0$, then we obtain from Proposition 6.5 that $\sum_{i=1}^{m} W_i^{\tilde{\pi}}(t) = 0$. The second assertion is a direct implication of the first one. □

The following proposition states the analogous version of Proposition 3.5.

Proposition 6.7 *Assume the service requirements are exponentially distributed. Let π and $\tilde{\pi}$ be two policies that satisfy Property 6.4 and assume the system is stable. If $c_1\mu_1 \geq c_2\mu_2 \geq \ldots \geq c_K\mu_K$, then*

$$\sum_{i=1}^{K} c_i \mathbb{E}(N_i^{\pi}(t)) \geq \sum_{i=1}^{K} c_i \mathbb{E}(N_i^{\tilde{\pi}}(t)), \quad \forall\ t \geq 0.$$

Proof: Assume at time $t = 0$ the conditions as stated in Proposition 6.5 are satisfied. From Proposition 6.5 we obtain $\sum_{i=1}^{m} W_i^{\pi}(t) \geq \sum_{i=1}^{m} W_i^{\tilde{\pi}}(t)$. Since π is non-anticipating and the service requirements are exponentially distributed we obtain

$$\sum_{i=1}^{m} \frac{1}{\mu_i} \mathbb{E}(N_i^{\pi}(t)) \geq \sum_{i=1}^{m} \frac{1}{\mu_i} \mathbb{E}(N_i^{\tilde{\pi}}(t)) \tag{24}$$

for $m \leq K$. Define $P_m^{\pi}(t) := \sum_{i=1}^{m} \frac{1}{\mu_i} \mathbb{E}(N_i^{\pi}(t))$. So $P_m^{\pi}(t) \geq P_m^{\tilde{\pi}}(t),\ m = 1, \ldots, K$ and hence

$$\begin{aligned}
\sum_{i=1}^{K} c_i \mathbb{E}(N_i^{\pi}(t)) &= (c_1\mu_1 - c_2\mu_2)P_1^{\pi}(t) + (c_2\mu_2 - c_3\mu_3)P_2^{\pi}(t) + \ldots + c_K\mu_K P_K^{\pi}(t) \\
&\geq (c_1\mu_1 - c_2\mu_2)P_1^{\tilde{\pi}}(t) + (c_2\mu_2 - c_3\mu_3)P_2^{\tilde{\pi}}(t) + \ldots + c_K\mu_K P_K^{\tilde{\pi}}(t) \\
&= \sum_{i=1}^{K} c_i \mathbb{E}(N_i^{\tilde{\pi}}(t)),
\end{aligned}$$

where we used that $c_1\mu_1 \geq c_2\mu_2 \geq \ldots \geq c_K\mu_K$. □

Example 6.8 (Optimality of the $c\mu$-rule:) *As mentioned before, for exponentially distributed service requirements, the $c\mu$-rule, i.e. the policy that gives preemptive priority to the class i with the maximum $c_i\mu_i$, minimizes the mean holding cost among all non-anticipating policies. For a time-varying multi-class single-server system, this was shown in [31]. In fact this also follows from Proposition 6.7. Assume $c_1\mu_1 \geq c_2\mu_2 \geq \ldots \geq c_K\mu_K$. Denote the $c\mu$-rule by $\tilde{\pi}$, and consider an arbitrary non-anticipating policy π. Whenever $\sum_{i=1}^{k} n_i^{\tilde{\pi}} > 0$ we have that $\sum_{i=1}^{k} s_i^{\tilde{\pi}}(\vec{n}^{\tilde{\pi}}) = C(t)$ and hence (20) is satisfied for these states. Now assume $\sum_{i=1}^{k} n_i^{\tilde{\pi}} = 0$. Since $n_i^{\tilde{\pi}} = 0$ for all $i \leq k$, the corresponding states $\vec{n}^{\pi}$ we have to consider in Property 6.4 should satisfy $n_i^{\pi} \leq n_i^{\tilde{\pi}} = 0$ for all $i \leq k$, that is $\sum_{i=1}^{k} n_i^{\pi} = 0$ as well. But then (20) is by definition satisfied. Hence Property 6.4 is satisfied and the optimality of the $c\mu$-rule follows now from Proposition 6.7.*

Proposition 6.7, combined with Property 6.4 gives sufficient conditions in order to compare the mean holding cost under two policies. When $K = 2$, Property 6.4 reduces to the rather natural condition $s_1^{\pi}(\vec{n}) \leq s_1^{\tilde{\pi}}(\vec{n})$ for all states $\vec{n}$. This is for example satisfied by either two DPS policies or two GPS policies when $\phi_1 \leq \tilde{\phi}_1$. Unfortunately, for more than two classes Property 6.4 fails to hold for any two DPS policies. For a GPS system, Property 6.4 is satisfied under more stringent conditions than the ones stated in Conjecture 6.2. For example, for the case of three classes it can be checked that for two GPS disciplines, GPS(ϕ) and GPS($\tilde{\phi}$), Property 6.4 is equivalent to

$$\frac{\phi_1}{\phi_1+\phi_2} \leq \tilde{\phi}_1, \qquad \frac{\phi_1}{\phi_1+\phi_3} \leq \frac{\tilde{\phi}_1}{\tilde{\phi}_1+\tilde{\phi}_3} \qquad \text{and} \qquad \phi_3 \geq \frac{\tilde{\phi}_3}{\tilde{\phi}_2+\tilde{\phi}_3}. \tag{25}$$

Hence, (25) is a sufficient condition to compare the mean holding cost under GPS(ϕ) and GPS($\tilde{\phi}$). If we choose as weights $\phi_i(r) = \Omega(r) \cdot r^{K-i}$, $r > 1$ (as considered in Example 6.3), equation (25) is equivalent to $1 \leq r$ and $\tilde{r} \geq r + r^2$. We would expect the comparison result already to hold for all $\tilde{r} \geq r$, so this shows that there is still a gap of length r^2. For an arbitrary number of classes, the sufficient conditions in order for Conjecture 6.2 to hold for GPS can be obtained as well, however, the derivations become very cumbersome.

In this section we used sample-path inequalities as given in (21) in order to compare the mean holding cost under two different policies. Property 6.4 is a sufficient (but not necessary) condition for these sample-path inequalities to hold. For DPS and GPS, this property is not (always) satisfied. In fact, the counterexample below illustrates for the case of three classes that the sample-path inequalities (21) do not need to hold for either two DPS policies or two GPS policies that satisfy the conditions of Conjecture 6.2. This indicates that for more than two classes Conjecture 6.2 may not be proved using sample-path arguments and requires a different kind of approach.

Example 6.9 (Counterexamples for DPS and GPS) *We give a counterexample for inequality (21) that is valid for both DPS and GPS, since the sample-path will exactly be the same under both policies. Consider a system with three classes, and consider the two policies with weight vectors $\phi = (2,1,1)$ and $\tilde{\phi} = (\infty,1,1)$, respectively. It is easy to verify that the vectors ϕ and $\tilde{\phi}$ satisfy the condition of Conjecture 6.2. Assume that at time $t = 0$ there is one user in every class, that is, $N^{\pi}(0) = N^{\tilde{\pi}}(0) = (1,1,1)$ and their service requirements are respectively 4, 10 and 1 under both policies π and $\tilde{\pi}$. At time $t = 6$ a class-3 user arrives with a strictly positive service requirement. Let us analyze the evolution under both disciplines over time:*

- *Policy π: In the interval $[0,4)$ all users share the capacity according to the weights. At time $t = 4$ the class-3 user departs the system and the remaining service requirements of the class-1 and the class-2 user are 2 and 9, respectively. In the interval $[4,6)$ the class-1 and class-2 users will share the capacity according to their weights, thus at time $t = 6$ the remaining service requirements for the class-1 and class-2 users are $\frac{2}{3}$ and $\frac{25}{3}$, respectively. It follows that $S_1^{\pi}(6)+S_2^{\pi}(6) = 4+10-\frac{2}{3}-\frac{25}{3} = 5$.*

- *Policy $\tilde{\pi}$: In the interval $[0,4)$ only class 1 will be served and it departs at time $t = 4$. In the interval $[4,6)$ the class-2 and class-3 users will equally share the capacity. At time $t = 6$ the class-3 user departs and the class-2 user has a remaining service requirement of 9. It follows that $S_1^{\tilde{\pi}}(6) + S_2^{\tilde{\pi}}(6) = 4 + 10 - 9 = 5$.*

Due to the new arrival at $t = 6$ it follows that $s_1^{\pi}(\vec{N}^{\pi}(6^+)) + s_2^{\pi}(\vec{N}^{\pi}(6^+)) = \frac{3}{4}$ whereas $s_1^{\tilde{\pi}}(\vec{N}^{\tilde{\pi}}(6^+)) + s_2^{\tilde{\pi}}(\vec{N}^{\tilde{\pi}}(6^+)) = \frac{1}{2}$. This together with the fact that $S_1^{\pi}(6) + S_2^{\pi}(6) = S_1^{\tilde{\pi}}(6) + S_2^{\tilde{\pi}}(6)$ implies that $S_1^{\pi}(6^+) + S_2^{\pi}(6^+) > S_1^{\tilde{\pi}}(6^+) + S_2^{\tilde{\pi}}(6^+)$, which contradicts (21) for $m = 2$.

In the following remark we explain that Conjecture 6.2 does not follow either from results in [24, 25] and hence that a novel approach is needed.

Remark 6.10 *Assume the processes* $\{\vec{N}^{\pi}(t)\}_{t\geq 0}$ *and* $\{\vec{N}^{\tilde{\pi}}(t)\}_{t\geq 0}$ *are two continuous-time Markov processes (hence Poisson arrivals with exponentially distributed service requirements and* $C(t) = C$*). From Remark 3.6 we readily see that the conditions on the policies* π *and* $\tilde{\pi}$ *in order to obtain* $\{\sum_{i=1}^{K} N_i^{\pi}(t)\}_{t\geq 0} \geq_{st} \{\sum_{i=1}^{K} N_i^{\tilde{\pi}}(t)\}_{t\geq 0}$ *for any initial states* $\sum_{i=1}^{K} N_i^{\pi}(0) \geq_{st} \sum_{i=1}^{K} N_i^{\tilde{\pi}}(0)$*, are only satisfied when* $\mu_i = \mu$ *for all* i*. Consider for example the two states* $\vec{n}^{\pi} = e_k$ *and* $\vec{n}^{\tilde{\pi}} = e_j$*, where* e_j *denotes a vector with the* j*-th component equal to 1 and all other components equal to 0. Then the condition as stated in Remark 3.6 becomes* $\sum_{i=1}^{K} \mu_i s_i^{\pi}(\vec{n}^{\pi}) = \mu_k \leq \sum_{i=1}^{K} \mu_i s_i^{\tilde{\pi}}(\vec{n}^{\tilde{\pi}}) = \mu_j$*, see also [24, 25]. However, for the states* $\vec{n}^{\pi} = e_j$ *and* $\vec{n}^{\tilde{\pi}} = e_k$ *we obtain similarly that we need* $\mu_j \leq \mu_k$*. So only when* $\mu_i = \mu$ *for all* i*, the conditions are satisfied, but this is not very interesting.*

The necessary and sufficient conditions in order to obtain a similar comparison result as in Proposition 6.5, i.e. $\{N_1^{\pi}(t)\}_{t\geq 0} \geq_{st} \{N_1^{\tilde{\pi}}(t)\}_{t\geq 0}$ *and* $\{N_K^{\pi}(t)\}_{t\geq 0} \leq_{st} \{N_K^{\tilde{\pi}}(t)\}_{t\geq 0}$ *given that* $N_1^{\pi}(0) \geq N_1^{\tilde{\pi}}(0)$ *and* $N_K^{\pi}(0) \leq N_K^{\tilde{\pi}}(0)$*, are*

$$s_1^{\pi}(\vec{n}^{\pi}) \leq s_1^{\tilde{\pi}}(\vec{n}^{\tilde{\pi}}) \quad \textit{for all} \quad n_1^{\pi} = n_1^{\tilde{\pi}} \textit{ and } n_K^{\pi} \leq n_K^{\tilde{\pi}}, \tag{26}$$

$$s_K^{\pi}(\vec{n}^{\pi}) \geq s_K^{\tilde{\pi}}(\vec{n}^{\tilde{\pi}}) \quad \textit{for all} \quad n_1^{\pi} \geq n_1^{\tilde{\pi}} \textit{ and } n_K^{\pi} = n_K^{\tilde{\pi}}, \tag{27}$$

see [24, 25]. In a queueing context, this can only be satisfied when policy $\tilde{\pi}$ *gives preemptive priority to class 1 (see equation (26) with states such that* $n_2^{\pi} = \ldots = n_K^{\pi} = 0$*) and policy* π *gives preemptive priority to class* K *(see equation (27) with states such that* $n_1^{\tilde{\pi}} = \ldots = n_{K-1}^{\tilde{\pi}} = 0$*). In particular, for any two GPS policies or two DPS policies (with non-degenerate weights) the inequalities (26) and (27) do not hold.*

7 Conclusion and future work

In this paper we have studied monotonicity properties for multi-class stochastic networks and have obtained comparison results for the performance under two different policies in terms of stability, mean holding cost and the mean overall delay. The results were obtained by using a natural coupling, namely by choosing the same realization of inter-arrival times and service requirements for both processes. Sample-path comparisons were obtained for the workload and the number of users of certain classes.

The results were applied to a linear network and a multi-class single-server system. In future work, it might be interesting to consider different types of networks, like a star or grid network, and use the same approach in order to compare the performance of different policies.

For the linear network we proved monotonicity results for the mean holding cost under α-fair policies. In the numerical section, we observed additional monotonicity properties. For instance we have strong evidence to believe that the total mean number of users in the system is monotone in $\alpha \in [1, \infty)$ when the other parameters are kept fixed. Another interesting observation from the numerical section is that the total mean number of users is monotone in μ_0 for given load ρ_0, when the other parameters are kept fixed. There is no hope that this latter property can be proved using sample-path arguments, since this requires the same realizations for the service requirements. When we compare the two stochastic processes for different values of μ_0, this can no longer be done.

For the single-server system, it is reasonable to expect that for popular weighted time-sharing policies like DPS and GPS, monotonicity results for expected performance measures hold under natural conditions on the weights, see Conjecture 6.2. We were able to prove this for some special cases using a sample-path argument. The other cases remain as a challenging topic for further research.

Acknowledgment

The authors are grateful to Matthieu Jonckheere (Eindhoven University of Technology, The Netherlands) for helpful discussions on stochastic comparisons.

References

[1] E. Altman, K.E. Avrachenkov, and U. Ayesta. A survey on discriminatory processor sharing. *Queueing Systems*, 53:53–63, 2006.

[2] K.E. Avrachenkov, U. Ayesta, P. Brown, and R. Núñez-Queija. Discriminatory processor sharing revisited. In *Proceedings of IEEE INFOCOM*, Miami, FL, USA, 2005.

[3] T. Bonald, S.C. Borst, and A. Proutière. Inter-cell coordination in wireless data networks. *European Transactions on Telecommunications*, 17:303–312, 2006.

[4] T. Bonald and L. Massoulié. Impact of fairness on Internet performance. In *Proceedings of ACM SIGMETRICS/Performance*, pages 82–91, Boston MA, USA, 2001.

[5] T. Bonald and A. Proutière. On stochastic bounds for monotonic processor sharing networks. *Queueing Systems*, 47:81–106, 2004.

[6] S.C. Borst, M. Jonckheere, and L. Leskelä. Stability of parallel queueing systems with coupled service rates. *Discrete Event Dynamic Systems*, 18:447–472, 2008.

[7] S.C. Borst, M. Mandjes, and M.J.G. van Uitert. Generalized processor sharing queues with heterogeneous traffic classes. *Advances in Applied Probability*, 35:806–845, 2003.

[8] S.K. Cheung, J.L. van den Berg, and R.J. Boucherie. Insensitive bounds for the moments of the sojourn time distribution in the M/G/1 processor-sharing queue. *Queueing Systems*, 53:7–18, 2006.

[9] J.G. Dai. On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. *Annals of Applied Probability*, 5:49–77, 1995.

[10] A. Demers, S. Keshav, and S. Shenker. Analysis and simulation of a Fair Queueing Algorithm. In *Proceedings of ACM SIGCOMM*, pages 1–12, 1989.

[11] M. El-Taha and S. Stidham. *Sample-path analysis of queueing systems.* Kluwer Academic Publishers, 1999.

[12] G. Fayolle, I. Mitrani, and R. Iasnogorodski. Sharing a processor among many job classes. *Journal of the ACM*, 27:519–532, 1980.

[13] W.N. Kang, F.P. Kelly, N.H. Lee, and R.J. Williams. Fluid and Brownian approximations for an Internet congestion control model. In *Proceedings of IEEE CDC*, pages 3938–3943, 2004.

[14] W.N. Kang, F.P. Kelly, N.H. Lee, and R.J. Williams. State space collapse and diffusion approximation for a network operating under a fair bandwidth sharing policy. *Annals of Applied Probability*, 2009. To appear.

[15] F.P. Kelly. *Stochastic Networks and Reversibility.* Wiley, Chichester, 1979.

[16] F.P. Kelly. Fairness and stability of end-to-end congestion control. *European Journal of Control*, 9:159–176, 2003.

[17] F.P. Kelly and R.J. Williams. Fluid model for a network operating under a fair bandwidth-sharing policy. *Annals of Applied Probability*, 14:1055–1083, 2004.

[18] G. van Kessel, R. Núñez-Queija, and S.C. Borst. Differentiated bandwidth sharing with disparate flow sizes. In *Proceedings of IEEE INFOCOM*, Miami, FL, USA, 2005.

[19] B. Kim and J. Kim. Comparison of DPS and PS systems according to DPS weights. *IEEE Communications Letters*, 10(7):558–560, 2006.

[20] L. Kleinrock. Time-shared systems: A theoretical treatment. *Journal of the ACM*, 14:242–261, 1967.

[21] P. Lieshout, S.C. Borst, and M. Mandjes. Heavy-traffic approximations for linear networks operating under alpha-fair bandwidth-sharing policies. In *Proceedings of ValueTools*, Pisa, Italy, 2006.

[22] J. Liu, A. Proutière, Y. Yi, M. Chiang, and V.H. Poor. Flow-level stability of data networks with non-convex and time-varying rate regions. In *Proceedings of ACM SIGMETRICS*, pages 239–250, San Diego, CA, USA, 2007.

[23] Z. Liu, P. Nain, and D. Towsley. Sample path methods in the control of queues. *Queueing Systems*, 21:293–335, 1995.

[24] F.J. López and G. Sanz. Markovian couplings staying in arbitrary subsets of the state space. *Journal of Applied Probability*, 39:197–212, 2002.

[25] W.A. Massey. Stochastic orderings for Markov processes on partially ordered spaces. *Mathematics of Operations Research*, 12:350–367, 1987.

[26] L. Massoulié and J.W. Roberts. Bandwidth sharing and admission control for elastic traffic. *Telecommunication Systems*, 15:185–201, 2000.

[27] L. Massoulié and J.W. Roberts. Bandwidth sharing: objectives and algorithms. *IEEE/ACM Transactions on Networking*, 10:320–328, 2002.

[28] S.P. Meyn and R.L. Tweedie. *Markov chains and stochastic stability.* Springer-Verlag, 1993.

[29] J. Mo and J. Walrand. Fair end-to-end window-based congestion control. *IEEE/ACM Transactions on Networking*, 8:556–567, 2000.

[30] A.M. Muller and D. Stoyan. *Comparison methods for stochastic models and risks.* J. Wiley & Sons, 2002.

[31] P. Nain and D. Towsley. Optimal scheduling in a machine with stochastic varying processing rate. *IEEE Transactions on Automatic Control*, 39:1853–1855, 1994.

[32] A.K. Parekh and R.G. Gallager. A generalized processor sharing approach to flow control in integrated services networks: the single-node case. *IEEE/ACM Transactions on Networking*, 1:344–357, 1993.

[33] R. Righter and J.G. Shanthikumar. Scheduling multiclass single server queueing systems to stochastically maximize the number of successful departures. *Probability in the Engineering and Informational Sciences*, 3:323–333, 1989.

[34] L.E. Schrage and L.W. Miller. The queue M/G/1 with the shortest remaining processing time discipline. *Operations Research*, 14:670–684, 1966.

[35] M. Shaked and J.G. Shanthikumar. *Stochastic orders and their applications.* Academic Press, 1993.

[36] D.R. Smith. A new proof of the optimality of the shortest remaining processing time discipline. *Operations Research*, 26:197–199, 1978.

[37] I.M. Verloop. *Efficient flow scheduling in resource-sharing networks.* Master Thesis, Utrecht University, available at www.cwi.nl/∼maaike/articles/thesisVerloop.pdf, 2005.

[38] I.M. Verloop, U. Ayesta, and S.C. Borst. Comparison of bandwidth-sharing policies in a linear network. In *Proceedings of ValueTools*, Athens, Greece, 2008.

[39] I.M. Verloop and R. Núñez-Queija. Assessing the efficiency of resource allocations in bandwidth-sharing networks. *Performance Evaluation*, 66:59–77, 2009.

[40] H. Viswanathan and K. Kumaran. Rate scheduling in multiple antenna downlink wireless systems. *IEEE Transactions on Communications*, 53:645–655, 2005.

[41] A. Wierman, M. Harchol-Balter, and T. Osogami. Nearly insensitive bounds on SMART scheduling. In *Proceedings of ACM SIGMETRICS*, 2005.

Appendix A: Proof of Lemma 5.1

For a given state $\vec{n}$, the α-fair allocation is the vector $(s_0, s_1, \ldots, s_L)$ that solves the optimization problem (7). If $n_i > 0$, then $s_i = C_i - s_0$. The objective function in (7) expressed in terms of the value s_0, is concave in s_0. Taking the derivative of (7) with respect to s_0 and setting it equal to zero, we obtain that $s_0^{\pi(\alpha,w)}(\vec{n})$ satisfies

$$w_0 \cdot n_0^\alpha \cdot (s_0^{\pi(\alpha,w)}(\vec{n}))^{-\alpha} = \sum_{i=1}^{L} w_i \cdot n_i^\alpha \cdot (C_i - s_0^{\pi(\alpha,w)}(\vec{n}))^{-\alpha},$$

or equivalently

$$1=\sum_{i=1}^{L}\frac{w_i}{w_0}\Big(\frac{n_i}{n_0}\frac{s_0^{\pi(\alpha,w)}(\vec{n})}{C_i-s_0^{\pi(\alpha,w)}(\vec{n})}\Big)^{\alpha}. \tag{28}$$

The function $\sum_{i=1}^{L}\frac{w_i}{w_0}\left(\frac{n_i}{n_0}\frac{s_0}{C_i-s_0}\right)^{\alpha}$ is non-decreasing in s_0. Hence, when either n_i or $\frac{w_i}{w_0}$ increases, by equality (28) the corresponding value of s_0 must decrease. Statements (i) and (iii) follow now immediately. Statement (ii) follows similarly by noting that

$$\begin{aligned}
\Big(\sum_{i=1}^{L}\frac{w_i}{w_0}\Big(\frac{n_i}{n_0}\frac{s_0^{\pi(\gamma,w)}(\vec{n})}{C_i-s_0^{\pi(\gamma,w)}(\vec{n})}\Big)^{\gamma}\Big)^{\frac{1}{\gamma}}=1 &= \Big(\sum_{i=1}^{L}\frac{w_i}{w_0}\Big(\frac{n_i}{n_0}\frac{s_0^{\pi(\beta,w)}(\vec{n})}{C_i-s_0^{\pi(\beta,w)}(\vec{n})}\Big)^{\beta}\Big)^{\frac{1}{\beta}}\\
&= \Big(\sum_{i=1}^{L}\frac{w_i}{w_0}\Big(\frac{n_i}{n_0}\frac{s_0^{\pi(\beta,w)}(\vec{n})}{C_i-s_0^{\pi(\beta,w)}(\vec{n})}\Big)^{\beta}\Big)^{\frac{r}{r\beta}}\\
&\geq \Big(\sum_{i=1}^{L}\frac{w_i}{w_0}\Big(\frac{n_i}{n_0}\frac{s_0^{\pi(\beta,w)}(\vec{n})}{C_i-s_0^{\pi(\beta,w)}(\vec{n})}\Big)^{r\beta}\Big)^{\frac{1}{r\beta}}\\
&= \Big(\sum_{i=1}^{L}\frac{w_i}{w_0}\Big(\frac{n_i}{n_0}\frac{s_0^{\pi(\beta,w)}(\vec{n})}{C_i-s_0^{\pi(\beta,w)}(\vec{n})}\Big)^{\gamma}\Big)^{\frac{1}{\gamma}},
\end{aligned}$$

with $r\beta=\gamma$ and $r>1$. Hence $s_0^{\pi(\beta,w)}(\vec{n})\leq s_0^{\pi(\gamma,w)}(\vec{n})$. □

Appendix B: Proof of Proposition 5.6

By the conjecture of [14], the scaled workload in a node is independent of α. In addition, it is stated that the diffusion scaled number of users, $\hat{\vec{N}}^{k,\pi(\alpha)}(t)$, converges in distribution as $k\to\infty$ to $\hat{\vec{N}}^{\pi(\alpha)}(t)=\Delta(\hat{\vec{V}}^{\pi(\alpha)}(t))$, where the lifting mapping $\Delta:\mathbf{R}_+^2\to\mathbf{R}_+^3$ is as defined in [14, 17]. This is equivalent to saying that there are $q_1(\alpha),q_2(\alpha)\geq 0$ such that

$$\hat{N}_i^{\pi(\alpha)}=\rho_i q_i(\alpha)^{\frac{1}{\alpha}},\quad i=1,2\quad\text{and}\quad \hat{N}_0^{\pi(\alpha)}=\rho_0(q_1(\alpha)+q_2(\alpha))^{\frac{1}{\alpha}}. \tag{29}$$

Using this representation for the number of users, we can describe the effect the parameter α has on the holding cost.

We compare the holding cost under two α-fair policies with parameters α_1 and α_2 for a given workload in both nodes. So we have at each time that the workload in a node is $\hat{V}_i^{\pi(\alpha)}=\hat{v}_i$, independent of α, $i=1,2$ (from now on we will drop the dependence on t). Using the representation as in (9) and the fact that $\rho_0+\rho_i=C_i$, this gives

$$(C_i-\rho_0)q_i(\alpha)^{1/\alpha}+\rho_0\frac{\mu_i}{\mu_0}(q_1(\alpha)+q_2(\alpha))^{1/\alpha}=\hat{v}_i. \tag{30}$$

Together with (29), the holding cost for an α-fair policy can be written as

$$\begin{aligned}
\sum_{i=0}^{2}c_i\hat{N}_i^{\pi(\alpha)} &= c_0\rho_0(q_1(\alpha)+q_2(\alpha))^{\frac{1}{\alpha}}+c_1(C_1-\rho_0)q_1(\alpha)^{\frac{1}{\alpha}}+c_2(C_2-\rho_0)q_2(\alpha)^{\frac{1}{\alpha}}\\
&= c_1\Big((C_1-\rho_0)q_1(\alpha)^{1/\alpha}+\rho_0\frac{\mu_1}{\mu_0}(q_1(\alpha)+q_2(\alpha))^{1/\alpha}\Big)\\
&\quad +c_2\Big((C_2-\rho_0)q_2(\alpha)^{1/\alpha}+\rho_0\frac{\mu_2}{\mu_0}(q_1(\alpha)+q_2(\alpha))^{1/\alpha}\Big)\\
&\quad +\frac{c_0\mu_0-c_1\mu_1-c_2\mu_2}{\mu_0}\rho_0(q_1(\alpha)+q_2(\alpha))^{\frac{1}{\alpha}}\\
&\overset{d}{=} c_1\hat{v}_1+c_2\hat{v}_2+\frac{c_0\mu_0-c_1\mu_1-c_2\mu_2}{\mu_0}\rho_0(1+f(\alpha)^{\alpha})^{\frac{1}{\alpha}}q_2(\alpha)^{\frac{1}{\alpha}},
\end{aligned} \tag{31}$$

where $f(\alpha):=\left(\frac{q_1(\alpha)}{q_2(\alpha)}\right)^{\frac{1}{\alpha}}$.

For $i = 2$, equation (30) gives

$$q_2(\alpha_1)^{1/\alpha_1}\Big(C_2 - \rho_0 + \rho_0\frac{\mu_2}{\mu_0}(1 + f(\alpha_1)^{\alpha_1})^{1/\alpha_1}\Big) = q_2(\alpha_2)^{1/\alpha_2}\Big(C_2 - \rho_0 + \rho_0\frac{\mu_2}{\mu_0}(1 + f(\alpha_2)^{\alpha_2})^{1/\alpha_2}\Big). \quad (32)$$

From (32) we conclude that

$$(1 + f(\alpha_1)^{\alpha_1})^{\frac{1}{\alpha_1}} q_2(\alpha_1)^{\frac{1}{\alpha_1}} < \ (=)\ (1 + f(\alpha_2)^{\alpha_2})^{\frac{1}{\alpha_2}} q_2(\alpha_2)^{\frac{1}{\alpha_2}} \quad (33)$$

if and only if

$$\frac{(1 + f(\alpha_1)^{\alpha_1})^{\frac{1}{\alpha_1}}}{(C_2 - \rho_0 + \rho_0\frac{\mu_2}{\mu_0}(1 + f(\alpha_1)^{\alpha_1})^{1/\alpha_1}} < \ (=)\ \frac{(1 + f(\alpha_2)^{\alpha_2})^{\frac{1}{\alpha_2}}}{(C_2 - \rho_0 + \rho_0\frac{\mu_2}{\mu_0}(1 + f(\alpha_2)^{\alpha_2})^{1/\alpha_2}},$$

if and only if

$$(1 + f(\alpha_1)^{\alpha_1})^{\frac{1}{\alpha_1}} < \ (=)\ (1 + f(\alpha_2)^{\alpha_2})^{\frac{1}{\alpha_2}}. \quad (34)$$

Let b be such that $\hat{v}_1 = b\hat{v}_2$. Assume without loss of generality $\frac{\rho_0/\mu_0}{(C_2-\rho_0)/\mu_2+\rho_0/\mu_0} \leq b \leq 1$. (For states with $b > 1$ the analysis is the same, with only the roles of nodes 1 and 2 interchanged.) Note that when $b = \frac{\rho_0/\mu_0}{(C_2-\rho_0)/\mu_2+\rho_0/\mu_0}$ we are on the edge of the cone as described in (10). In Lemma 1 (see below) we prove that $(1+f(\alpha)^{\alpha})^{\frac{1}{\alpha}}$ is indeed strictly decreasing in α when $\frac{\rho_0/\mu_0}{(C_2-\rho_0)/\mu_2+\rho_0/\mu_0} < b \leq 1$ and is constant when $b = \frac{\rho_0/\mu_0}{(C_2-\rho_0)/\mu_2+\rho_0/\mu_0}$, the edge of the cone. Assuming the probability mass is not all concentrated on the edge of the cone, we conclude from (31) and the equivalence between (33) and (34), that the mean holding cost is strictly decreasing (strictly increasing) in α when $c_1\mu_1 + c_2\mu_2 < c_0\mu_0$ ($c_1\mu_1 + c_2\mu_2 > c_0\mu_0$). □

The following lemma is used in the proof of Proposition 5.6.

Lemma 1 *The function* $(1 + f(\alpha)^{\alpha})^{1/\alpha}$ *with* $f(\alpha) = \left(\frac{q_1(\alpha)}{q_2(\alpha)}\right)^{\frac{1}{\alpha}}$, *is strictly decreasing in* α *when* $\frac{\rho_0/\mu_0}{(C_2-\rho_0)/\mu_2+\rho_0/\mu_0} < b \leq 1$ *and is constant when* $b = \frac{\rho_0/\mu_0}{(C_2-\rho_0)/\mu_2+\rho_0/\mu_0}$. *Here* b *satisfies* $\hat{v}_1 = b\hat{v}_2$.

Proof: From $\hat{v}_1 = b\hat{v}_2$ and (30) we obtain the relation

$$(C_1 - \rho_0)q_1(\alpha_i)^{\frac{1}{\alpha_i}} + (1-b)\rho_0\frac{\mu_1}{\mu_0}(q_1(\alpha_i) + q_2(\alpha_i))^{\frac{1}{\alpha_i}} \quad = \quad b(C_2 - \rho_0)\frac{\mu_1}{\mu_2}q_2(\alpha_i)^{\frac{1}{\alpha_i}},$$

hence when we divide both sides by $q_2(\alpha_i)^{\frac{1}{\alpha_i}}$, we obtain

$$(C_1 - \rho_0)f(\alpha_i) + (1-b)\rho_0\frac{\mu_1}{\mu_0}(1 + f(\alpha_i)^{\alpha_i})^{\frac{1}{\alpha_i}} \quad = \quad b(C_2 - \rho_0)\frac{\mu_1}{\mu_2}. \quad (35)$$

By (35) we have that $f(\alpha) = 0$ if and only if $b = \frac{\rho_0/\mu_0}{(C_2-\rho_0)/\mu_2+\rho_0/\mu_0}$.
Assume $b = \frac{\rho_0/\mu_0}{(C_2-\rho_0)/\mu_2+\rho_0/\mu_0}$. Then $f(\alpha) = 0$ and hence the function $(1 + f(\alpha)^{\alpha})^{\frac{1}{\alpha}}$ is constant.
Now assume $\frac{\rho_0/\mu_0}{(C_2-\rho_0)/\mu_2+\rho_0/\mu_0} < b \leq 1$. So $f(\alpha) > 0$ for all α. Take $\alpha_1 < \alpha_2$ and let $r > 1$ be such that $\alpha_2 = r\alpha_1$. Then

$$(1 + f(\alpha_2)^{\alpha_2})^{\frac{1}{\alpha_2}} = (1^{r\alpha_1} + f(r\alpha_1)^{r\alpha_1})^{\frac{1}{r\alpha_1}} < (1^{\alpha_1} + f(r\alpha_1)^{\alpha_1})^{\frac{r}{r\alpha_1}} = (1 + f(r\alpha_1)^{\alpha_1})^{\frac{1}{\alpha_1}}, \quad (36)$$

since $1 + f(r\alpha_1) > 1$. Suppose $f(\alpha_2) = f(r\alpha_1) \leq f(\alpha_1)$. From (36), we then obtain $(1 + f(\alpha_2)^{\alpha_2})^{\frac{1}{\alpha_2}} < (1 + f(\alpha_1)^{\alpha_1})^{\frac{1}{\alpha_1}}$. However, from (35) we know that when $f(\alpha_2) \leq f(\alpha_1)$, then $(1 + f(\alpha_2)^{\alpha_2})^{\frac{1}{\alpha_2}} \geq (1 + f(\alpha_1)^{\alpha_1})^{\frac{1}{\alpha_1}}$, hence we have a contradiction. So we conclude that $f(\alpha_2) > f(\alpha_1)$, and hence $f(\alpha)$ is strictly increasing in α and from (35) it then follows that $(1 + f(\alpha)^{\alpha})^{\frac{1}{\alpha}}$ is strictly decreasing in α. □