Comparison of bandwidth-sharing policies in a linear network

I.M. Verloop
CWI
P.O. Box 94079
1090 GB Amsterdam
The Netherlands
maaike@cwi.nl

U. Ayesta*
LAAS-CNRS
Université de Toulouse
7 Avenue Colonel Roche
31077 Toulouse Cedex
France
urtzi@laas.fr

S.C. Borst†
Department of Mathematics & Computer Science,
Eindhoven University of Technology
P.O. Box 513
5600 MB Eindhoven
The Netherlands
s.c.borst@tue.nl

ABSTRACT

In bandwidth-sharing networks, users of various classes require service from different subsets of shared resources simultaneously. These networks have been proposed to analyze the performance of wired and wireless networks. 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 in a linear bandwidth-sharing network. This allows us to compare the performance of the system under various policies in terms of stability, the mean overall delay and the weighted mean number of users.

For the important family of weighted α-fair policies, we derive stability results and establish monotonicity of the weighted mean number of users with respect to the fairness parameter α and the relative weights. In order to broaden the comparison results, we investigate a heavy-traffic regime and perform numerical experiments.

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. To gain insights into the performance of the system, researchers have resorted to deriving qualitative properties of the stochastic processes, among them stability, comparison results and performance bounds.

Stability of stochastic systems is a well-founded theory [18, 5]. Recently new results have been derived for systems with state-dependent (and time-varying) capacities. For example, in [13] the stability conditions for utility-based allocation policies in a time-varying scenario are characterized. In [4] 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 [24, 20]. In particular, stochastic comparison is often used. In the seminal paper [16] (see also [15]) 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 are not satisfied in the examples we will study in this paper. In this paper we will consider a special case of stochastic ordering: We use a sample-path approach to compare two stochastic networks, that is for both networks we consider the same realizations of the arrival processes and service requirements (see [6, 14] for more details).

A related research direction is to obtain bounds for the stochastic process of interest. In a recent paper [3] 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 any user).

Our main interest is in stochastic processes that arise in so-called bandwidth-sharing networks introduced in [17] 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 [19] 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

*The work of U. Ayesta was financially supported by The Netherlands Organisation for Scientific Research (NWO, grant B 62-640).

†Also affiliated with Bell Laboratories, Alcatel-Lucent, P.O. Box 636, Murray Hill, NJ 07974, USA

ValueTools 2008, October 21-23, 2008, Athens, GREECE.

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4336
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4336

($\alpha \to 0$), 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 $\alpha = 2$ [9]. In [2] 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 certain probabilistic limiting regimes. For example, in [7, 8, 10] 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.

A simple but important case of bandwidth-sharing networks is the so-called linear network. A linear network is the canonical model to study the interaction between traffic that traverses multiple links and the cross-traffic it meets on its route. For $\alpha = 1$, the joint equilibrium distribution of the number of users in a linear network is known [17]. In [12] approximations are given for the mean number of users under general weighted α-fair policies when one or more of the nodes are in heavy traffic.

In this paper we consider a linear bandwidth-sharing network with general arrival and service processes. The capacity of the various nodes may vary in time. The main goal of this paper is to give sufficient conditions on two 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 [16, 15], which can be explained from the fact that we only compare the two processes starting in the same initial state, as opposed to starting from any two ordered initial states as in [16, 15]. From the performance point of view, starting from the same initial state does not diminish the applicability of the results, since the steady-state behavior does not depend on the initial state. Furthermore, our result is a pure sample-path comparison, and as a consequence it holds for arbitrary arrival processes, service time processes and rate region variations.

We then consider the family of weighted α-fair policies. From our sample-path result we obtain stability results and, under certain restrictions on the service requirements, we prove monotonicity of the weighted mean number of users with respect to the fairness parameter α and the relative weights. To completely investigate all service requirement parameters, we consider a two-node linear network in a heavy-traffic regime and fully characterize the monotonicity results by making use of a conjecture in [7, 8].

The remainder of the paper is organized as follows. In Section 2 the linear bandwidth-sharing network is introduced and in Section 3 the comparison results are stated. Then we focus on the family of weighted α-fair policies in Section 4. We obtain stability results and, under certain restrictions on the service requirements, prove monotonicity of the weighted mean number of users with respect to α and the relative weights. In Section 5 we consider a heavy-traffic regime and obtain monotonicity results for the weighted mean number of users for all service requirement parameters. We also present numerical experiments that provide further insight into the performance of the α-fair policies. The conclusions and future research directions may be found in Section 6.

2. MODEL DESCRIPTION

We consider a linear bandwidth-sharing network with L nodes and $L+1$ classes of users, see Figure 1. 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 $1/\mu_i$, $i = 0, \ldots, L$. Then $\rho_i = \frac{\lambda_i}{\mu_i}$ represents the offered work of class i per time unit. The inter-arrival times and service requirements are mutually independent random variables.

The capacity of node i at time t is equal to $C_i(t)$. Class-i users require service at node i only, $i = 1, \ldots, L$, while class-0 users require service at all nodes simultaneously. When $C_i(t) = C$ for all i and all t, we call it a symmetric linear network. Otherwise the model is called an asymmetric linear network with possibly time-varying capacities.

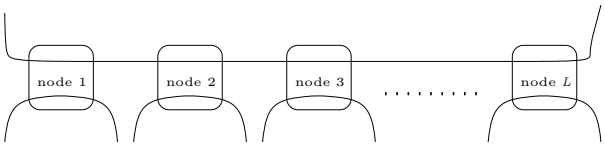

Figure 1: Linear network.

EXAMPLE 2.1 (WIRED AND WIRELESS NETWORKS). *As mentioned in the Introduction, linear bandwidth-sharing networks model the dynamic interaction among competing data flows that traverse several links in a wired network.*

In a wireless network we can think of the following simple example. 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 which we will obtain later in this paper can be applied to a wireless network if coordination between base stations is possible. This has recently been proposed in for example [1, 28].

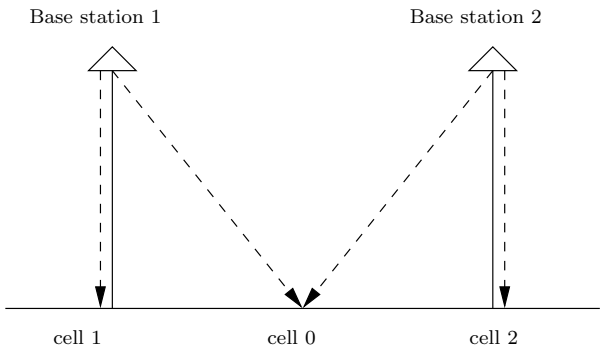

Figure 2: Two base stations.

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4336
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4336

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

$$N^\pi(t) := \sum_{i=0}^{L} N_i^\pi(t)$$

as the total number of users in the system at time 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 further define N_i^π, W_i^π and N^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 $\vec{N} = (N_0, N_1, \ldots, N_L)$. Hence the allocation given to class i can only depend on the time and on the number of users present in the system. The allocation vector $\vec{s}^\pi(t, N) = (s_0^\pi(t, \vec{N}), \ldots, s_L^\pi(t, \vec{N}))$ has to lie in the rate region

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

where $C_i(t)$ may depend on the time t, but not on the state $\vec{N}$ itself. In the remainder of the paper we suppress the dependence on t and write $s_i^\pi(\vec{N})$ instead of $s_i^\pi(t, \vec{N})$. Let

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

be the cumulative amount of service received by class i during the time interval $[0, t]$.

The service discipline within a particular class, the intra-class policy, is the First Come First Served discipline (FCFS).

REMARK 2.2. *It is worth observing that the FCFS assumption is not common in bandwidth-sharing networks, where the intra-class policy is typically assumed to be Processor-Sharing (PS). However, in case the service requirements are exponentially distributed, as is also usually assumed in bandwidth-sharing networks, the stochastic behavior in fact does not depend on the intra-class policy as long as it is non-anticipating. A policy is called non-anticipating when the discipline is not based on any knowledge of the actual realizations of the service requirements. This implies that when the service requirements are exponentially distributed, the results we obtain (by assuming FCFS) will also be valid for non-anticipating policies like PS, the Last Come First Served discipline and the Foreground Background discipline.*

REMARK 2.3. *When the service requirements are exponentially distributed, the arrival processes are Poisson and the capacities do not vary in time ($C_i(t) = C_i$), the process $\{N_0^\pi(t), N_1^\pi(t), \ldots, N_L^\pi(t)\}_{t\geq 0}$ is a continuous-time Markov process. The transition rates are given by:*

$$(N_0, \ldots, N_i, \ldots, N_L) \to (N_0, \ldots, N_i + 1, \ldots, N_L)$$

at rate λ_i and

$$(N_0, \ldots, N_i, \ldots, N_L) \to (N_0, \ldots, N_i - 1, \ldots, N_L)$$

at rate $\mu_i s_i^\pi(\vec{N})\mathbf{1}_{(N_i>0)}$. As indicated in Remark 2.2, the transition rates are independent of the non-anticipating intra-class policy used.

For any policy π we assume that $N_i = 0$ forces $s_i^\pi(N) = 0$, i.e. no capacity is given to a class when there are no users present. Furthermore, $s_i^\pi(N) = C_i(t) - s_0^\pi(N)$ when $N_i > 0$, $i = 1 \ldots, L$, i.e. the remaining capacity in node i is given fully to class i at time t. This implies that no capacity is unnecessarily left unused by a policy. However, this does not ensure that a policy gives 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 $1, \ldots, L$ (note that this policy satisfies the conditions on $\vec{s}^\pi$ as stated in the beginning of this paragraph) 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.

Our goal in this paper is to compare the performance of the network under different policies. First of all, we will be interested in whether a policy can achieve a stable system. Another important performance measure we consider is the (weighted) number of users present in the system. Because of Little's law, a policy that minimizes the total mean number of users present in the system, minimizes the mean overall sojourn time as well.

3. COMPARISON OF POLICIES

In this section we consider the behavior of the 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 $s_0^\pi(\vec{N}^\pi) \leq s_0^{\tilde{\pi}}(\vec{N}^{\tilde{\pi}})$, when $N_0^\pi = N_0^{\tilde{\pi}}$ and $N_i^\pi \geq N_i^{\tilde{\pi}}$, $i = 1, \ldots, L$.*

In particular, note that Property 3.1 is implied by the following property.

PROPERTY 3.1' *Let π and $\tilde{\pi}$ be two 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 $i \neq 0$.*

These properties basically state that higher priority is given to class 0 under policy $\tilde{\pi}$ compared to π. In the remainder of this section we will see that 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 3.5). These results arise from the fact that when class 0 is served, it simultaneously uses capacity in all nodes. Hence, we can prove that giving more preference to class 0 makes better use of the available capacity and 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)). The main challenge in the proof of Proposition 3.2 is that the flip side of giving higher priority to class 0 is that classes 1 and 2 will receive lower preference, and therefore contain more users. Class 0 can then receive less service later on, so that the ultimate impact on the

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4336
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4336

number of class-0 users and the workload in a node is not obvious.

We first establish the 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. If at time $t = 0$, we have $\vec{N}^{\pi}(0) = \vec{N}^{\tilde{\pi}}(0) = \vec{N}(0)$ and at time $t = 0$ the k-th class-i user has the same remaining service requirement under both policies π and $\tilde{\pi}$, for all k and i, then for all $t \geq 0$,*

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

(ii) $S_0^{\pi}(t) + S_i^{\pi}(t) \leq S_0^{\tilde{\pi}}(t) + S_i^{\tilde{\pi}}(t)$,

and hence

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

(iv) $W_0^{\pi}(t) + W_i^{\pi}(t) \geq W_0^{\tilde{\pi}}(t) + W_i^{\tilde{\pi}}(t)$.

PROOF. Denote by $B_{j,n}$ the service requirement of the n-th class-j user, $j = 0, 1, \ldots, L$. When the user was already in the system at time $t = 0$, $B_{j,n}$ denotes the remaining service requirement at time $t = 0$. Denote by $F_j(s) := \sup\{n : \sum_{m=1}^{n} B_{j,m} < s\}$ the number of class-j service completions as function of the amount of service received by class j. Thus $D_j^{\pi}(t) = F_j(S_j^{\pi}(t))$ represents the number of class-j service completions during the time interval $[0, t]$. Denote by $Q_j(t)$ the number of class-j users arriving during the time interval $[0, t]$. Denote by $A_j(0, t) := \sum_{m=1}^{Q_j(t)} B_{j,N_j(0)+m}$ the amount of class-j work arriving during the time interval $[0, t]$.

Thus, the total number of class-j users present at time t can be written as

$$N_j^{\pi}(t) = N_j(0) + Q_j(t) - D_j^{\pi}(t) \tag{1}$$

and the class-j workload can be written as

$$W_j^{\pi}(t) = W_j(0) + A_j(0, t) - S_j^{\pi}(t).$$

Note that $N_j(0)$, $W_j(0)$ and the functions $F_j(\cdot), Q_j(\cdot)$ and $A_j(0, t)$ are independent of the policy, and the function $F_j(\cdot)$ is non-decreasing. Hence, inequality (i) implies (iii) and inequality (ii) implies inequality (iv). It suffices to prove that inequalities (i) and (ii) hold.

We will 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, that is $S_0^{\pi}(t) = S_0^{\tilde{\pi}}(t)$ and $s_0^{\pi}(\vec{N}^{\pi}(t)) > s_0^{\tilde{\pi}}(\vec{N}^{\tilde{\pi}}(t))$ (with strict inequality), but $S_0^{\pi}(t) + S_i^{\pi}(t) \leq S_0^{\tilde{\pi}}(t) + S_i^{\tilde{\pi}}(t)$ for all $i = 1, \ldots, L$. Hence from (1) we obtain $N_0^{\pi}(t) = N_0^{\tilde{\pi}}(t)$ and $N_i^{\pi}(t) \geq N_i^{\tilde{\pi}}(t)$ for all $i = 1, \ldots, L$. Together with Property 3.1 this contradicts the initial assumption.

Next, assume that inequality (ii) is the first one to be violated, i.e., $S_0^{\pi}(t) + S_i^{\pi}(t) = S_0^{\tilde{\pi}}(t) + S_i^{\tilde{\pi}}(t)$ 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 $S_0^{\pi}(t) \leq S_0^{\tilde{\pi}}(t)$ and $S_0^{\pi}(t) + S_j^{\pi}(t) \leq S_0^{\tilde{\pi}}(t) + S_j^{\tilde{\pi}}(t)$ for all $j \neq 0, i$. By (1) we have $N_0^{\pi}(t) \geq N_0^{\tilde{\pi}}(t)$ and $N_i^{\pi}(t) \leq N_i^{\tilde{\pi}}(t)$.

First assume $N_i^{\tilde{\pi}}(t) > 0$. Since we made the assumption that no capacity is unnecessarily left unused ($s_i(\vec{N}) = C_i(t) - s_0(\vec{N})$ when $N_i > 0$), it then follows that $s_0^{\pi}(\vec{N}^{\pi}(t)) + s_i^{\pi}(\vec{N}^{\pi}(t)) \leq C_i(t) = s_0^{\tilde{\pi}}(\vec{N}^{\tilde{\pi}}(t)) + s_i^{\tilde{\pi}}(\vec{N}^{\tilde{\pi}}(t))$ which contradicts the initial assumption.

If $N_i^{\tilde{\pi}}(t) = 0$, then $N_i^{\pi}(t) = 0$ as well and $S_i^{\pi}(t) = S_i^{\tilde{\pi}}(t)$. Hence $S_0^{\pi}(t) = S_0^{\tilde{\pi}}(t)$ and $S_j^{\pi}(t) \leq S_j^{\tilde{\pi}}(t)$ for all j. By (1) we obtain $N_0^{\pi}(t) = N_0^{\tilde{\pi}}(t)$ and $N_j^{\pi}(t) \geq N_j^{\tilde{\pi}}(t)$ for all j. 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 in fact not require any distributional or independence assumptions with respect to the inter-arrival times and service requirements.*

Assume for the moment Poisson arrivals, exponentially distributed service requirements and that the capacities do not vary in time, i.e. $C_i(t) = C_i$. The process $\{N_0(t), \ldots, N_L(t)\}_{t \geq 0}$ is a continuous-time Markov process, see Remark 2.3. Hence Proposition 3.2 (iii) then states in fact a sample-path wise pre-ordering on two continuous-time Markov processes $\{\vec{N}^{\pi}(t)\}_{t \geq 0}$ and $\{\vec{N}^{\tilde{\pi}}(t)\}_{t \geq 0}$ starting from the same initial state.

There is a broad range of literature on the existence of orderings of stochastic processes. An important ordering is the stochastic ordering $\leq_{st}$ ([20, 24]). Note that 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. $X(t) \stackrel{d}{=} X'(t)$ and $Y(t) \stackrel{d}{=} Y'(t)$, 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 [16, Theorem 5.3] and [15, 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, the necessary and sufficient conditions on the policies π and $\tilde{\pi}$ to obtain

$$\{N_0^{\pi}(t)\}_{t \geq 0} \geq_{st} \{N_0^{\tilde{\pi}}(t)\}_{t \geq 0}, \tag{2}$$

for any two ordered initial states $N_0^{\pi}(0) \geq N_0^{\tilde{\pi}}(0)$, are

$$s_0^{\pi}(\vec{N}^{\pi}) \leq s_0^{\tilde{\pi}}(\vec{N}^{\tilde{\pi}}) \text{ when } N_0^{\pi} = N_0^{\tilde{\pi}}. \tag{3}$$

It can immediately be seen that Property 3.1 is a weaker condition than (3). Interestingly, we will show that for α-fair bandwidth-sharing policies, Property 3.1 is satisfied, but (3) is not satisfied. Note that the stochastic ordering result in (2) holds for any two initial states that are ordered, whereas in Proposition 3.2 the initial state is the same. This explains the fact that Property 3.1 can be weaker than (3). Since we are interested in performance metrics in steady state (stability and mean number of users), the assumption that the processes have the same initial state is not restrictive. In addition, Proposition 3.2 is not restricted to Markov processes, hence it applies as well for generally distributed arrival processes, service requirements and time-varying rate regions.

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4336
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4336

Property 3.1 and Proposition 3.2 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.

COROLLARY 3.4. *Let π be a policy in a linear network with L nodes that satisfies the following property:*

$$s_0^\pi(N_0, N_1, \ldots, N_L) \leq s_0^\pi(N_0, M_1, \ldots, M_{L-1}, 0)$$

for all $N_i \geq M_i, i = 1, \ldots, L-1$.

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: $s_0^\pi(N_0, \ldots, N_{L-1}) := s_0^\pi(N_0, \ldots, N_{L-1}, 0)$. Then

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

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

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

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 $x \geq 0$, $s_0^{\tilde{\pi}}(N_0, N_1, \ldots, N_{L-1}, x) := s_0^\pi(N_0, N_1, \ldots, N_{L-1})$. Hence

$$\begin{aligned} s_0^\pi(N_0, \ldots, N_L) &\leq s_0^\pi(N_0, M_1, \ldots, M_{L-1}, 0) \\ &= s_0^\pi(N_0, M_1, \ldots, M_{L-1}) \\ &= s_0^{\tilde{\pi}}(N_0, M_1, \ldots, M_{L-1}, x) \end{aligned}$$

for all x and all $N_i \geq M_i$, $i = 1, \ldots, L-1$. This implies that policies π and $\tilde{\pi}$ satisfy Property 3.1 and from Proposition 3.2 the result follows. □

In the next two sections, Proposition 3.2 will be used to readily derive results for the stability and the weighted mean number of users present in a linear bandwidth-sharing network.

3.1 Stability

Recall that the stability conditions depend on the policy being used. We note that 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 3.5. *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 and $\sum_{i=1}^L c_i\mu_i \leq c_0\mu_0$, the sample-path comparison established in Proposition 3.2 will allow us to show that giving more priority to class 0 decreases the weighted mean number of users.

PROPOSITION 3.6. *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*

$$\sum_{i=0}^L c_i \mathbb{E}(N_i^\pi(t)) \geq \sum_{i=0}^L c_i \mathbb{E}(N_i^{\tilde{\pi}}(t)), \quad \forall\, t \geq 0. \tag{4}$$

PROOF. 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 expectation 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

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

□

Note that by Remark 2.2, Proposition 3.6 holds for any non-anticipating intra-class policy, so not only for FCFS.

REMARK 3.7. *One natural choice for the weights c_i in (4) 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 3.6 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 3.8. *We only obtain a comparison result in terms of the* mean *number of users, 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.*

Assume $\vec{N}^\pi(t)$ and $\vec{N}^{\tilde{\pi}}(t)$ are two Markov processes. The necessary and sufficient conditions in order to obtain $N^\pi(t) \geq_{st} N^{\tilde{\pi}}(t)$ for any ordered initial states $N^\pi(0) \geq N^{\tilde{\pi}}(0)$, 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

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4336
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4336

with $N^{\pi} = N^{\tilde{\pi}}$, [15, 16]. (Recall that we denote by $N = \sum_{i=0}^{L} N_i$ the total number of users.) In a queueing context this condition is rather strong. For example, for the linear network we consider in this paper, we need for states with $\vec{N}^{\pi} = (0, 1, \ldots, 1)$ and $\vec{N}^{\tilde{\pi}} = (L, 0, \ldots, 0)$ that $\sum_{i=1}^{L} \mu_i \leq \mu_0$, but for states with $\vec{N}^{\pi} = (1, 0, \ldots, 0)$ and $\vec{N}^{\tilde{\pi}} = (0, \ldots, 0, 1, 0, \ldots, 0)$ we need $\mu_0 \leq \mu_i$, $i = 1, \ldots, L$. 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 on the total number of users as in the framework of [15, 16] does not hold.

A natural objective in queueing networks is to minimize the total/weighted number of users in the system. Classical results for a single-server system indicate that giving preference to "small" users is beneficial in terms of the total/weighted number of users present in the system [23, 25, 21, 22]. For exponentially distributed service requirements, the $c\mu$-rule, i.e. giving priority to the class with the maximum instantaneous weighted departure rate $c_i\mu_i$, minimizes the weighted mean number of users among all non-anticipating policies [21]. 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 [27].

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 weighted mean number of users present in the system. Using Proposition 3.6 it can be proved that this is indeed the case.

COROLLARY 3.9. *Consider an asymmetric 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 weighted mean number of users $\sum_{i=0}^{L} c_i\mathbb{E}(N_i(t))$, 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 3.1' is satisfied and from Proposition 3.6 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 [27] it was proved that for a symmetric linear network, policy π^*, as defined in Corollary 3.9, 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 $N^{\pi}(t) \geq_{st} N^{\pi^*}(t)$ given that $\vec{N}^{\pi}(0) = \vec{N}^{\pi^*}(0)$.

4. WEIGHTED α-FAIR POLICIES

Weighted α-fair policies are an important family of policies that have received a lot of attention in recent years. The weighted-α fair allocation is the solution to the following optimization problem:

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

Denote the weighted α-fair discipline with weights $w = (w_0, w_1, \ldots, w_L)$ and parameter α by $\pi^{\alpha,w}$ and the corresponding allocation vector by $\vec{s}^{(\alpha,w)}(\vec{N})$. 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. Throughout this section we assume exponentially distributed service requirement. Thus, the results we obtain will also be valid if the intra-class policy is PS, see Remark 2.2.

In order to compare two α-fair policies we only need to check whether Property 3.1' holds. In [2] it was shown that for a symmetric linear network with unit capacity for all nodes the weighted α-fair allocation is given by

$$s_0^{(\alpha,w)}(\vec{N}) = \frac{(w_0 N_0^{\alpha})^{1/\alpha}}{(w_0 N_0^{\alpha})^{1/\alpha} + (\sum_{i=1}^{L} w_i N_i^{\alpha})^{1/\alpha}} \tag{8}$$

and $s_i^{(\alpha,w)}(\vec{N}) = 1 - s_0^{(\alpha,w)}(\vec{N})$ for all i with $N_i > 0$. Using (8), it can be checked that Property 3.1' is satisfied for a symmetric linear network when comparing policies $\pi^{\beta,w}$ and $\pi^{\gamma,\tilde{w}}$ with $\beta \leq \gamma$ and $\frac{w_0}{w_i} \leq \frac{\tilde{w}_0}{\tilde{w}_i}$, $i = 1, \ldots, L$ (see also [12, 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 3.1' is satisfied then as well. The proof may be found in the technical report [26].

LEMMA 4.1. *The following results hold in an asymmetric linear network:*

(i) $s_0^{(\alpha,w)}(\vec{N})$ is non-increasing in N_i, $i = 1, ..., L$.

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

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

Since Property 3.1' holds for weighted α-fair policies, our framework allows us to gain insights into the performance of such policies in linear networks, see Sections 4.1 and 4.2 below.

Note that the stochastic comparison results in [15, Theorem 2] and [16, Theorem 5.3] are not applicable here. As we already saw in Remark 3.8, there is no stochastic ordering possible for any ordered initial states on the total number of users present in the system. Also, an ordering on the number of class-0 users for any ordered initial states is not possible, since equation (3) is not satisfied for the class of weighted α-fair policies in linear networks. Consider for example a 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

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4336
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4336

α-fair policies. From (8) we see that if M tends to ∞ then $s_0^{\vec{\pi}}(\vec{N}^{\vec{\pi}})$ tends to 0. Hence (3) cannot hold for any pair of α-fair policies.

In [3] 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 any user). This property fails to hold for a linear network under α-fair policies. For example, removing a class-1 user implies that class 1 gets less capacity and class 0 gets more. This however implies that classes $i = 2, \ldots, L$ obtain less capacity as well and hence a class-i user gets less capacity, $i = 2, \ldots, L$. The only requirement in Property 3.1' is that removing a class-i user, $i \neq 0$, increases the capacity allocated to the class-0 users. As shown in Lemma 4.1, this holds under natural conditions on the parameters of weighted α-fair policies.

REMARK 4.2. *Note that from Lemma 4.1 and Corollary 3.4 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.*

4.1 Stability

In [2] 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 $\rho_0 + \rho_i < C_i$, $i = 1, \ldots, L$. Corollary 3.5 and Lemma 4.1 allow us to derive stability results for a linear network *with* time-varying capacities.

COROLLARY 4.3. *Assume $\beta \leq \gamma$ and $\frac{w_0}{w_i} \leq \frac{\bar{w}_0}{\bar{w}_i}$, $i = 1, \ldots, L$. Let the service requirements be exponentially distributed. If policy $\pi^{\beta,w}$ gives a stable system, then policy $\pi^{\gamma,\bar{w}}$ gives a stable system as well.*

PROOF. Note that α-fair policies have PS as an 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 take FCFS as intra-class policy. From Lemma 4.1 we obtain that Property 3.1 is satisfied, hence the result in Corollary 3.5 applies. □

In [13] 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. In addition, in every state the rate region is convex. The authors characterize the stability conditions and show that the stability region is non-increasing in the value of α. Interestingly, Corollary 4.3 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 4.4. *Assume Poisson arrivals and exponentially distributed service requirements. Consider an asymmetric 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 $C_i(t)$.*

Then the policy $\pi^{\alpha,w}$ with $w_i \leq w_0, i = 1, \ldots, L$ gives a stable system whenever possible. The stability conditions are given by $\rho_0 + \rho_i < \overline{C}_i$, $i = 1, \ldots, L$.

PROOF. In [13] 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 4.3 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. From Corollary 4.3 we can then conclude that the same holds for policy $\pi^{\alpha,w}$ with $w_i \leq w_0, i = 1, \ldots, L$. □

4.2 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 asymmetric 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 4.1. 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 weighted mean number of users indeed decreases when more preference is given to class 0. More precisely, the weighted mean number of users is non-increasing in α and in $\frac{w_0}{w_i}$, $i = 1, \ldots, L$.

COROLLARY 4.5. *Assume exponentially distributed service requirements with $\sum_{i=1}^{L} c_i\mu_i \leq c_0\mu_0$. If $\beta \leq \gamma$ and $\frac{w_0}{w_i} \leq \frac{\bar{w}_0}{\bar{w}_i}$, $i = 1, \ldots, L$, then*

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

PROOF. From Lemma 4.1 we obtain that $\pi^{\beta,w}$ and $\pi^{\gamma,\bar{w}}$ satisfy Property 3.1'. The result then follows from Proposition 3.6. □

When $\sum_{i=1}^{L} c_i\mu_i > c_0\mu_0$, a trade-off arises between the above-described effects. In the next section we will investigate this further.

5. MONOTONICITY RESULTS FOR α-FAIR POLICIES

In the previous section, monotonicity results of the weighted mean number of users were derived for the family of α-fair policies and exponentially distributed service requirements with $\sum_{i=1}^{L} c_i\mu_i \leq c_0\mu_0$. In this section we will explore the case $\sum_{i=1}^{L} c_i\mu_i > c_0\mu_0$ for a two-node linear network ($L = 2$).

When $c_1\mu_1 + c_2\mu_2 > c_0\mu_0$, 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 4.1 we see that this can be done by choosing α small. However, at the same time this uses the available capacity in each node less efficiently, as proved in Proposition 3.2 (iv). Thus a trade-off arises between the two effects, which makes the analysis difficult. In

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4336
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4336

Section 5.1 we will consider a heavy-traffic regime and establish (over the whole range of μ_0) monotonicity results in α for the weighted mean scaled number of users. In Section 5.2 we perform numerical experiments for a normally loaded system and observe in particular that when $\mu_1 + \mu_2 > \mu_0$, the total mean number of users is not necessarily monotone in α when $\alpha < 1$.

5.1 Heavy-traffic regime

In this section we study the monotonicity 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 [7, 8, 10], 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, see [7, 8] for full details.

Assume a heavy-traffic setting $\rho_i + \rho_0 = C_i$ for $i = 1, 2$. Define the diffusion scaled processes as follows:

$$\begin{aligned}\hat{n}_i^{k,(\alpha)}(t) &:= \frac{N_i^{\pi^{\alpha,\vec{1}}}(kt)}{\sqrt{k}},\ i = 0, 1, 2, \quad \text{and} \\ \hat{v}_i^{k,(\alpha)}(t) &:= \frac{N_0^{\pi^{\alpha,\vec{1}}}(kt)/\mu_0 + N_i^{\pi^{\alpha,\vec{1}}}(kt)/\mu_i}{\sqrt{k}} \\ &= \hat{n}_0^{k,(\alpha)}(t)/\mu_0 + \hat{n}_i^{k,(\alpha)}(t)/\mu_i,\ i = 1, 2.\end{aligned}$$

Here $\hat{v}_i^{k,(\alpha)}(t)$ can be seen as the total workload in node i under the diffusion scaling. In [8, Conjecture 5.1] it is conjectured that for an arbitrary bandwidth-sharing network, the diffusion scaled workload process $\hat{\vec{v}}^{k,(\alpha)}(t)$ converges in distribution as $k \to \infty$ to $\vec{v}^{(\alpha)}(t)$, where $\vec{v}^{(\alpha)}(t)$ is a semi-martingale 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 [7, 8] 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 $\alpha \neq 1$. The workload cone for a two-node linear network under an α-fair policy with unit weights is given by

$$\begin{aligned}&\{\vec{v} : v_i = \frac{\rho_0}{\mu_0}(q_1 + q_2)^{\frac{1}{\alpha}} + \frac{\rho_i}{\mu_i} q_i^{\frac{1}{\alpha}},\ q_1, q_2 \geq 0,\ i = 1, 2\} \\ &= \{\vec{v} : v_1 \geq 0, v_1 \frac{\rho_0/\mu_0}{(C_1 - \rho_0)/\mu_1 + \rho_0/\mu_0} \leq v_2, \\ &\qquad v_2 \leq v_1 \frac{(C_2 - \rho_0)/\mu_2 + \rho_0/\mu_0}{\rho_0/\mu_0}\},\end{aligned}$$

which is independent of the parameter α. Hence, the process $\vec{v}^{(\alpha)}(t)$ is independent of α as well. The diffusion scaled number of users, $\hat{\vec{n}}^{k,(\alpha)}(t)$, converges in distribution as $k \to \infty$ to some process $\vec{n}^{(\alpha)}(t)$ which does depend on α (this process is specified in [7]).

Since the process of the total workload in a node does not depend on α, we can derive monotonicity results for the weighted mean number of users present in the system over the whole range of the parameter μ_0. We can express the weighted number of users in the system as follows:

$$\begin{aligned}\sum_{i=0}^{L} c_i \hat{n}_i^{(\alpha)}(t) &= \frac{c_0\mu_0 - \sum_{i=1}^{2} c_i\mu_i}{\mu_0} \hat{n}_0^{(\alpha)}(t) \\ &\quad + \sum_{i=1}^{2} c_i\mu_i \cdot (\frac{1}{\mu_0}\hat{n}_0^{(\alpha)}(t) + \frac{1}{\mu_i}\hat{n}_i^{(\alpha)}(t)) \\ &\stackrel{d}{=} \frac{c_0\mu_0 - \sum_{i=1}^{2} c_i\mu_i}{\mu_0} \hat{n}_0^{(\alpha)}(t) + \sum_{i=1}^{2} c_i\mu_i \hat{v}_i^{(\alpha)}(t).\end{aligned} \tag{9}$$

From Proposition 3.2 we know that $N_0^{\pi^\alpha}(t)$ is decreasing in α, and hence $\hat{n}_0^{(\alpha)}(t)$ is decreasing in α as well. Together with the fact that $\hat{v}_i^{(\alpha)}(t)$ is independent of α and by taking expectations in (9), we have 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 $\mathbb{E}(\sum_{i=0}^{2} c_i\hat{n}_i^{(\alpha)}(t))$ is non-increasing or non-decreasing in α respectively. In fact, when in addition we use the characterization of $\vec{\hat{n}}^{(\alpha)}(t)$, we are able to derive a stronger monotonicity result. The proof may be found in the technical report [26].

PROPOSITION 5.1. *Assume $\rho_i + \rho_0 = C_i$, $i = 1, 2$.*

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

5.2 Numerical results

In this section we present numerical experiments to provide further insight into the performance of α-fair policies. We simulate 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 $w_j = c_j = 1$, $j = 0, 1, 2$.

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 4.5, 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 4.1 (ii). These two opposite effects might cause that the total mean number of users is not monotone in α. Note that for the heavy-traffic regime as considered in Section 5.1, the workload in a node is independent of the parameter α. Hence, there was no trade-off and we were able to prove the monotonicity result 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, with exception of a few

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4336
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4336

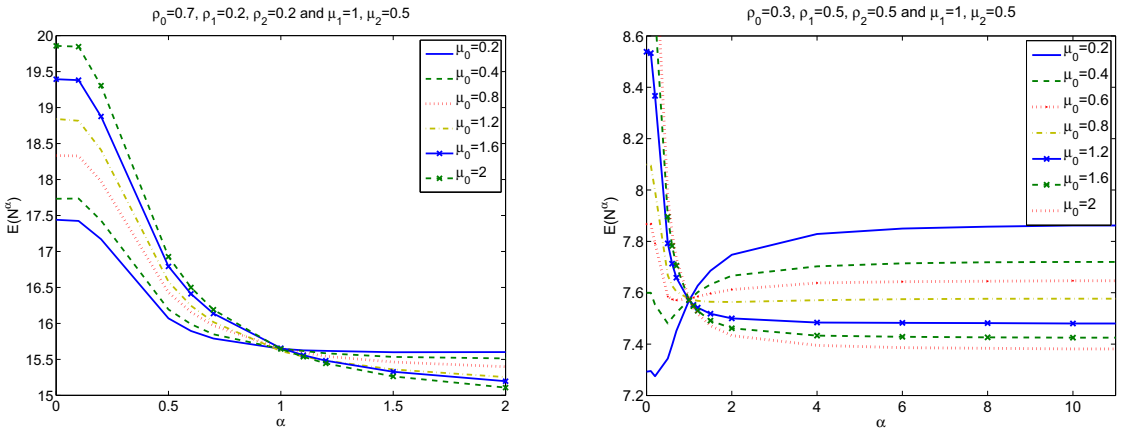

Figure 3: Total mean number of users under α-fair policies in a two-node linear network with a) $\rho_0 = 0.7, \rho_1 = 0.2$ and $\rho_2 = 0.2$, and b) $\rho_0 = 0.3, \rho_1 = 0.5$ and $\rho_2 = 0.5$.

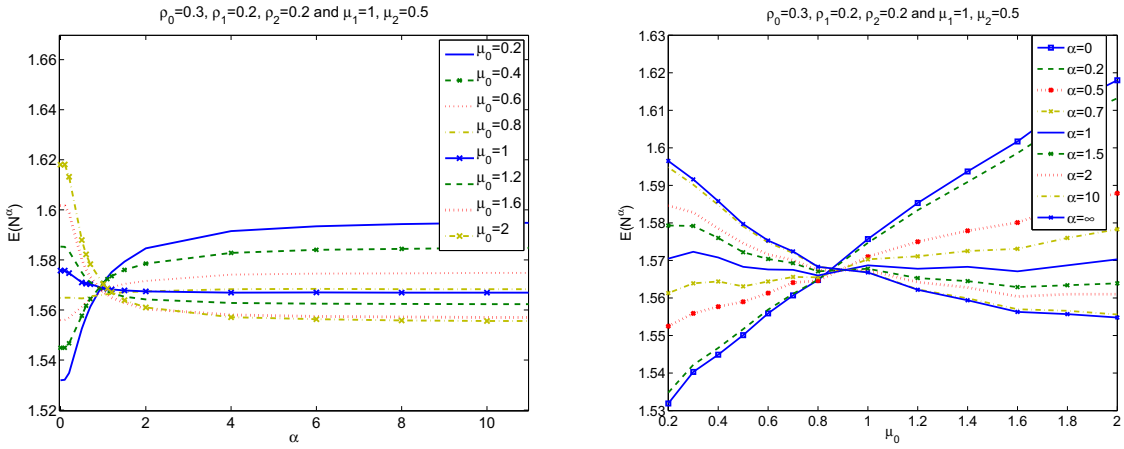

Figure 4: Total mean number of users under α-fair policies in a two-node linear network with $\rho_0 = 0.3, \rho_1 = 0.2$ and $\rho_2 = 0.2$.

points, the total mean number of users is increasing in μ_0 when $\alpha < 1$ and decreasing in μ_0 when $\alpha > 1$, respectively. Intuitively, the monotonicity can be explained as follows. Note that if $\alpha = 1$, the policy reduces to PF. For PF with unit weights, the mean total number of users is exactly known ([17]) and equals

$$\mathbb{E}(N^{\pi^{1,\vec{1}}}) = \frac{\rho_1}{1-\rho_0-\rho_1} + \frac{\rho_2}{1-\rho_0-\rho_2} + \frac{\rho_0}{1-\rho_0}\Big(1 + \frac{\rho_1}{1-\rho_0-\rho_1} + \frac{\rho_2}{1-\rho_0-\rho_2}\Big).$$

This is independent of the parameters μ_0, μ_1 and μ_2 for given values of ρ_0, ρ_1 and ρ_2. When $\alpha > 1$, we observe from Lemma 4.1 (ii) 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.

6. CONCLUSION AND FUTURE WORK

In this paper we studied linear bandwidth-sharing networks and obtained comparison results for the performance under two different policies in terms of both stability and the weighted mean number of users. 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 in certain classes.

We proved monotonicity results for the weighted mean number of users under α-fair policies when $c_0\mu_0 \geq \sum_{i=1}^{L} c_i\mu_i$. In the numerical section, we demonstrated additional monotonicity properties. For instance we have strong evidence to believe that the total mean number of users in the system is

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4336
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4336

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, see Figure 4 b). Similar monotonicity results have been discussed for a single-server queue in [3, 11], but to the best of our knowledge there does not exist any proof. There is no hope that this monotonicity 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.

In future work, it might be interesting to consider different types of networks, like a star or a grid network. Also a multi-class single-server queue is worth studying. In the case of two classes, the single-server system is a linear network with one node ($L = 1$). For this case, the results developed in this paper can be used to derive that the weighted mean number of users is monotone with respect to the ratio of the weights for both Discriminatory Processor Sharing (DPS) and Generalized Processor Sharing (GPS). In [26] we extend our sample-path approach to a single-server system with more than two classes.

Acknowledgment

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

7. REFERENCES

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

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

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

[4] S.C. Borst, M. Jonckheere, and L. Leskelä. Stability of parallel queueing systems with coupled service rates. *Discrete Event Dynamic Systems*, 2008. To appear.

[5] 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.

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

[7] 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 2004*, pages 3938–3943, 2004.

[8] 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. *Preprint*, 2007.

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

[10] 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.

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

[12] 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 2006*, Pisa, Italy, 2006.

[13] 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 2007*, pages 239–250, San Diego, CA, USA, 2007.

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

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

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

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

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

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

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

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

[22] 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*, 3:323–334, 1989.

[23] 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.

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

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

[26] I.M. Verloop, U. Ayesta, and S.C. Borst. Monotonicity properties for multi-class stochastic systems. *CWI research report*, 2008.

[27] I.M. Verloop, S.C. Borst, and R. Núñez-Queija. Delay optimization in bandwidth-sharing networks. In *Proceedings of CISS 2006*, Princeton University, 2006.

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

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4336
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4336