Centrum voor Wiskunde en Informatica

REPORTRAPPORT

PNA

Probability, Networks and Algorithms

Probability, Networks and Algorithms

Heavy-traffic delay minimization in bandwidth-sharing networks

I.M. Verloop, S.C. Borst

REPORT PNA-E0609 AUGUST 2006

P.O. Box 94079, 1090 GB Amsterdam (NL)
Kruislaan 413, 1098 SJ Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

ISSN 1386-3711

Heavy-traffic delay minimization in bandwidth-sharing networks

ABSTRACT

Bandwidth-sharing networks as considered by Massoulie & Roberts provide a natural modeling framework for describing the dynamic flow-level interaction among elastic data transfers. Although valuable stability results have been obtained, crucial performance metrics such as flow-level delays and throughputs in these models have remained intractable in all but a few special cases. In particular, it is not well understood to what extent flow-level delays and throughputs achieved by standard bandwidth-sharing mechanisms such as alpha-fair strategies leave potential room for improvement. In order to gain a better understanding of the latter issue, we set out to determine the scheduling policies that minimize the mean delay in some simple linear bandwidth-sharing networks. While admittedly simple, linear networks provide a useful model for flows that traverse several links and experience bandwidth contention from independent cross-traffic. Even for linear topologies it is rarely possible however to explicitly identify optimal policies except in a few limited cases with exponentially distributed flow sizes. Rather than aiming for strictly optimal policies, we therefore focus on a class of relatively simple priority-type strategies that only separate large flows from small ones. To benchmark the performance of these strategies, we compare them with Proportional Fair as the prototypical alpha-fair policy, and establish that the mean delay may be reduced by an arbitrarily large factor when the load is sufficiently high. In addition, we show the above strategies to be asymptotically optimal for flow size distributions with bounded support. Numerical experiments reveal that even at fairly moderate load values the performance gains can be significant.

2000 Mathematics Subject Classification: 68M20, 60K25, 90B18

Keywords and Phrases: bandwidth-sharing networks; delay optimization; alpha-fair strategies; heavy traffic; size-based scheduling

Note: The work of the first author was carried out in the framework of the project "Efficient flow scheduling in resource-sharing networks with variable service rates" funded by The Netherlands Organisation for Scientific Research (NWO).

Heavy-Traffic Delay Minimization in Bandwidth-Sharing Networks

Maaike Verloop[1], Sem Borst[1,2,3]

[1]CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

[2]Department of Mathematics & Computer Science, Eindhoven University of Technology
P.O. Box 513, 5600 MB Eindhoven, The Netherlands

[3]Bell Laboratories, Lucent Technologies, P.O. Box 636, Murray Hill, NJ 07974, USA

Abstract

Bandwidth-sharing networks as considered by Massoulié & Roberts provide a natural modeling framework for describing the dynamic flow-level interaction among elastic data transfers. Although valuable stability results have been obtained, crucial performance metrics such as flow-level delays and throughputs in these models have remained intractable in all but a few special cases. In particular, it is not well understood to what extent flow-level delays and throughputs achieved by standard bandwidth-sharing mechanisms such as α-fair strategies leave potential room for improvement.

In order to gain a better understanding of the latter issue, we set out to determine the scheduling policies that minimize the mean delay in some simple linear bandwidth-sharing networks. While admittedly simple, linear networks provide a useful model for flows that traverse several links and experience bandwidth contention from independent cross-traffic. Even for linear topologies it is rarely possible however to explicitly identify optimal policies except in a few limited cases with exponentially distributed flow sizes. Rather than aiming for strictly optimal policies, we therefore focus on a class of relatively simple priority-type strategies that only separate large flows from small ones. To benchmark the performance of these strategies, we compare them with Proportional Fair as the prototypical α-fair policy, and establish that the mean delay may be reduced by an arbitrarily large factor when the load is sufficiently high. In addition, we show the above strategies to be asymptotically optimal for flow size distributions with bounded support. Numerical experiments reveal that even at fairly moderate load values the performance gains can be significant.

1 Introduction

Over the past several years, the processor-sharing discipline has emerged as a useful paradigm for evaluating the flow-level performance of elastic data transfers competing for bandwidth on a single bottle-neck link, see for instance [1, 12]. Bandwidth-sharing networks as considered by Massoulié & Roberts [10] provide a natural extension for modeling the dynamic interaction among competing elastic flows that traverse several links along their source-destination paths. Bonald & Massoulié [2] showed that a wide class of α-fair bandwidth-sharing policies as introduced by Mo & Walrand [11] achieve stability in such networks under the simple (and necessary) condition that no individual link is overloaded, see also [18] for instance. While stability is arguably the most fundamental performance criterion, flow-level delays and throughputs are obviously crucial metrics too. Although useful approximations, bounds [3] and heavy-traffic limits [8] have been obtained, the latter performance metrics have largely remained intractable in all but a few special cases. In particular, it is not well understood to what extent the flow-level delays and throughputs achieved by common bandwidth-sharing mechanisms leave potential room for improvement.

The scope for improving flow-level delays and throughputs has been the focus of intense efforts in a somewhat distinct strand of research on size-based scheduling strategies. Several studies have demonstrated that the Shortest Remaining Processing Time first (SRPT) discipline can achieve significant performance improvements for heavy-tailed service requirements compared to First-Come First-Served or Processor Sharing. The SRPT discipline has therefore been adopted as an effective mechanism for improving the performance of web servers [4, 6]. A critical issue associated with size-based scheduling in general and SRPT in particular, is that it relies on (partial) knowledge of (remaining) service requirements. While such information is usually available in web servers, it is impractical to obtain in Internet routers. An alternative strategy which has hence been advocated for scheduling data flows is the Least Attained Service first (LAS) discipline also known as Foreground-Background Processor Sharing [9, 13, 14, 15].

Nearly all studies on the performance gains from size-based scheduling strategies such as SRPT and LAS have considered single-server settings. Single-server systems provide reasonable models for web servers, but they do not accurately capture scenarios where users require service from several resources simultaneously. Such concurrent resource possession arises in the above-mentioned bandwidth-sharing networks, where data flows traverse several links between their source-destination pairs and consume bandwidth on each of them for the duration of the transfer. (Even though individual packets travel across the network on a hop-by-hop basis, when we view the system behavior on a somewhat longer time scale, a data flow claims roughly equal bandwidth on each of the links along its path since the amount of buffering at intermediate nodes is typically quite limited.)

While single-server systems provide tractable results and useful insights, they do not exhibit the potential non-work-conserving behavior that may occur in scenarios with concurrent resource possession. There are various indications that priority mechanisms in such scenarios may cause starvation effects with possibly severe consequences. For example, Yang & De Veciana [21, 22] demonstrated that SRPT scheduling in network scenarios may yield considerable performance improvements in terms of mean delays and throughputs, but also observed that flows on long routes with large sizes may sustain a marked performance degradation. Recently, it was shown that size-based scheduling strategies such as SRPT and LAS may in fact unnecessarily fail to achieve stability in network settings, even at arbitrarily low loads [20].

In conclusion, the results for size-based scheduling in single-server models do not provide a good indication for the scope for improvement over common bandwidth-sharing mechanisms in network scenarios. In order to gain better insight into the latter issue, we will set out to determine scheduling policies that minimize the mean delay in bandwidth-sharing networks with a linear topology. While admittedly simple, linear networks provide a useful model for flows that traverse several links and experience bandwidth contention from independent cross-traffic. Even for linear topologies, however, it is barely feasible to explicitly obtain optimal policies, except in a few restrictive cases with exponentially distributed flow sizes [19].

In case of general flow size distributions, optimal policies may be exceedingly complicated or even totally intractable. Rather than seeking strictly optimal policies, we will therefore focus on a class of relatively simple priority-type strategies that only distinguish small and large flows on each of the routes. We will examine the performance of these strategies in heavy-traffic conditions where each of the links is near-critically loaded. Although the link utilization may not always be that high, a heavy-traffic regime is relevant to consider because at low load the performance will tend to be satisfactory no matter what. Also, even when the typical link utilization is relatively low, the load might fluctuate over time and exhibit significant surges, causing severe congestion periods or even temporary overload conditions. In particular, we compare the performance of the above strategies with that of Proportional Fair as the prototypical α-fair policy, and demonstrate that the reduction in the mean flow delay and thus the improvement in user throughput becomes arbitrarily large as the load approaches the critical value. For flow size distributions with bounded support, we show the above strategies in addition to be asymptotically optimal. Numerical experiments indicate that even at reasonably moderate load values the performance gains can be substantial.

As a final comment, it is worth emphasizing that there is a fundamental trade-off between achievable performance and implementation complexity. Although promising methods for obtaining flow size estimates and supporting flow-aware scheduling have been proposed, the actual implementation of size-based scheduling strategies in high-speed routers arguably involves major challenges. In the present paper we do not aim to pursue implementation issues in any depth, but rather focus on deliberately simple strategies in an effort to evaluate the scope for performance gains. Gaining quantitative insight into the achievable improvements in an ideal situation is meant to serve as first step towards determining whether the potential benefit is sufficient to even bother contemplating implementation aspects.

The remainder of the paper is organized as follows. In Section 2 we provide a detailed model description and introduce notation. We gather some useful preliminaries in Section 3. In Section 4 we develop a heavy-traffic analysis of a single-node system in order to illuminate the key observations and mathematical constructs in the simplest possible context. In Section 5 we then turn the attention to linear bandwidth-sharing networks as described above. Subsection 5.1 deals with the case where all the flows on the long route are granted priority over the large flows on the short routes. In Subsection 5.2 we address the case where the flows on the short routes, when simultaneously present, are favored over the large flows on the long route. In Section 6 we present the numerical experiments that we conducted to validate the analytical findings and in particular compare the performance of the above strategies with that of a Proportional Fair policy.

2 Model description and notation

We consider a linear network with L nodes, see Figure 1. For convenience, we assume each of the nodes to have a unit service rate. In order to present the results in the simplest possible setting, we focus on a traffic scenario with $L+1$ classes, where class i requires service at node i only, $i = 1, \ldots, L$, while class 0 requires service at all L nodes simultaneously. Class-i users arrive according to a Poisson process of rate λ_i, and have generally distributed service requirements B_i with distribution function $B_i(x) = \mathbb{P}(B_i < x)$, $i = 0, \ldots, L$. Define $M_i := \inf\{m : \mathbb{P}(B_i > m) = 0\}$ as the maximum possible value of B_i, with $M_i = \infty$ in case B_i has infinite support. Throughout, we assume $\mathbb{E}(B_i^2) < \infty$. Denote by $p_i := \lambda_i/\lambda$ the fraction of class-i users, with $\lambda = \sum_{i=0}^{L} \lambda_i$ the total arrival rate. Let the traffic load of class i be $\rho_i := \lambda_i \mathbb{E}(B_i)$, thus the load at node i is $\rho_0 + \rho_i$.

Denote by Π the class of all (possibly preemptive) policies. For a given policy $\pi \in \Pi$, denote by $N_i^\pi(t)$ the number of class-i users at time t and by $W_i^\pi(t)$ their total residual amount of work. Define $N^\pi(t) := \sum_{i=0}^{L} N_i^\pi(t)$ as the total number of users in the system at time t. Denote by $N_{i,<m_i}^\pi(t)$ and $N_{i,\geq m_i}^\pi(t)$ the number of class-i users with original service requirement smaller than m_i and larger than or equal to m_i, respectively. Similarly, we define $W_{i,<m_i}^\pi(t)$ and

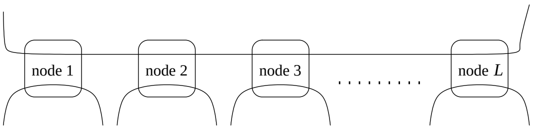

Figure 1: Linear network with L nodes.

$W^{\pi}_{i,\geq m_i}(t)$ as the amount of work consisting of class-i users with original service requirement smaller than m_i and larger than or equal to m_i, respectively. Denote by $\rho_{i,m_i} := \lambda_i \mathbb{P}(B_i < m_i)\mathbb{E}(B_i|B_i < m_i) = \lambda_i \int_0^{m_i^-} y\mathrm{d}B_i(y)$ the load composed of class-i users with original service requirement smaller than m_i. We further define N^{π}_i, W^{π}_i, N^{π}, $N^{\pi}_{i,<m_i}$, $N^{\pi}_{i,\geq m_i}$, $W^{\pi}_{i,<m_i}$ and $W^{\pi}_{i,\geq m_i}$ as random variables with the corresponding steady-state distributions (when they exist). Throughout, we assume $\rho_0 + \rho_i < 1$ for all $i = 1, \ldots, L$, which are obviously necessary conditions for stability. In fact, these are known [2] to be sufficient for the family of α-fair bandwidth-sharing policies as introduced by [11], provided $\alpha > 0$. (For conciseness, these conditions will be referred to as the 'standard' conditions.) As mentioned earlier, the flow-level performance of α-fair policies is in general intractable. In the special case of $\alpha \to 1$, i.e., Proportional Fairness (PF), however, the joint distribution of the numbers of users of the various classes in the above-described network has a product-form and is insensitive to the service requirement distributions, see [10]. In particular, the mean numbers of users are given by $\mathbb{E}(N_0^{PF}) = \frac{\rho_0}{1-\rho_0}(1+\sum_{i=1}^{L} \frac{\rho_i}{1-\rho_0-\rho_i})$ and $\mathbb{E}(N_i^{PF}) = \frac{\rho_i}{1-\rho_0-\rho_i}$ for $i = 1, \ldots, L$. Thus the mean total number of users is

$$\mathbb{E}(N^{PF}) = \frac{1}{1-\rho_0}\Big(\rho_0 + \sum_{i=1}^{L} \frac{\rho_i}{1-\rho_0-\rho_i}\Big). \tag{1}$$

3 Preliminaries

In the remainder of the paper we seek to identify policies that minimize the mean total number of users in the system. Because of Little's law, minimizing the mean number of users is equivalent to minimizing the mean sojourn time, and thus also equivalent to maximizing the user throughput defined as the ratio between the mean service requirement and the mean sojourn time. We distinguish the following classes of policies.

- $\bar{\Pi} \subseteq \Pi$ is the class of non-anticipating policies: $\bar{\pi} \in \bar{\Pi}$ if $\bar{\pi}$ uses no knowledge of the (remaining) service requirements.
- $\hat{\Pi} \subseteq \Pi$ is the class of work-conserving policies: $\hat{\pi} \in \hat{\Pi}$ if $\hat{\pi}$ utilizes the full service rate at any node i that is backlogged.
- $\Pi^* \subseteq \hat{\Pi}$: $\pi^* \in \Pi^*$ if π^* gives preemptive priority to class 0 whenever it is backlogged. Otherwise, all other classes with a backlog are served simultaneously.
- $\Pi^{**} \subseteq \hat{\Pi}$: $\pi^{**} \in \Pi^{**}$ if π^{**} serves classes $i = 1, \ldots, L$ whenever they are all simultaneously backlogged. Otherwise class 0 is served. When class 0 is non-backlogged, all other classes with a backlog are served simultaneously.

Observe that under any policy $\hat{\pi} \in \hat{\Pi}$ the total workload in any node i behaves as that of a single work-conserving server offered traffic from classes 0 and i. It immediately follows that any policy $\hat{\pi} \in \hat{\Pi}$ ensures stability under the 'standard' conditions mentioned earlier. In addition, the Pollaczek-Khintchine formula gives

$$\mathbb{E}(W_0^{\hat{\pi}}) + \mathbb{E}(W_i^{\hat{\pi}}) = \frac{\lambda_0 \mathbb{E}({B_0}^2) + \lambda_i \mathbb{E}({B_i}^2)}{2(1-\rho_0-\rho_i)}, \tag{2}$$

for any policy $\hat{\pi} \in \hat{\Pi}$. It further follows that any policy $\hat{\pi} \in \hat{\Pi}$ minimizes the total workload in any node i at every point in time. More specifically, if $W_0^{\hat{\pi}}(0) + W_i^{\hat{\pi}}(0) \leq_{st} W_0^{\pi}(0) + W_i^{\pi}(0)$ for some arbitrary policy $\pi \in \Pi$, then for all $t \geq 0$

$$W_0^{\hat{\pi}}(t) + W_i^{\hat{\pi}}(t) \leq_{st} W_0^{\pi}(t) + W_i^{\pi}(t), \tag{3}$$

with $\leq_{st}$ denoting the standard stochastic ordering.
Note that the definition of the policies in Π^* and Π^{**} only pertains to how the service rate is allocated among the various classes, and not to the scheduling among users within classes. Since $\Pi^*, \Pi^{**} \subseteq \hat{\Pi}$, all policies in these two classes satisfy inequality (3) for all $i = 1, \ldots, L$.
In addition, under a policy $\pi^* \in \Pi^*$, class 0 does not notice the presence of other classes. The mean amount of class-0 work is therefore given by the Pollaczek-Khintchine formula:

$$\mathbb{E}(W_0^{\pi^*}) = \frac{\lambda_0 \mathbb{E}({B_0}^2)}{2(1-\rho_0)}.$$

Substituting the latter equation in (2),

$$\mathbb{E}(W_i^{\pi^*}) = \frac{\lambda_0 \mathbb{E}({B_0}^2) + \lambda_i \mathbb{E}({B_i}^2)}{2(1-\rho_0-\rho_i)} - \frac{\lambda_0 \mathbb{E}({B_0}^2)}{2(1-\rho_0)}.$$

For a policy $\pi^{**} \in \Pi^{**}$ there are no closed-form expressions available for the individual mean workloads of the various classes. For $L = 2$, determining these amounts to solving a boundary-value problem [5]: the service rate allocated to any class i depends on the workloads of the other classes. However, we can compare (sample-path wise) the workloads of the various classes under different policies. Call $W_{0,j,k}^{\pi}(t) := W_0^{\pi}(t) + W_j^{\pi}(t) + W_k^{\pi}(t)$ the aggregate workload in nodes j and k. Note that the aggregate workload differs from the sum of the workloads in the two nodes as the workload of class 0 is only counted once. Besides minimizing the workload in any individual node at every point in time, a policy $\pi^{**} \in \Pi^{**}$ also minimizes the aggregate workload in at least one pair of nodes (these need not always be the same pair of nodes) as is formalized in the following lemma.

Lemma 3.1 *Let $\pi^{**} \in \Pi^{**}$ and $\pi \in \Pi$. If for $t = 0$ there exist nodes j and k with $j \neq k$, such that*

$$W_{0,j,k}^{\pi^{**}}(t) \leq W_{0,j,k}^{\pi}(t) \tag{4}$$

and the arrival and service requirement sequences are identical for both policies, then, for any $t > 0$, there exist j and k (not necessarily the same as at time $t = 0$) with $j \neq k$ such that (4) holds.

For $L = 2$, the above lemma implies that the class of policies Π^{**} *stochastically* minimize the total workload in the system. We note that there is no policy that achieves the same for $L > 2$. The proof of Lemma 3.1 can be found in [19].

The above results hold for arbitrary service requirement distributions and scheduling disciplines

within classes. In the remainder of the section we focus on the particular case of exponentially distributed service requirements, with $\mu_i = 1/\mathbb{E}(B_i)$, and restrict the attention to the class of non-anticipating policies $\bar{\Pi}$. In this case, scheduling within classes (without knowledge of the actual service requirements) does not affect the distribution of the numbers of users. We will show that, roughly speaking, for relatively 'large' values of μ_0, i.e., when class-0 users are relatively small, policies in either class $\bar{\Pi} \cap \Pi^*$ or $\bar{\Pi} \cap \Pi^{**}$ minimize the mean number of users at every point in time, among all policies in $\bar{\Pi}$.
To put our results into perspective, we recall that the 'μ-rule', i.e., granting priority to the user with the highest service rate, is known to stochastically minimize the number of users [16] in a single-server system. The rationale behind this rule is that it maximizes the departure rate at all times. In the present network context, besides trying to maximize the total departure rate, we must also take into account that when serving class $i \neq 0$ while another class $j \neq 0$ is non-backlogged but class 0 is backlogged may leave node j underutilized. For example, if $\mu_i > \mu_0$ for all $i = 1, \ldots, L$, then giving priority to classes $i = 1, \ldots, L$ myopically maximizes the total departure rate, but such a discipline unnecessarily causes instability [20] when $\rho_0 > \prod_{j=1}^{L}(1-\rho_j)$. In general, there can be a trade-off between maximizing the total departure rate and using the full capacity in each node whenever that node is backlogged. It is precisely in those cases where these two objectives are compatible, that we can identify relatively simple policies that minimize the total number of users.
The following two propositions [19] state that, in certain cases, policies in either class Π^* or Π^{**} stochastically minimize the total number of users at every point in time among all non-anticipating policies.

Proposition 3.2 *Let $\pi^* \in \bar{\Pi} \cap \Pi^*$. Assume $W_i^{\pi^*}(0) \leq_{st} W_i^{\pi}(0)$ for all i. Let the service requirements be exponentially distributed with $\mu_i = 1/\mathbb{E}(B_i)$ and $\sum_{i=1}^{L} \mu_i \leq \mu_0$. Then $N^{\pi^*}(t) \leq_{st} N^{\pi}(t)$ for all $\pi \in \bar{\Pi}$ and for all $t \geq 0$.*

Proposition 3.3 *Let $\pi^{**} \in \bar{\Pi} \cap \Pi^{**}$. Assume $W_i^{\pi^{**}}(0) \leq_{st} W_i^{\pi}(0)$ for all i. Let the service requirements be exponentially distributed with $\mu_i = 1/\mathbb{E}(B_i)$ and $\sum_{i=1}^{L} \mu_i \geq \mu_0 \geq \sum_{i=1, i\neq j}^{L} \mu_i$ for all $j \neq 0$. Then $N^{\pi^{**}}(t) \leq_{st} N^{\pi}(t)$ for all $\pi \in \bar{\Pi}$ and for all $t \geq 0$.*

Propositions 3.2 and 3.3 extend to the case where class-i users have hyperexponentially distributed service requirements with parameters p_{ij}, $\sum_{j=1}^{K_i} p_{ij} = 1$ and μ_{ij}, $j = 1, \ldots, K_i$. Optimality in expectation can then be established when either $\sum_{i=1}^{L} \mu_i^{max} \leq \mu_0^{min}$ or $\sum_{i=1}^{L} \mu_i^{min} \geq \mu_0^{max}$ and $\mu_0^{min} \geq \sum_{i=1, i\neq j}^{L} \mu_i^{max}$, for all $j \neq 0$, with $\mu_i^{min} = \min_{j=1,\ldots,K_i} \mu_{ij}$ and $\mu_i^{max} = \max_{j=1,\ldots,K_i} \mu_{ij}$.
The above results provide a strong notion of optimality, but involve correspondingly stringent assumptions. In the next sections, we will seek to find policies that are optimal under significantly milder conditions, although only in a suitable asymptotic sense. In particular, we will identify a wide class of policies that are optimal in a heavy-traffic regime. We will compare these policies with Proportional Fairness (PF) as a prototypical α-fair policy to assess the potential performance improvement, and demonstrate that the relative improvement can be arbitrarily large.

4 Single-node system in heavy traffic

Although the issue of concurrent resource possession only arises in network scenarios, we first present a heavy-traffic analysis of a single-node system in order to illustrate the key concepts and insights in the simplest possible context. In the next section we will return to the linear network model.

Consider a single-server system, where users arrive according to a Poisson process of rate λ and have service requirements B with $B(x) := \mathbb{P}(B \leq x)$. Denote the load by $\rho := \lambda\mathbb{E}(B) < 1$ and define $\rho_m := \lambda \int_0^{m^-} y\mathrm{d}B(y)$. For every policy π, the mean workload in the system obeys the lower bound $\mathbb{E}(W^\pi) \geq \frac{\lambda\mathbb{E}(B^2)}{2(1-\rho)}$, assuming $\mathbb{E}(B^2) < \infty$, with equality when policy π is work-conserving. We analyze a heavy-traffic regime where the system is critically loaded, i.e., $\lambda \uparrow \lambda^* := \frac{1}{\mathbb{E}(B)}$ (so $\rho \uparrow 1$ as ρ implicitly depends on λ). Motivated by classical heavy-traffic theory, we will consider the workload and number of users scaled by $1 - \rho$.

In order to improve the overall user performance, we can exploit the variability in service demands, and give precedence to small users over large ones. Specifically, we introduce a class of policies $\Pi_m \subseteq \hat{\Pi}\backslash\bar{\Pi}$ which use a simple threshold m to determine whether a user is small or large, and give preemptive priority to users with (original) service requirement smaller than m. Among users with an (original) service requirement larger than m, service is non-preemptive, i.e., the service of a user of size larger than m cannot be preempted by the service of another user of size larger than m.

Under policies in the class Π_m, the small users do not notice the presence of the large users, and experience similar performance as in a system without any large users. Now observe that the load in the latter system is ρ_m, and remains bounded away from 1, even when the load ρ approaches 1 as $\lambda \uparrow \lambda^*$ (assuming that there are in fact large users, i.e., $\mathbb{P}(B < m) < 1$). Hence, the small users are 'shielded' from the heavy-traffic conditions, as is formalized in the next proposition, which shows that the number of small users remains bounded as the load approaches the capacity.

Proposition 4.1 *For a policy $\pi_m \in \Pi_m$ with $\mathbb{P}(B < m) < 1$, it holds that $\mathbb{E}(N^{\pi_m}_{<m}) = \mathrm{O}(1)$ as $\lambda \uparrow \lambda^*$.*

Proof: Consider a policy $\pi_m \in \Pi_m$. Users of size smaller than m do not notice the presence of users of size larger than m. Therefore, $\mathbb{E}(W^{\pi_m}_{<m}) = \frac{\lambda\mathbb{P}(B<m)\mathbb{E}(B^2|B<m)}{2(1-\rho_m)} \leq \frac{\lambda m^2}{2(1-\rho_m)}$. The condition $\mathbb{P}(B < m) < 1$ is equivalent to $\rho_m < \rho$, so $\lim_{\lambda\uparrow\lambda^*} \rho_m < 1$. Hence, $\mathbb{E}(W^{\pi_m}_{<m}) = \mathrm{O}(1)$ as $\lambda \uparrow \lambda^*$. Now suppose that service is non-preemptive among users of size smaller than m as well (this assumption is not essential, see Remark 4.2 below). Then $(\mathbb{E}(N^{\pi_m}_{<m})-1)\mathbb{E}(B|B < m) \leq \mathbb{E}(W^{\pi_m}_{<m})$, which implies $\mathbb{E}(N^{\pi_m}_{<m}) = \mathrm{O}(1)$ as $\lambda \uparrow \lambda^*$. □

Remark 4.2 *The assumption in the proof of Proposition 4.1 that service is non-preemptive among users of size smaller than m, is not crucial. Instead, we could use that the preemptive Longest Remaining Processing Time (LRPT) discipline maximizes sample-path wise the number of users among all work-conserving policies. This follows from the fact that under LRPT all users leave together at the end of the busy period. Since users of size smaller than m receive preemptive priority under policy π_m, their mean number can be bounded by the mean number of users under LRPT in a system with only users of size smaller than m. The latter is given by $\mathbb{E}(N^{LRPT}_{<m}) = \lambda\mathbb{P}(B < m)(\frac{\mathbb{E}(B|B<m)}{1-\rho_m} + \frac{\lambda\mathbb{P}(B<m)\mathbb{E}(B^2|B<m)}{2(1-\rho_m)^2})$, see [7]. The result now follows by noting that $\lim_{\lambda\uparrow\lambda^*} \rho_m < 1$.*

Proposition 4.1 implies that the scaled mean number of small users tends to zero in heavy traffic. The number of large users can be bounded in terms of the total workload in the system. Hence, we obtain an upper bound for the scaled mean total number of users in the system, as provided in the next proposition.

Proposition 4.3 *For a policy $\pi_m \in \Pi_m$ with $\mathbb{P}(B < m) < 1$, it holds that $\lim_{\lambda\uparrow\lambda^*}(1-\rho)\mathbb{E}(N^{\pi_m}) \leq \frac{\lambda^*\mathbb{E}(B^2)}{2m}$.*

Proof: Consider a policy $\pi_m \in \Pi_m$. Note that $\mathbb{E}(W^{\pi_m}_{\geq m}) \geq m(\mathbb{E}(N^{\pi_m}_{\geq m}) - 1)$, because service is non-preemptive among users of size larger than m. Proposition 4.1 implies in particular that the scaled mean number of users smaller than m converges to zero. This yields $\lim_{\lambda\uparrow\lambda^*}(1-\rho)\mathbb{E}(N^{\pi_m}) = \lim_{\lambda\uparrow\lambda^*}(1-\rho)\mathbb{E}(N^{\pi_m}_{\geq m}) = \lim_{\lambda\uparrow\lambda^*}(1-\rho)(\mathbb{E}(N^{\pi_m}_{\geq m}) - 1) \leq \lim_{\lambda\uparrow\lambda^*}(1-\rho)\frac{\mathbb{E}(W^{\pi_m}_{\geq m})}{m} \leq \lim_{\lambda\uparrow\lambda^*}(1-\rho)\frac{\mathbb{E}(W^{\pi_m})}{m} = \lim_{\lambda\uparrow\lambda^*}(1-\rho)\frac{\lambda\mathbb{E}(B^2)}{2(1-\rho)m} = \frac{\lambda^*\mathbb{E}(B^2)}{2m}$. □

4.1 Comparison with Processor Sharing

The next proposition provides a comparison of the policies in the class $\cup_m\Pi_m$ with the Processor-Sharing (PS) discipline, which corresponds to the PF policy in a single-node system.

Proposition 4.4 *Let $\pi_m \in \Pi_m$. When B has infinite support, we have*

$$\lim_{m\to\infty}\lim_{\lambda\uparrow\lambda^*}\frac{\mathbb{E}(N^{\pi_m})}{\mathbb{E}(N^{PS})} = 0.$$

When B has finite support, we have

$$\lim_{m\uparrow M}\lim_{\lambda\uparrow\lambda^*}\frac{\mathbb{E}(N^{\pi_m})}{\mathbb{E}(N^{PS})} \leq \frac{\lambda^*\mathbb{E}(B^2)}{2M}.$$

Proof: It is well-known that $\mathbb{E}(N^{PS}) = \rho/(1-\rho)$, so $\lim_{\lambda\uparrow\lambda^*}(1-\rho)\mathbb{E}(N^{PS}) = 1$. Invoking Proposition 4.3, we obtain $\lim_{\lambda\uparrow\lambda^*}\frac{\mathbb{E}(N^{\pi_m})}{\mathbb{E}(N^{PS})} \leq \frac{\lambda^*\mathbb{E}(B^2)}{2m}$ for any m such that $\rho_m < \rho$, which proves both assertions. □

In case B has infinite support, it may be deduced that a policy from the class $\cup_m\Pi_m$ can outperform PS by an arbitrarily large factor. In case B has finite support, the ratio $\lambda^*\mathbb{E}(B^2)/M$ can be arbitrarily small for a wide range of service requirement distributions since $\mathbb{E}(B^2) \leq \frac{1}{\lambda}\Big[k\rho_k + M(1-\rho_k)\Big]$. These two findings may be intuitively explained as follows. Under the PS discipline, the total workload is distributed across users of various sizes, in proportion to their share in the total load, and hence the total number of users grows linearly with the workload as $\lambda \uparrow \lambda^*$. In contrast, under policies in the class $\cup_m\Pi_m$ the overwhelming fraction of the workload is contributed by users of size larger than m as $\lambda \uparrow \lambda^*$. Thus, as the value of m increases, the entire workload is concentrated in fewer and fewer users compared to the PS discipline.

Remark 4.5 *The assumption that service is non-preemptive among users of size larger than m under policies in the class Π_m, is not essential. For example, for the first statement of Proposition 4.4 to hold, it is sufficient that $\lim_{m\to\infty}\lim_{\lambda\uparrow\lambda^*}\frac{\mathbb{E}(W)}{\mathbb{E}(N^{\pi_m}_{\geq m})} = \infty$, which is valid under considerably milder assumptions. Then we have*

$$\lim_{m\to\infty}\lim_{\lambda\uparrow\lambda^*}\frac{\mathbb{E}(N^{\pi_m}_{\geq m})}{\mathbb{E}(N^{PS})} = \lim_{m\to\infty}\lim_{\lambda\uparrow\lambda^*}\frac{\mathbb{E}(N^{\pi_m}_{\geq m})}{\mathbb{E}(W)}\frac{\mathbb{E}(B^2)}{2\mathbb{E}(B)} = 0,$$

since $\mathbb{E}(W) = \frac{\mathbb{E}(B^2)}{2\mathbb{E}(B)}\mathbb{E}(N^{PS})$.

Remark 4.6 *Note that policies in Π_m rely on knowledge of the service requirements, which is not always easy to obtain. Instead, we could consider policies that give preemptive priority to users with* ***attained*** *service less than m. Let $\tilde{\pi}_m$ be such a policy. Denote by $\tilde{\rho}_m = \lambda \int_0^m y \mathrm{d}B(y) + \lambda m \mathbb{P}(B \geq m) < \rho$ the load due to users truncated at size m (users larger than or equal to m contribute an amount m, rather than zero as in ρ_m). Let $\tilde{N}^{\tilde{\pi}_m}_{<m}$ and $\tilde{N}^{\tilde{\pi}_m}_{\geq m}$ denote the number of users with attained service less than m and larger than or equal to m, respectively. We define $\tilde{W}^{\tilde{\pi}_m}_{<m}$ as the amount of work in the system consisting of users with attained service smaller than m, with their service requirement truncated at size m. Furthermore, let $\tilde{W}^{\tilde{\pi}_m}_{\geq m} = W^{\tilde{\pi}_m} - \tilde{W}^{\tilde{\pi}_m}_{<m}$.*
Since users with attained service less than m do not notice the presence of users that have attained more than m, we can upper bound the former by considering a system where users have service requirement $\min(B, m)$ and where we apply the LRPT discipline. This gives, $\mathbb{E}(\tilde{N}^{\tilde{\pi}_m}_{<m}) \leq \lambda(\frac{\mathbb{E}(\min(B,m))}{1-\tilde{\rho}_m} + \frac{\lambda \mathbb{E}((\min(B,m))^2)}{(1-\tilde{\rho}_m)^2})$. Using the fact that $\lim_{\lambda \uparrow \lambda^*} \tilde{\rho}_m < 1$ for $m < \infty$, we obtain $\lim_{\lambda \uparrow \lambda^*}(1-\rho)\mathbb{E}(\tilde{N}^{\tilde{\pi}_m}_{<m}) = 0$. Furthermore,

$$\begin{aligned}\lim_{\lambda \uparrow \lambda^*}(1-\rho)\mathbb{E}(\tilde{N}^{\tilde{\pi}_m}_{\geq m}) &= \lim_{\lambda \uparrow \lambda^*}(1-\rho)\frac{\mathbb{E}(W^{\tilde{\pi}_m}_{\geq m})}{\mathbb{E}(B|B>m)-m} \leq \lim_{\lambda \uparrow \lambda^*}(1-\rho)\frac{\mathbb{E}(W)}{\mathbb{E}(B|B>m)-m} \\ &= \lim_{\lambda \uparrow \lambda^*}(1-\rho)\mathbb{E}(W)\frac{\mathbb{P}(B>m)}{\int_m^\infty \mathbb{P}(B>y)\mathrm{d}y} = \frac{\mathbb{P}(B>m)}{\int_m^\infty \mathbb{P}(B>y)\mathrm{d}y}\lim_{\lambda \uparrow \lambda^*}(1-\rho)\mathbb{E}(W),\end{aligned}$$

where we need $\mathbb{E}(B) < \infty$ in the third equality. For service requirement distributions with

$$\lim_{m \to \infty} \frac{\mathbb{P}(B>m)}{\int_m^\infty \mathbb{P}(B>y)\mathrm{d}y} = 0, \tag{5}$$

we then obtain $\lim_{m\to\infty}\lim_{\lambda \uparrow \lambda^*} \frac{\mathbb{E}(N^{\tilde{\pi}_m})}{\mathbb{E}(N^{PS})} = \frac{\mathbb{E}(\tilde{N}^{\tilde{\pi}_m}_{\geq m})}{\mathbb{E}(N^{PS})} = 0$, i.e. these non-anticipating policies can outperform PS by an arbitrarily large factor.
An important class of policies that satisfy condition (5) are heavy-tailed distributions, i.e. $\lim_{y\to\infty}\frac{\mathbb{P}(B>y+z)}{\mathbb{P}(B>y)} = 1$ for any z, that have a decreasing failure rate (DFR), i.e. $\frac{f_B(y)}{1-B(y)}$ is decreasing in y (with $f_B(\cdot)$ the density function of B). Note that many heavy-tailed distributions, such as Pareto, satisfy the DFR property. Since the service requirement distribution is of type DFR, the function $\frac{\mathbb{P}(B>m+z)}{\mathbb{P}(B>m)}$ is monotone in m. From the monotone convergence theorem we then obtain

$$\begin{aligned}\lim_{m\to\infty}\frac{\int_m^\infty \mathbb{P}(B>y)\mathrm{d}y}{\mathbb{P}(B>m)} &= \lim_{m\to\infty}\int_0^\infty \frac{\mathbb{P}(B>m+z)}{\mathbb{P}(B>m)}\mathrm{d}z \\ &= \int_0^\infty \lim_{m\to\infty}\frac{\mathbb{P}(B>m+z)}{\mathbb{P}(B>m)}\mathrm{d}z = \infty,\end{aligned}$$

where the third equality follows from the heavy-tailed assumption.

4.2 Optimality properties

The next proposition shows that for any policy π, there exists a policy in $\cup_m \Pi_m$ that performs at least as well as π in heavy traffic. In other words, the class of policies $\cup_m \Pi_m$ is asymptotically optimal. This may be heuristically interpreted as follows. As mentioned above, under policies in the class $\cup_m \Pi_m$ the vast bulk of the workload is concentrated in users of size larger than m, while at the same time the total workload is minimal. In case B has finite support and the value of m is close to M, it is not possible to achieve a smaller number of users for the given workload. (When B has infinite support, it may be possible to reduce the number of users for a given workload yet further, by allowing preemptive service among large users.)

Proposition 4.7 *Let $\pi_m \in \Pi_m$. If B has finite support, then for any policy $\pi \in \Pi$,*

$$\lim_{m \uparrow M} \lim_{\lambda \uparrow \lambda^*} \frac{\mathbb{E}(N^{\pi_m})}{\mathbb{E}(N^{\pi})} \leq 1.$$

Proof: For any policy $\pi \in \Pi$,

$$\mathbb{E}(N^{\pi}) \geq \frac{\mathbb{E}(W^{\pi})}{M} \geq \frac{\lambda \mathbb{E}(B^2)}{2(1-\rho)M}. \tag{6}$$

Applying Proposition 4.3, taking $m < M$, we obtain

$$\lim_{\lambda \uparrow \lambda^*} (1-\rho)\mathbb{E}(N^{\pi_m}) \leq \frac{\lambda^* \mathbb{E}(B^2)}{2m}. \tag{7}$$

Comparing (6) and (7), and letting $m \uparrow M$ yields the assertion. □

5 Linear network in heavy traffic

In Section 3 we showed that in case of exponential service requirements, with relatively small class-0 users, policies in either class Π^* or Π^{**} are optimal among all non-anticipating policies. We will now explore whether, in a heavy-traffic regime, these results extend to more general service requirement distributions, now also allowing anticipating policies.

As described in Section 2, the linear network consists of L nodes and $L+1$ classes of users. We impose that $p_1\mathbb{E}(B_1) = \ldots = p_L\mathbb{E}(B_L)$, so that $\rho_1 = \ldots = \rho_L$. We analyze a heavy-traffic regime where each node is critically loaded, i.e., $\rho_0 + \rho_i =: \rho \uparrow 1$ for all $i = 1, \ldots, L$. This is equivalent to $\lambda \uparrow \lambda^* := (p_0\mathbb{E}(B_0) + p_i\mathbb{E}(B_i))^{-1}$. We will consider the workload and number of users scaled by $1-\rho$.

Just like for the single-node system in Section 4, we focus on simple priority-type strategies. In Section 5.1 we analyze a class of policies where class 0 is favored, while in Section 5.2 policies are studied which simultaneously favor classes $i = 1, \ldots, L$.

5.1 Favoring class 0

We first consider policies that serve either class-i users of size smaller than m_i for all $i = 1, \ldots, L$ simultaneously or class-0 users. If that is not possible, then classes $i = 1, \ldots, L$ are served, with class-i users with service requirement smaller than m_i receiving priority. Other than that, the priority structure within each of the classes is not essential for the analysis. Service is non-preemptive among class-i users of original size larger than m_i, i.e., the service of a class-i user of size larger than m_i cannot be preempted by the service of another class-i user of size larger than m_i. We denote this class of anticipating policies by $\Pi^*_{\boldsymbol{m}}$, where $\boldsymbol{m} \equiv (m_1, \ldots, m_L)$. We adopt the notation $\boldsymbol{m} \uparrow \boldsymbol{M}$ and $\boldsymbol{m} \to \infty$ to indicate that $m_i \uparrow M_i$ for all $i = 1, \ldots, L$ and $m_i \to \infty$ for all $i = 1, \ldots, L$, respectively (order is irrelevant).

Under policies in the class $\Pi^*_{\boldsymbol{m}}$, class-0 users and small class-i users do not notice the presence of the large class-i users, and experience similar performance as in a system without any large class-i users. Now observe that the load at node i in the latter system is $\rho_0 + \rho_{i,m_i}$, and remains bounded away from 1, even when the load ρ approaches 1 as $\lambda \uparrow \lambda^*$ (provided that there are in fact large class-i users, i.e., $\mathbb{P}(B_i < m_i) < 1$). Hence, the class-0 users and small class-i users are 'immune' from the heavy-traffic conditions, as is proved in the next proposition, which shows that the number of class-0 users and small class-i users remains bounded as the load approaches the capacity.

Proposition 5.1 *For a policy $\pi^*_{\boldsymbol{m}} \in \Pi^*_{\boldsymbol{m}}$ with $\mathbb{P}(B_i < m_i) < 1$ for all $i = 1, \dots, L$, it holds that $\mathbb{E}(N_0^{\pi^*_{\boldsymbol{m}}}) = \mathrm{O}(1)$ and $\mathbb{E}(N_{i,<m_i}^{\pi^*_{\boldsymbol{m}}}) = \mathrm{O}(1)$ as $\lambda \uparrow \lambda^*$.*

Proof: Consider a policy $\pi^*_{\boldsymbol{m}} \in \Pi^*_{\boldsymbol{m}}$. Class-0 users and class-i users ($i = 1, \dots, L$) of size smaller than m_i do not notice the presence of class-i users of size larger than m_i. Policy $\pi^*_{\boldsymbol{m}}$ is work-conserving, and therefore $\mathbb{E}(W_0^{\pi^*_{\boldsymbol{m}}}) + \mathbb{E}(W_{i,<m_i}^{\pi^*_{\boldsymbol{m}}}) = \lambda \frac{p_0 \mathbb{E}(B_0^2) + p_i \mathbb{P}(B_i < m_i) \mathbb{E}(B_i^2 | B_i < m_i)}{1 - \rho_0 - \rho_{i,m_i}}$. The condition $\mathbb{P}(B_i < m_i) < 1$ is equivalent to $\rho_{i,m_i} < \rho_i < 1 - \rho_0$, so $\lim_{\lambda \uparrow \lambda^*} 1 - \rho_0 - \rho_{i,m_i} > 0$. Hence, we conclude that

$$\mathbb{E}(W_0^{\pi^*_{\boldsymbol{m}}}) + \mathbb{E}(W_{i,<m_i}^{\pi^*_{\boldsymbol{m}}}) = \mathrm{O}(1), \text{ as } \lambda \uparrow \lambda^*. \tag{8}$$

Now suppose that service among class-0 users and class-i users of size smaller than m_i, $i = 1, \dots, L$, is non-preemptive as well (this assumption is not essential, see Remark 5.2 below). Then $(\mathbb{E}(N_0^{\pi^*_{\boldsymbol{m}}}) - 1)\mathbb{E}(B_0) \leq \mathbb{E}(W_0^{\pi^*_{\boldsymbol{m}}})$ and $(\mathbb{E}(N_{i,<m_i}^{\pi^*_{\boldsymbol{m}}}) - 1)\mathbb{E}(B_i | B_i < m_i) \leq \mathbb{E}(W_{i,<m_i}^{\pi^*_{\boldsymbol{m}}})$. Together with (8) this proves both claims. □

Remark 5.2 *In a similar way as in the single-node case, Proposition 5.1 can also be proved without the non-preemptive assumption with regard to class-0 users and class-i users smaller than m_i. Under policy $\pi^*_{\boldsymbol{m}}$, these users do not notice the presence of class-i users of size larger than m_i. Since each node is work-conserving we can therefore upper bound the mean number of users by considering the LRPT discipline in a system with only class-0 users and class-i users smaller than m_i. This gives*

$$\mathbb{E}(N_0^{\pi^*_{\boldsymbol{m}}}) + \mathbb{E}(N_{i,<m_i}^{\pi^*_{\boldsymbol{m}}}) \leq \lambda(p_0 + p_i \mathbb{P}(B_i < m_i))\Big(\frac{\mathbb{E}(\bar{B})}{1 - \rho_0 - \rho_{i,<m_i}} + \frac{\lambda(p_0 + p_i \mathbb{P}(B_i < m_i))\bar{B}^2}{2(1 - \rho_0 - \rho_{i,<m_i})^2}\Big),$$

with $\bar{B} = \frac{p_0}{p_0 + p_i \mathbb{P}(B_i < m_i)} \mathbb{E}(B_0) + \frac{p_i \mathbb{P}(B_i < m_i)}{p_0 + p_i \mathbb{P}(B_i < m_i)} \mathbb{E}(B_i | B_i < m_i)$ and $\bar{B}^2 = \frac{p_0}{p_0 + p_i \mathbb{P}(B_i < m_i)} \mathbb{E}(B_0^2) + \frac{p_i \mathbb{P}(B_i < m_i)}{p_0 + p_i \mathbb{P}(B_i < m_i)} \mathbb{E}(B_i^2 | B_i < m_i)$. The result now follows by noting that $\lim_{\lambda \uparrow \lambda^*} 1 - \rho_0 - \rho_{i,m_i} > 0$.

Proposition 5.1 implies that the scaled mean number of class-0 users and small class-i users tends to zero in heavy traffic. The number of large class-i users can be bounded in terms of the total workload at node i. Hence, we obtain an upper bound for the scaled mean total number of users in the system, as provided in the next proposition.

Proposition 5.3 *For a policy $\pi^*_{\boldsymbol{m}} \in \Pi^*_{\boldsymbol{m}}$ with $\mathbb{P}(B_i < m_i) < 1$ for all $i = 1, \dots, L$, it holds that*

$$\lim_{\lambda \uparrow \lambda^*} (1 - \rho)\mathbb{E}(N^{\pi^*_{\boldsymbol{m}}}) \leq \lambda^* \sum_{i=1}^{L} \frac{p_0 \mathbb{E}(B_0^2) + p_i \mathbb{E}(B_i^2)}{2m_i}.$$

Proof: It follows from Proposition 5.1 that $\lim_{\lambda \uparrow \lambda^*}(1-\rho)\mathbb{E}(N^{\pi^*_{\boldsymbol{m}}}) = \lim_{\lambda \uparrow \lambda^*}(1-\rho)\sum_{i=1}^{L} \mathbb{E}(N_{i,\geq m_i}^{\pi^*_{\boldsymbol{m}}})$. Since service is non-preemptive among class-i users of original size larger than m_i, we have $\mathbb{E}(W_{i,\geq m_i}^{\pi^*_{\boldsymbol{m}}}) \geq m_i(\mathbb{E}(N_{i,\geq m_i}^{\pi^*_{\boldsymbol{m}}}) - 1)$. Hence, $\lim_{\lambda \uparrow \lambda^*}(1-\rho)(\mathbb{E}(N_{i,\geq m_i}^{\pi^*_{\boldsymbol{m}}}) - 1) \leq \lim_{\lambda \uparrow \lambda^*}(1-\rho)\frac{\mathbb{E}(W_{i,\geq m_i}^{\pi^*_{\boldsymbol{m}}})}{m_i} \leq \lim_{\lambda \uparrow \lambda^*}(1-\rho)\frac{\mathbb{E}(W_i^{\pi^*_{\boldsymbol{m}}}) + \mathbb{E}(W_0^{\pi^*_{\boldsymbol{m}}})}{m_i} = \lambda^* \frac{p_0 \mathbb{E}(B_0^2) + p_i \mathbb{E}(B_i^2)}{2m_i}$ for $i = 1, \dots, L$, which proves the statement. □

5.1.1 Comparison with Proportional Fairness

We now compare the performance of the policies in the class $\cup_{\boldsymbol{m}}\Pi^*_{\boldsymbol{m}}$ with that of PF as a natural extension of PS.

Proposition 5.4 *Let $\pi^*_{\boldsymbol{m}} \in \Pi^*_{\boldsymbol{m}}$.*
When $B_1, \ldots, B_L$ have infinite support, we have

$$\lim_{\boldsymbol{m}\to\infty} \lim_{\lambda\uparrow\lambda^*} \frac{\mathbb{E}(N^{\pi^*_{\boldsymbol{m}}})}{\mathbb{E}(N^{PF})} = 0.$$

When $B_1, \ldots, B_L$ have finite support, we have

$$\lim_{\boldsymbol{m}\uparrow\boldsymbol{M}} \lim_{\lambda\uparrow\lambda^*} \frac{\mathbb{E}(N^{\pi^*_{\boldsymbol{m}}})}{\mathbb{E}(N^{PF})} \leq \frac{\lambda^*}{L}\sum_{i=1}^{L}\frac{p_0\mathbb{E}(B_0^2)+p_i\mathbb{E}(B_i^2)}{2M_i}.$$

Proof: From (1),

$$\lim_{\lambda\uparrow\lambda^*}(1-\rho)\mathbb{E}(N^{PF}) = \lim_{1-\rho_0-\rho_i\downarrow 0}(1-\rho_0-\rho_i)\mathbb{E}(N^{PF}) = \lim_{1-\rho_0-\rho_i\downarrow 0}\frac{\sum_{i=1}^{L}\rho_i}{1-\rho_0} = \frac{L(1-\rho_0)}{1-\rho_0} = L. \tag{9}$$

Together with Proposition 5.3, this proves the assertion. □

We deduce that when $B_1, \ldots, B_L$ have infinite support, there exists a policy $\pi^*_m \in \cup_{\boldsymbol{m}}\Pi^*_{\boldsymbol{m}}$ that outperforms PF by an arbitrarily large factor in a heavy-traffic regime. This may be intuitively explained as follows. Under the PF discipline, the total workload is distributed across users of various sizes, and hence the total number of users grows linearly with the workload as $\lambda \uparrow \lambda^*$. In contrast, under policies in the class $\cup_m\Pi_m$ the dominant fraction of the workload is contributed by class-i users of size larger than m_i as $\lambda \uparrow \lambda^*$. Thus, as the value of m_i increases, the entire workload is concentrated in fewer and fewer users compared to the PF discipline.
Comparing Propositions 4.4 and 5.4 we observe that the relative improvement over the PF policy achieved by policies in the class $\pi^*_{\boldsymbol{m}}$ is equal to the average relative improvement that would have been obtained over the PS policy by policies in the class Π_m in each of the L nodes separately.

5.1.2 Optimality properties

We now assume that B_i has finite support for all classes $i = 0, \ldots, L$, with $\sum_{i=1}^{L}\frac{1}{M_i} \leq \frac{1}{M_0}$. The next proposition shows that for any policy $\pi \in \Pi$, there exists a policy in $\cup_{\boldsymbol{m}}\Pi^*_{\boldsymbol{m}}$ that performs at least as well in heavy-traffic conditions. This may be heuristically interpreted as follows. As mentioned above, under policies in the class $\cup_{\boldsymbol{m}}\Pi^*_{\boldsymbol{m}}$ the lion share of the workload is composed of class-i users of size larger than m_i, while at the same time the total workload in each node is minimized. In case $\sum_{i=1}^{L}\frac{1}{M_i} \leq \frac{1}{M_0}$, and the value of m_i is close to M_i, it is not possible to achieve a smaller total number of users for the given workload as attained under policies in $\cup_{\boldsymbol{m}}\Pi^*_{\boldsymbol{m}}$.

Proposition 5.5 *Let $\pi^*_{\boldsymbol{m}} \in \Pi^*_{\boldsymbol{m}}$. Assume $M_i < \infty$ for $i = 0, \ldots, L$ and $\sum_{i=1}^{L}\frac{1}{M_i} \leq \frac{1}{M_0}$. Then for any policy $\pi \in \Pi$,*

$$\lim_{\boldsymbol{m}\uparrow\boldsymbol{M}} \lim_{\lambda\uparrow\lambda^*} \frac{\mathbb{E}(N^{\pi^*_{\boldsymbol{m}}})}{\mathbb{E}(N^{\pi})} \leq 1.$$

Proof: Policy $\pi^*_{\boldsymbol{m}} \in \Pi^*_{\boldsymbol{m}}$ is work-conserving in all nodes. Therefore we have for any policy $\pi \in \Pi$,

$$\mathbb{E}(W_0^{\pi^*_{\boldsymbol{m}}}) + \mathbb{E}(W_i^{\pi^*_{\boldsymbol{m}}}) \leq \mathbb{E}(W_0^{\pi}) + \mathbb{E}(W_i^{\pi}). \tag{10}$$

Proposition 5.1 implies that $\lim_{\lambda\uparrow\lambda^*}(1-\rho)\mathbb{E}(N_{i,<m_i}^{\pi^*_{\boldsymbol{m}}}) = 0$. In conjunction with $(\mathbb{E}(N_{i,\geq m_i}^{\pi^*_{\boldsymbol{m}}}) - 1)m_i \leq \mathbb{E}(W_{i,\geq m_i}^{\pi^*_{\boldsymbol{m}}})$, this yields that $\lim_{\lambda\uparrow\lambda^*}(1-\rho)\mathbb{E}(N_i^{\pi^*_{\boldsymbol{m}}})m_i \leq \lim_{\lambda\uparrow\lambda^*}(1-\rho)\mathbb{E}(W_{i,>m_i}^{\pi^*_{\boldsymbol{m}}})$. It also follows from Proposition 5.1 that $\lim_{\lambda\uparrow\lambda^*}(1-\rho)\mathbb{E}(N_0^{\pi^*_{\boldsymbol{m}}}) = 0$. Furthermore, we have $\mathbb{E}(W_i^{\pi}) \leq \mathbb{E}(N_i^{\pi})M_i$ for every policy π. Together with (10), this results in

$$(1-\rho)\Big(\mathbb{E}(N_0^{\pi^*_{\boldsymbol{m}}})M_0 + \mathbb{E}(N_i^{\pi^*_{\boldsymbol{m}}})m_i\Big) \leq (1-\rho)\Big(\mathbb{E}(N_0^{\pi})M_0 + \mathbb{E}(N_i^{\pi})M_i\Big) + o(1-\rho), \tag{11}$$

for $i = 1, \ldots, L$. Also,

$$0 = (1-\rho)\mathbb{E}(N_0^{\pi^*_{\boldsymbol{m}}})M_0 \leq (1-\rho)\mathbb{E}(N_0^{\pi})M_0 + o(1-\rho). \tag{12}$$

Letting $m_i \uparrow M_i$, $i = 1, \ldots, L$, multiplying (11) by $\frac{1}{M_i}$ for all $i = 1, \ldots, L$, and (12) by $\frac{1}{M_0} - \sum_{i=1}^{L} \frac{1}{M_i} \geq 0$, and summing these $L+1$ inequalities, gives

$$\lim_{\boldsymbol{m}\uparrow\boldsymbol{M}} \lim_{\lambda\uparrow\lambda^*} (1-\rho) \sum_{i=0}^{L} \mathbb{E}(N_i^{\pi^*_{\boldsymbol{m}}}) \leq \lim_{\lambda\uparrow\lambda^*} (1-\rho) \sum_{i=0}^{L} \mathbb{E}(N_i^{\pi}). \ \Box$$

5.2 Favoring classes $i = 1, \ldots, L$ simultaneously

We now consider policies that serve classes $i = 1, \ldots, L$ simultaneously or serve class-0 users of size smaller than m_0. If that is not feasible, then class-0 users of size larger than m_0 are served. Within class i, users of size smaller than m_i receive priority, $i = 0, \ldots, L$. Other than that, the priority structure within each of the classes is irrelevant for the analysis. Service is non-preemptive among class-i users of size larger than m_i: their service may not be interrupted by the service of other class-i users of size larger than m_i. We denote this class of anticipating policies by $\Pi^{**}_{\boldsymbol{m}}$, where $\boldsymbol{m} \equiv (m_0, \ldots, m_L)$. As before, we use the notation $\boldsymbol{m} \uparrow \boldsymbol{M}$ and $\boldsymbol{m} \to \infty$ to indicate that $m_i \uparrow M_i$ for all $i = 0, \ldots, L$ and $m_i \to \infty$ for all $i = 0, \ldots, L$, respectively (order is irrelevant).

Under policies in the class $\Pi^{**}_{\boldsymbol{m}}$, the number of small class-0 users as well as the number of small class-i users remain bounded as the load approaches the capacity, just like in Proposition 5.1, but this is far more difficult to prove now. While these users indeed receive some degree of preferred treatment, it is no longer the case that they do not notice the presence of the large users. Observe that simultaneous service of large class-i users can have precedence over service of small class-0 users, and that small class-i users must be simultaneously present in order to receive priority over large class-0 users. Hence, in order to prove the above assertion, we need more elaborate arguments as provided in Lemma 5.6 below. Denote by $\hat{W}_i^c(t)$ the workload at time t in a reference system with class-i traffic only, service rate c, and with $\hat{W}_i^c(0) = 0$. Define $U_j^d(t) := \sup_{0\leq s\leq t}\{d(t-s) - A_j(s,t)\}$.

Lemma 5.6 *Let $\delta < 1 - \rho_j$, for all $j = 1, \ldots, L$, and $\pi^{**}_{\boldsymbol{m}} \in \Pi^{**}_{\boldsymbol{m}}$. Assume $W_{0,<m_0}^{\pi^{**}_{\boldsymbol{m}}}(0) = W_{i,<m_i}^{\pi^{**}_{\boldsymbol{m}}}(0) = 0$, for a certain $i \in \{1, \ldots, L\}$. Then at time $t \geq 0$, there exists a $j^* \in \{1, \ldots, L\}$, such that*

$$W_{0,<m_0}^{\pi^{**}_{\boldsymbol{m}}}(t) + W_{i,<m_i}^{\pi^{**}_{\boldsymbol{m}}}(t) \leq \hat{W}_{0,<m_0}^{\rho_0-\delta}(t) + \hat{W}_{i,<m_i}^{\rho_{i,m_i}-\delta}(t) + \hat{W}_{j^*}^{\rho_{j^*}+\delta}(t) + U_{j^*}^{\rho_{j^*}-\delta}(t).$$

The proof of Lemma 5.6 may be found in Appendix A.

Proposition 5.7 *For a policy $\pi^{**}_{\boldsymbol{m}} \in \Pi^{**}_{\boldsymbol{m}}$ with $\mathbb{P}(B_i < m_i) < 1$, $i = 0, \ldots, L$, it holds that $\mathbb{E}(N_{0,<m_0}^{\pi^{**}_{\boldsymbol{m}}}) = \mathrm{O}(1)$ and $\mathbb{E}(N_{i,<m_i}^{\pi^{**}_{\boldsymbol{m}}}) = \mathrm{O}(1)$ as $\lambda \uparrow \lambda^*$.*

Proof: Using Lemma 5.6, we obtain that

$$\mathbb{E}(W_{0,<m_0}^{\pi^{**}_{\boldsymbol{m}}}) + \mathbb{E}(W_{i,<m_i}^{\pi^{**}_{\boldsymbol{m}}}) \leq \mathbb{E}(\hat{W}_{0,<m_0}^{\rho_0-\delta}) + \mathbb{E}(\hat{W}_{i,<m_i}^{\rho_i-\delta}) + \sum_{j=1}^{L} \mathbb{E}(\hat{W}_j^{\rho_j+\delta}) + \sum_{j=1}^{L} \mathbb{E}(U_j^{\rho_j-\delta}).$$

For δ small enough, $\mathbb{E}(\hat{W}_{j,<m_j}^{\rho_j-\delta}) = \lambda \frac{p_j \mathbb{P}(B_j<m_j)\mathbb{E}(B_j^2|B_j<m_j)}{2(\rho_j-\delta-\rho_{j,<m_j})}$ and $\mathbb{E}(\hat{W}_j^{\rho_j+\delta}) = \lambda \frac{p_j \mathbb{E}(B_j^2)}{2\delta}$, which means $\mathbb{E}(\hat{W}_{j,<m_j}^{\rho_j-\delta}) = \mathbb{E}(\hat{W}_j^{\rho_j+\delta}) = \mathrm{O}(1)$. Furthermore, we can equivalently replace $U_{j^*}^{\rho_{j^*}-\delta}$ by the supremum of a random walk with drift $\rho_{j^*} - \delta - \rho_{j^*} = -\delta < 0$. The drift is negative, independently of λ, which implies that the mean of the supremum is finite in heavy traffic. Hence $\mathbb{E}(U_{j^*}^{\rho_j-\delta}) = \mathrm{O}(1)$. Together, this gives $\mathbb{E}(W_{0,<m_0}^{\pi^{**}_{\boldsymbol{m}}}) + \mathbb{E}(W_{i,<m_i}^{\pi^{**}_{\boldsymbol{m}}}) = \mathrm{O}(1)$. Using similar arguments as in the proof of Proposition 5.1, then yields the assertion. □

Proposition 5.7 implies that the scaled mean number of small class-0 and small class-i users tends to zero in heavy traffic. The number of large class-0 and large class-i users can be bounded in terms of the total workload at node i. Hence, we obtain an upper bound for the scaled mean total number of users in the system, as provided in the next proposition.

Proposition 5.8 *For a policy $\pi^{**}_{\boldsymbol{m}} \in \Pi^{**}_{\boldsymbol{m}}$ with $\mathbb{P}(B_i < m_i) < 1$ for all $i = 0, \ldots, L$, it holds that $\lim_{\lambda\uparrow\lambda^*}(1-\rho)\mathbb{E}(N^{\pi^{**}_{\boldsymbol{m}}}) \leq \lambda^* \frac{Lp_0\mathbb{E}(B_0^2)+\sum_{i=1}^{L} p_i \mathbb{E}(B_i^2)}{2\min(m_0,m_1,\ldots,m_L)}$.*

Proof: Proposition 5.7 indicates that the number of class-i users smaller than m_i, $i = 0, \ldots, L$, under policy $\pi^{**}_{\boldsymbol{m}}$ remains bounded as $\lambda \uparrow \lambda^*$. Since furthermore service is non-preemptive among class-i users of size larger than m_i, it follows that $(1-\rho)(\mathbb{E}(N_0^{\pi^{**}_{\boldsymbol{m}}}) + \mathbb{E}(N_i^{\pi^{**}_{\boldsymbol{m}}})) \leq (1-\rho)\frac{\mathbb{E}(W_0^{\pi^{**}_{\boldsymbol{m}}})+\mathbb{E}(W_i^{\pi^{**}_{\boldsymbol{m}}})}{\min(m_0,m_i)} + o(1-\rho)$ as $\lambda \uparrow \lambda^*$. So for any m_i such that $\rho_{i,m_i} < \rho_i$, the scaled mean total number of users is $\lim_{\lambda\uparrow\lambda^*}(1-\rho)\mathbb{E}(N^{\pi^{**}_{\boldsymbol{m}}}) \leq \lim_{\lambda\uparrow\lambda^*}(1-\rho)\Big(L\mathbb{E}(N_0^{\pi^{**}_{\boldsymbol{m}}}) + \sum_{i=1}^{L}\mathbb{E}(N_i^{\pi^{**}_{\boldsymbol{m}}})\Big) \leq \lim_{\lambda\uparrow\lambda^*}(1-\rho)\Big(\frac{\sum_{i=1}^{L}(\mathbb{E}(W_0^{\pi^{**}_{\boldsymbol{m}}})+\mathbb{E}(W_i^{\pi^{**}_{\boldsymbol{m}}}))}{\min(m_0,m_1,\ldots,m_L)}\Big) = \lambda^* \frac{Lp_0\mathbb{E}(B_0^2)+\sum_{i=1}^{L} p_i\mathbb{E}(B_i^2)}{2\min(m_0,m_1,\ldots,m_L)}$. □

5.2.1 Comparison with Proportional Fairness

As before, we now compare the performance of the policies in the class $\cup_{\boldsymbol{m}}\Pi^{**}_{\boldsymbol{m}}$ with that of PF.

Proposition 5.9 *Let $\pi^{**}_{\boldsymbol{m}} \in \Pi^{**}_{\boldsymbol{m}}$.*
When $B_0, B_1, \ldots, B_L$ have infinite support, we have

$$\lim_{\boldsymbol{m}\to\infty} \lim_{\lambda\uparrow\lambda^*} \frac{\mathbb{E}(N^{\pi^{**}_{\boldsymbol{m}}})}{\mathbb{E}(N^{PF})} = 0.$$

When $B_0, B_1, \ldots, B_L$ have finite support, we have

$$\lim_{\boldsymbol{m}\uparrow\boldsymbol{M}} \lim_{\lambda\uparrow\lambda^*} \frac{\mathbb{E}(N^{\pi^{**}_{\boldsymbol{m}}})}{\mathbb{E}(N^{PF})} = \frac{\lambda^*}{L}\frac{Lp_0\mathbb{E}(B_0^2) + \sum_{i=1}^{L} p_i\mathbb{E}(B_i^2)}{2\min(M_0, M_1, \ldots, M_L)}.$$

Proof: Proposition 5.8, together with (9), gives the result. □

As before, we conclude that when $B_0, B_1, \ldots, B_L$ have infinite support, there exists a policy $\pi_m^{**} \in \cup_{\boldsymbol{m}} \Pi_{\boldsymbol{m}}^{**}$ that outperforms PF by an arbitrarily large factor in heavy-traffic conditions.

5.2.2 Optimality properties

We now assume that B_i has finite support for all classes, with $\sum_{i=1}^{L} \frac{1}{M_i} \geq \frac{1}{M_0}$ and $\frac{1}{M_0} \geq \sum_{l=1, l \neq i} \frac{1}{M_l}$ for all $i = 0, \ldots, L$. The next proposition shows that for any policy $\pi \in \Pi$, there exists a policy in $\cup_{\boldsymbol{m}} \Pi_{\boldsymbol{m}}^{**}$ that performs at least as well in heavy-traffic conditions. As before, these policies manage to simultaneously minimize the workload in each of the nodes and concentrate the entire workload in users of maximum size (which in general may not be possible to accomplish).

Proposition 5.10 *Assume $M_i < \infty$ for all $i = 0, \ldots, L$, and $\sum_{i=1}^{L} \frac{1}{M_i} \geq \frac{1}{M_0}$ and $\frac{1}{M_0} \geq \sum_{l=1, l \neq i} \frac{1}{M_l}$ for all $i = 0, \ldots, L$. Let $\pi_{\boldsymbol{m}}^{**} \in \Pi_{\boldsymbol{m}}^{**}$. Then for any policy $\pi \in \Pi$,*

$$\lim_{\boldsymbol{m} \uparrow \boldsymbol{M}} \lim_{\lambda \uparrow \lambda^*} \frac{\mathbb{E}(N^{\pi_{\boldsymbol{m}}^{**}})}{\mathbb{E}(N^{\pi})} \leq 1.$$

The proof may be found in Appendix B. The idea of the proof may be described as follows. Instead of proving the result for an arbitrary policy in $\Pi_{\boldsymbol{m}}^{**}$, we first focus on a policy $\pi^p \in \Pi_{\boldsymbol{m}}^{**} \cap \Pi^{**}$. Lemma 3.1 holds for π^p, which will allow us to prove the optimality of this policy. Since the scaled workloads for policies in class $\Pi_{\boldsymbol{m}}^{**}$ are identical in heavy traffic, the optimality result applies to all policies in this class.

6 Numerical experiments

In the previous sections we compared the performance of policies in classes $\Pi_{\boldsymbol{m}}^{*}$ and $\Pi_{\boldsymbol{m}}^{**}$ with that of PF in a heavy-traffic regime. We will now focus on a subset of the class $\Pi_{\boldsymbol{m}}^{*}$, and conduct numerical experiments to illustrate the analytical findings and assess the scope for performance gains. We specifically examine those policies $\pi \in \Pi_{\boldsymbol{m}}^{*} \cap \Pi^{*}$ that serve class-i users of original size smaller than m_i, $i \neq 0$ and class-0 users in a non-preemptive fashion. Because of the non-preemptive feature, the following upper bounds hold:

$$\mathbb{E}(N_0^{\pi}) \leq 1 + \frac{\mathbb{E}(W_0^{\pi})}{\mathbb{E}(B_0)}, \quad \mathbb{E}(N_{i,<m_i}^{\pi}) \leq 1 + \frac{\mathbb{E}(W_{i,<m_i}^{\pi})}{\mathbb{E}(B_i | B_i < m_i)} \quad \text{and} \quad \mathbb{E}(N_{i,\geq m_i}^{\pi}) \leq 1 + \frac{\mathbb{E}(W_{i,\geq m_i}^{\pi})}{\mathbb{E}(B_i | B_i \geq m_i)}.$$

In case of exponentially distributed service requirements, the 1 in the right-hand side of the first equation may in fact be omitted. Since class 0 receives preemptive priority, its mean workload is

$$\mathbb{E}(W_0^{\pi}) = \frac{\lambda p_0 \mathbb{E}(B_0^2)}{2(1 - \rho_0)}.$$

Class-0 and class-i users of size smaller than m_i, $i \neq 0$, are served in a work-conserving manner, and therefore

$$\mathbb{E}(W_{i,<m_i}^{\pi}) = \frac{\lambda(p_0 \mathbb{E}(B_0^2) + p_i \mathbb{P}(B_i < m_i) \mathbb{E}(B_i^2 | B_i < m_i))}{2(1 - \rho_0 - \rho_{i,m_i})}.$$

Policy π is work-conserving, hence

$$\mathbb{E}(W^{\pi}_{i,\geq m_i}) = \frac{\lambda(p_0\mathbb{E}(B_0^2) + p_i\mathbb{E}(B_i^2))}{2(1-\rho_0-\rho_i)} - \mathbb{E}(W^{\pi}_{i,<m_i}) - \mathbb{E}(W^{\pi}_0).$$

In the numerical experiments, we considered a system with two nodes, $p_0 = 0.5$, $p_1 = 0.25$, $p_2 = 0.25$ and $m_1 = m_2 = m$. We studied both exponentially and Pareto distributed service requirements. In the former case, we took $\mu_0 = 2$, $\mu_1 = 1$, $\mu_2 = 1$, while in the latter case we chose $\alpha_0 = 10$, $\alpha_1 = 3$, $\alpha_2 = 3$ ($\mathrm{d}B_i(x) = \alpha_i x^{-(\alpha_i+1)}\mathrm{d}x$).
In Figures 2 and 3 we plotted the upper bound for the scaled mean number of class-0 users as a function of ρ. Note that the scaled mean number of class-0 users does not depend on m and as ρ increases it converges to zero. In Figures 4 and 5 we plotted the upper bound for the scaled mean number of class-1 users smaller than m_1 as a function of ρ. Again, as ρ increases, it converges to zero. Furthermore, we observe for a large fixed ρ, a horizontal asymptote as m grows large. This asymptote can be found by interchanging the order of limits, i.e. $\lim_{\lambda\uparrow\lambda^*}\lim_{m\to\infty}(1-\rho)\mathbb{E}(N^{\pi}_{1,<m_1}) = \lambda^*(p_0\mathbb{E}(B_0^2) + p_1\mathbb{E}(B_1^2))/(2\mathbb{E}(B_1))$. In Figures 6 and 7 we plotted the upper bound for the scaled mean number of class-1 users larger than m_1.

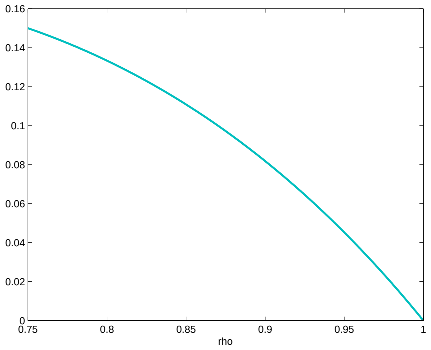

Figure 2: Upper bound for $(1-\rho)N^{\pi}_0$, for exponential service requirements.

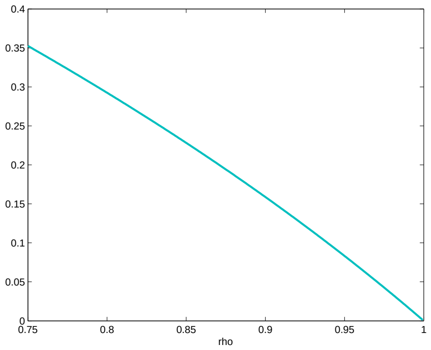

Figure 3: Upper bound for $(1-\rho)N^{\pi}_0$, for Pareto service requirements.

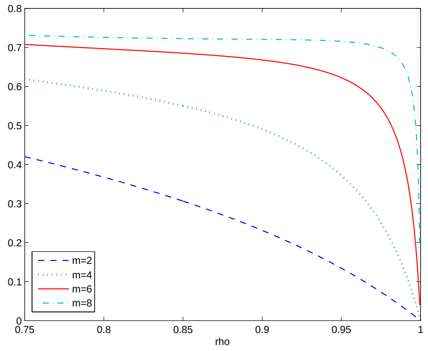

Figure 4: Upper bound for $(1-\rho)N^{\pi}_{1,<m_1}$, for exponential service requirements.

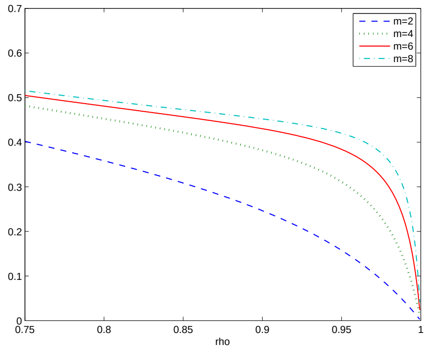

Figure 5: Upper bound for $(1-\rho)N^{\pi}_{1,<m_1}$, for Pareto service requirements.

The three bounds specified above provide an upper bound for the total mean number of users under policy π. In Figures 8 and 9 we plot the ratio between this upper bound and the total mean

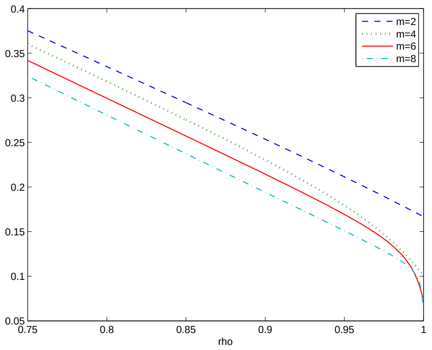

Figure 6: Upper bound for $(1-\rho)N^{\pi}_{1,\geq m_1}$, for exponential service requirements.

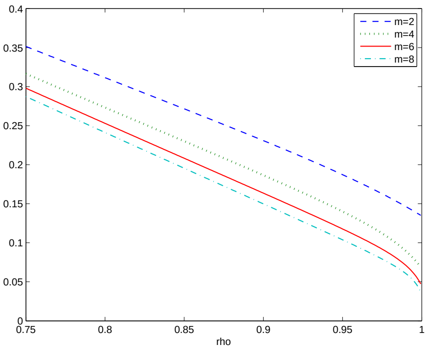

Figure 7: Upper bound for $(1-\rho)N^{\pi}_{1,\geq m_1}$, for Pareto service requirements.

number of users under PF as a function of ρ for exponentially and Pareto distributed service requirements, respectively. In both cases, the discipline with threshold $m = 2$ gives already a substantial performance improvement for a load of 0.85. It is worth observing here that we have pursued deliberately simple policies in order to obtain provable asymptotic performance guarantees. There are clearly more sophisticated policies conceivable that will typically achieve larger gains, but may be too complex to allow any explicit performance guarantees.

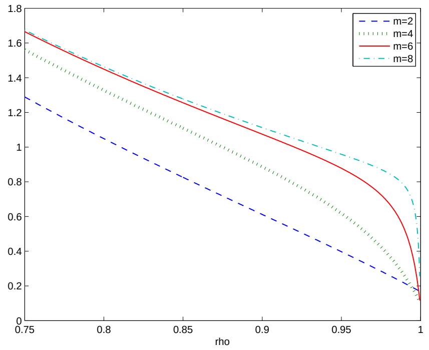

Figure 8: Upper bound for $\mathbb{E}(N^{\pi})/\mathbb{E}(N^{PF})$, for exponential service requirements.

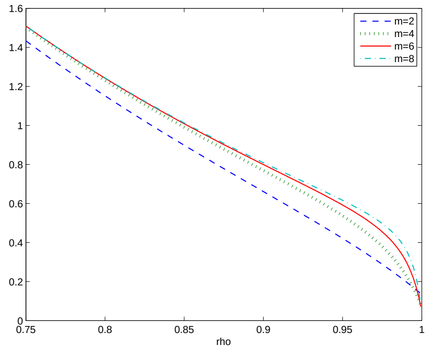

Figure 9: Upper bound for $\mathbb{E}(N^{\pi})/\mathbb{E}(N^{PF})$, for Pareto service requirements.

7 Summary and conclusions

Although valuable stability results have been obtained for the class of α-fair bandwidth-sharing strategies, it is not well understood to what extent the flow-level delays and throughputs leave potential room for improvement. In order to gain a better understanding of the latter issue, we set out to determine the scheduling policies that minimize the mean delay in some simple linear bandwidth-sharing networks. Rather than aiming for strictly optimal policies, we focused on a class of relatively simple priority-type strategies that only separate large flows from small ones. To benchmark the performance of these strategies, we compared them with Proportional Fair as the prototypical α-fair policy, and established that the mean delay may be reduced by an arbitrarily large factor when the load is sufficiently high. In addition, we showed the above strategies to be asymptotically optimal for flow size distributions with bounded support. Numerical experiments revealed that even at fairly moderate load values the performance gains can be significant. It is worth recalling here that we have focused on deliberately simple policies in order to obtain explicit asymptotic performance guarantees. There are evidently more advanced policies imaginable that will typically yield larger gains, but may be too complicated to allow any strict performance guarantees.

References

[1] Ben Fredj, S., Bonald, T., Proutière, A., Regnié, G., Roberts, J.W. (2001). Statistical bandwidth sharing: a study of congestion at the flow level. In: *Proc. ACM SIGCOMM 2001.*

[2] Bonald, T., Massoulié, L. (2001). Impact of fairness on Internet performance. In: *Proc. ACM SIGMETRICS & Performance 2001 Conf.*, Boston MA, USA, 82–91.

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

[4] Chen, X., Heidemann, J. (2003). Preferential treatment for short flows to reduce Web latency. *Comp. Netw.* **41**, 779–794.

[5] Cohen, J.W., Boxma, O.J. (1983). *Boundary Value Problems in Queueing System Analysis.* North-Holland, Amsterdam.

[6] Harchol-Balter, M., Schroeder, B., Bansal, N., Agrawal, M. (2003). Size-based scheduling to improve Web performance. *ACM Trans. Comp. Syst.* **21**, 207–233.

[7] Harchol-Balter, M., Sigman, K., Wierman, A. (2002). Asymptotic convergence of scheduling policies with respect to slowdown. In: *Proc. Performance 2002 Conf.*, Rome, Italy, 241–256.

[8] Kelly, F.P., Williams, R.J. (2004). Fluid model for a network operating under a fair bandwidth-sharing policy. *Ann. Appl. Prob.* **14**, 1055–1083.

[9] Kleinrock, L. (1976). *Queueing Systems, Vol. II: Computer Applications.* Wiley, New York.

[10] Massoulié, L., Roberts, J.W. (2000). Bandwidth sharing and admission control for elastic traffic. *Telecommun. Syst.* **15**, 185–201.

[11] Mo, J., Walrand, J. (2000). Fair end-to-end based congestion control. *IEEE/ACM Trans. Netw.* **8**, 556–567.

[12] Núñez-Queija, R. (2000). *Processor-Sharing Models for Integrated-Services Networks.* Ph.D. Thesis Eindhoven University of Technology.

[13] Rai, I.A., Urvoy-Keller, G., Biersack, E.W. (2003). Analysis of LAS scheduling for job size distributions with high variance. In: *Proc. ACM SIGMETRICS 2003 Conf.*, San Diego CA, USA, 218–228.

[14] Rai, I.A., Urvoy-Keller, G., Biersack, E.W. (2004). LAS scheduling approach to avoid bandwidth hogging in heterogeneous TCP networks. In: *7th IEEE International Conference on High-Speed Networks and Multimedia Communications HSNMC'04*, Toulouse, France.

[15] Rai, I.A., Urvoy-Keller, G., Vernon, M.K., Biersack, E.W. (2004). Performance analysis of LAS-based scheduling disciplines in a packet-switched network. In: *Proc. ACM SIGMETRICS & Performance 2004 Conf.*, New York NY, USA, 106–117.

[16] Righter, R., Shanthikumar, J.G. (1989). Scheduling multiclass single-server queueing systems to stochastically maximize the number of successful departures. *Prob. Eng. Inf. Sc.* **3**, 323–333.

[17] Tijms, H.C. (2003). *A First Course in Stochastic Models.* Wiley, England.

[18] De Veciana, G., Lee, T.-L., Konstantopoulos, T. (2001). Stability and performance analysis of networks supporting elastic services. *IEEE/ACM Trans. Netw.* **9**, 2–14.

[19] Verloop, I.M., Borst, S.C., Núñez-Queija, R. (2006). Delay optimization in bandwidth-sharing networks. In: *Proc. CISS 2006.*

[20] Verloop, I.M., Borst, S.C., Núñez-Queija, R. (2005). Stability of size-based scheduling disciplines in resource-sharing networks. In: *Proc. Performance 2005 Conf.*, Juan-les-Pins, France, 247–262.

[21] Yang, S., De Veciana, G. (2002). Size-based adaptive bandwidth allocation: optimizing the average QoS for elastic flows. In: *Proc. IEEE Infocom 2002*, New York NY, USA, 657–666.

[22] Yang, S., De Veciana, G. (2004). Enhancing both network and user performance for networks supporting best-effort traffic. *IEEE/ACM Trans. Netw.* **12**, 349–360.

Appendix A: Proof of Lemma 5.6

Assume $\delta < 1 - \rho_j$, for all $j = 1, \ldots, L$, and $\pi^{**}_{\boldsymbol{m}} \in \Pi^{**}_{\boldsymbol{m}}$. Let $W^{\pi^{**}_{\boldsymbol{m}}}_{0,<m_0}(0) = W^{\pi^{**}_{\boldsymbol{m}}}_{i,<m_i}(0) = 0$, for a certain $i \in \{1, \ldots, L\}$. We will prove that at time $t \geq 0$, there is a $j^* \in \{1, \ldots, L\}$, such that

$$W^{\pi^{**}_{\boldsymbol{m}}}_{0,<m_0}(t) + W^{\pi^{**}_{\boldsymbol{m}}}_{i,<m_i}(t) \leq \hat{W}^{\rho_0-\delta}_{0,<m_0}(t) + \hat{W}^{\rho_{i,m_i}-\delta}_{i,<m_i}(t) + \hat{W}^{\rho_{j^*}+\delta}_{j^*}(t) + U^{\rho_{j^*}-\delta}_{j^*}(t). \tag{13}$$

For convenience, we will assume that among class-i users of size smaller than m_i service is non-preemptive, although this is not essential in any way for the assertion to hold.
Define $s_1 := \sup\{s \leq t : W^{\pi^{**}_{\boldsymbol{m}}}_{0,<m_0}(s) + \min(W^{\pi^{**}_{\boldsymbol{m}}}_1(s), \ldots, W^{\pi^{**}_{\boldsymbol{m}}}_L(s)) = 0\}$ and $s_2 := \sup\{s \leq t : W^{\pi^{**}_{\boldsymbol{m}}}_{i,<m_i}(s) = 0\}$. Note that $W^{\pi^{**}_{\boldsymbol{m}}}_{0,<m_0}(s_1) = 0$, $W^{\pi^{**}_{\boldsymbol{m}}}_{i,<m_i}(s_2) = 0$, and $W^{\pi^{**}_{\boldsymbol{m}}}_{j^*}(s_1) = 0$ for some $j^* \in \{1, \ldots, L\}$. Denote by $B_i(s,t)$ the total amount of service given to class-i users during the time interval $[s,t]$, and denote by $B_{i,<m_i}(s,t)$ the portion of service that is given to class-i users of size smaller than m_i. Then,

$$W^{\pi^{**}_{\boldsymbol{m}}}_{i,<m_i}(t) = W^{\pi^{**}_{\boldsymbol{m}}}_{i,<m_i}(s_2) + A_{i,<m_i}(s_2,t) - B_{i,<m_i}(s_2,t) = A_{i,<m_i}(s_2,t) - B_{i,<m_i}(s_2,t),$$

and

$$B_{0,<m_0}(s,t) + B_{i,<m_i}(s,t) = t - s, \tag{14}$$

with $s := \max(s_1, s_2)$.
We distinguish between two cases: $s_1 \leq s_2$ and $s_1 \geq s_2$. If $s_1 \leq s_2$, then from (14) we obtain

$$B_{0,<m_0}(s_2,t) + B_{i,<m_i}(s_2,t) = t - s_2.$$

By definition,

$$W_{0,<m_0}^{\pi_{\boldsymbol{m}}^{**}}(t) = W_{0,<m_0}^{\pi_{\boldsymbol{m}}^{**}}(s_2) + A_{0,<m_0}(s_2,t) - B_{0,<m_0}(s_2,t)$$

and

$$W_{0,<m_0}^{\pi_{\boldsymbol{m}}^{**}}(s_2) = W_{0,<m_0}^{\pi_{\boldsymbol{m}}^{**}}(s_1) + A_{0,<m_0}(s_1,s_2) - B_{0,<m_0}(s_1,s_2) = A_{0,<m_0}(s_1,s_2) - B_{0,<m_0}(s_1,s_2).$$

In the interval (s_1,t) there is continuously work present of class-0 users of size smaller than m_0 and work of class-j users, $j=1,\dots,L$. Therefore, under policy $\pi_{\boldsymbol{m}}^{**}$, for all $j \in \{1,\dots,L\}$,

$$B_{0,<m_0}(s_1,s_2) + B_j(s_1,s_2) = s_2 - s_1.$$

Furthermore,

$$B_{j^*}(s_1,s_2) = W_{j^*}^{\pi_{\boldsymbol{m}}^{**}}(s_1) + A_{j^*}(s_1,s_2) - W_{j^*}^{\pi_{\boldsymbol{m}}^{**}}(s_2) = A_{j^*}(s_1,s_2) - W_{j^*}^{\pi_{\boldsymbol{m}}^{**}}(s_2).$$

Combining the above equations, we obtain

$$\begin{aligned}
&W_{0,<m_0}^{\pi_{\boldsymbol{m}}^{**}}(t) + W_{i,<m_i}^{\pi_{\boldsymbol{m}}^{**}}(t)\\
&= A_{0,<m_0}(s_1,t) + A_{i,<m_i}(s_2,t) + A_{j^*}(s_1,s_2) - (t-s_1) - W_{j^*}^{\pi_{\boldsymbol{m}}^{**}}(s_2)\\
&\le A_{0,<m_0}(s_1,t) + A_{i,<m_i}(s_2,t) + A_{j^*}(s_1,s_2) - (t-s_1)\\
&\le A_{0,<m_0}(s_1,t) + A_{i,<m_i}(s_2,t) + A_{j^*}(s_1,t) - A_{j^*}(s_2,t)\\
&\quad -(\rho_0-\delta)(t-s_1) - (\rho_i+\delta)(t-s_1) - (\rho_i-\delta)(t-s_2) + (\rho_i-\delta)(t-s_2)\\
&\le \sup_{s\le t}\{A_{0,<m_0}(s,t) - (\rho_0-\delta)(t-s)\} + \sup_{s\le t}\{A_{i,<m_i}(s,t) - (\rho_i-\delta)(t-s)\}\\
&\quad + \sup_{s\le t}\{A_{j^*}(s,t) - (\rho_i+\delta)(t-s)\} + \sup_{s\le t}\{(\rho_i-\delta)(t-s) - A_{j^*}(s,t)\},
\end{aligned}$$

yielding (13).
Now assume $s_1 \ge s_2$. From (14) we obtain

$$B_{0,<m_0}(s_1,t) + B_{i,<m_i}(s_1,t) = t - s_1.$$

Furthermore,

$$W_{0,<m_0}^{\pi_{\boldsymbol{m}}^{**}}(t) = W_{0,<m_0}^{\pi_{\boldsymbol{m}}^{**}}(s_1) + A_{0,<m_0}(s_1,t) - B_{0,<m_0}(s_1,t) = A_{0,<m_0}(s_1,t) - B_{0,<m_0}(s_1,t)$$

and

$$B_{j^*}(s_2,s_1) = W_{j^*}^{\pi_{\boldsymbol{m}}^{**}}(s_2) + A_{j^*}(s_2,s_1) - W_{j^*}^{\pi_{\boldsymbol{m}}^{**}}(s_1) \ge A_{j^*}(s_2,s_1).$$

There is continuously work present of class-i users of size smaller than m_i in the interval (s_2,t). Hence, for all $j \in \{1,\dots,L\}$,

$$B_{i,<m_i}(s_2,s_1) \ge B_j(s_2,s_1).$$

Combining the above equations, we obtain

$$\begin{aligned}
&W_{0,<m_0}^{\pi_{\boldsymbol{m}}^{**}}(t) + W_{i,<m_i}^{\pi_{\boldsymbol{m}}^{**}}(t)\\
&= A_{0,<m_0}(s_1,t) + A_{i,<m_i}(s_2,t) - B_{i,<m_i}(s_2,s_1) - (t-s_1)\\
&\le A_{0,<m_0}(s_1,t) + A_{i,<m_i}(s_2,t) - A_{j^*}(s_2,s_1) - (t-s_1)\\
&\le A_{0,<m_0}(s_1,t) + A_{i,<m_i}(s_2,t) + A_{j^*}(s_1,t) - A_{j^*}(s_2,t)\\
&\quad -(\rho_0-\delta)(t-s_1) - (\rho_i+\delta)(t-s_1) + (\rho_i-\delta)(t-s_2) - (\rho_i-\delta)(t-s_2)\\
&\le \sup_{s\le t}\{A_{0,<m_0}(s,t) - (\rho_0-\delta)(t-s)\} + \sup_{s\le t}\{A_{i,<m_i}(s,t) - (\rho_i-\delta)(t-s)\}\\
&\quad + \sup_{s\le t}\{A_{j^*}(s,t) - (\rho_i+\delta)(t-s)\} + \sup_{s\le t}\{(\rho_i-\delta)(t-s) - A_{j^*}(s,t)\},
\end{aligned}$$

which again yields (13). This concludes the proof. □

Appendix B: Proof of Proposition 5.10

Take $\pi^p \in \Pi^{**}_{\boldsymbol{m}} \cap \Pi^{**}$. In Lemma 3.1 it is proved that for every policy $\pi \in \Pi$, there are at time t classes $j \neq k \in \{1, \ldots, L\}$, such that

$$W_0^{\pi^p}(t) + W_j^{\pi^p}(t) + W_k^{\pi^p}(t) \leq W_0^{\pi}(t) + W_j^{\pi}(t) + W_k^{\pi}(t). \tag{15}$$

Furthermore, π^p is work-conserving in all nodes. Therefore,

$$W_0^{\pi^p}(t) + W_i^{\pi^p}(t) \leq W_0^{\pi}(t) + W_i^{\pi}(t). \tag{16}$$

Multiplying (15) by $\sum_{i=1}^{L} \frac{1}{M_i} - \frac{1}{M_0} \geq 0$ and (16) by $\frac{1}{M_0} - \sum_{l=1, l\neq i} \frac{1}{M_l} \geq 0$ for $i = j, k$ and by $\frac{1}{M_i}$ for all $i = 1 \ldots L$ with $i \neq j, k$, and summing these $L+1$ inequalities results in $\sum_{i=0}^{L} \frac{1}{M_i} W_i^{\pi^p}(t) \leq \sum_{i=0}^{L} \frac{1}{M_i} W_i^{\pi}(t)$, hence

$$\sum_{i=0}^{L} \frac{1}{M_i} \mathbb{E}(W_i^{\pi^p}) \leq \sum_{i=0}^{L} \frac{1}{M_i} \mathbb{E}(W_i^{\pi}). \tag{17}$$

We now extend this result to policies in $\Pi^{**}_{\boldsymbol{m}}$ by analyzing the scaled workloads. Note that under both policy $\pi^{**}_{\boldsymbol{m}} \in \Pi^{**}_{\boldsymbol{m}}$ and policy π^p every node operates in a work-conserving manner with respect to work consisting of class-0 users smaller than m_0 and class-i users, which implies

$$W_{0,<m_0}^{\pi^{**}_{\boldsymbol{m}}}(t) + W_i^{\pi^{**}_{\boldsymbol{m}}}(t) = W_{0,<m_0}^{\pi^p}(t) + W_i^{\pi^p}(t).$$

Since policies $\pi^{**}_{\boldsymbol{m}}$ and π^p are work-conserving in all nodes with respect to the total amount of work, it follows that

$$W_{0,\geq m_0}^{\pi^{**}_{\boldsymbol{m}}}(t) = W_{0,\geq m_0}^{\pi^p}(t). \tag{18}$$

Observing that $\pi^{**}_{\boldsymbol{m}}, \pi^p \in \Pi^{**}_{\boldsymbol{m}}$, we obtain from Proposition 5.7 that $\lim_{\lambda\uparrow\lambda^*}(1-\rho)(\mathbb{E}(W^{\pi}_{0,<m_0}) + \mathbb{E}(W^{\pi}_{i,<m_i})) = 0$ for $\pi \in \{\pi^{**}_{\boldsymbol{m}}, \pi^p\}$. Together with (18) and the fact that $\pi^{**}_{\boldsymbol{m}}$ and π^p are work-conserving, this implies that for $i = 1, \ldots, L$,

$$\lim_{\lambda\uparrow\lambda^*} (1-\rho)\mathbb{E}(W_{i,\geq m_i}^{\pi^{**}_{\boldsymbol{m}}}) = \lim_{\lambda\uparrow\lambda^*} (1-\rho)\mathbb{E}(W_{i,\geq m_i}^{\pi^p}). \tag{19}$$

Using (18) and (19) and the fact that $\lim_{\lambda\uparrow\lambda^*}(1-\rho)\Big(\mathbb{E}(W_{0,<m_0}^{\pi^{**}_{\boldsymbol{m}}}) + \mathbb{E}(W_{i,<m_i}^{\pi^{**}_{\boldsymbol{m}}})\Big) = 0$, we obtain from (17) that

$$\lim_{\lambda\uparrow\lambda^*} (1-\rho) \sum_{i=0}^{L} \frac{1}{M_i} \mathbb{E}(W_i^{\pi^{**}_{\boldsymbol{m}}}) \leq \lim_{\lambda\uparrow\lambda^*} (1-\rho) \sum_{i=0}^{L} \frac{1}{M_i} \mathbb{E}(W_i^{\pi}). \tag{20}$$

Class-i users of size larger than m_i are served in a non-preemptive way, which means $(\mathbb{E}(N_{i,\geq m_i}^{\pi^{**}_{\boldsymbol{m}}}) - 1)m_i \leq \mathbb{E}(W_{i,\geq m_i}^{\pi^{**}_{\boldsymbol{m}}})$. Proposition 5.7 shows that under policy $\pi^{**}_{\boldsymbol{m}}$, all scaled class-$i$ work is composed of users of size m_i or larger, $i = 0, \ldots, L$, hence $\lim_{\lambda\uparrow\lambda^*}(1-\rho)\mathbb{E}(N_i^{\pi^{**}_{\boldsymbol{m}}})m_i \leq \lim_{\lambda\uparrow\lambda^*}(1-\rho)\mathbb{E}(W_i^{\pi^{**}_{\boldsymbol{m}}})$. Noting that $\mathbb{E}(W_i^{\pi}) \leq \mathbb{E}(N_i^{\pi})M_i$ and substituting into (20), we obtain

$$\lim_{\boldsymbol{m}\uparrow\boldsymbol{M}} \lim_{\lambda\uparrow\lambda^*} (1-\rho)(\sum_{i=0}^{L} \mathbb{E}(N_i^{\pi^{**}_{\boldsymbol{m}}})) \leq \lim_{\lambda\uparrow\lambda^*} (1-\rho)(\sum_{i=0}^{L} \mathbb{E}(N_i^{\pi})),$$

which concludes the proof. □