Performance of TCP-Friendly Streaming Sessions in the Presence of Heavy-Tailed Elastic Flows*

René Bekker[†,⋆], Sem Borst[†,⋆,‡], Rudesindo Núñez-Queija[†,⋆]

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

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

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

Abstract

We consider a fixed number of streaming sessions which share a bottleneck link with a dynamic population of elastic flows. Motivated by extensive measurement studies, we assume that the sizes of the elastic flows exhibit heavy-tailed characteristics. The elastic flows are TCP-controlled, while the transmission rates of the streaming applications are governed by a so-called TCP-friendly rate control protocol. TCP-friendly rate control protocols provide a promising mechanism for avoiding severe fluctuations in the transmission rate, while ensuring fairness with competing TCP-controlled flows.

Adopting the Processor-Sharing (PS) discipline to model the bandwidth sharing, we investigate the asymptotic tail distribution of the deficit in service received by the streaming sessions compared to a nominal service target. The latter metric provides an indication for the quality experienced by the streaming applications. The results yield valuable qualitative insight into the occurrence of persistent quality disruption for the streaming users. We also examine the delay performance of the elastic flows by exploiting a useful relationship with a Processor-Sharing queue with permanent customers.

1 Introduction

Over the past decade, TCP has gained ubiquity as the predominant congestion control mechanism in the Internet. While TCP is adequate for best-effort elastic traffic, such as file transfers and Web browsing sessions, it is less suitable for supporting delay-sensitive streaming applications. In particular, the inherent fluctuations in the window size adversely impact the user-perceived quality of real-time streaming applications. As a potential alternative, UDP could be used to avoid the wild oscillations in the transmission rate.

*This work was financially supported by a grant from Philips Research and by the Dutch Ministry of Economic Affairs (via its agency SENTER; project EQUANET)

Since UDP does not respond to congestion, it may cause severe packet losses however, and give rise to unfairness in the competition for bandwidth with TCP-controlled flows.

Discriminatory packet scheduling mechanisms provide a further alternative to achieve some form of prioritization of streaming applications. However, the implementation of scheduling mechanisms is surrounded with substantial controversy, because it entails major complexity and scalability issues. In addition, prioritization of streaming applications may cause performance degradation and even starvation of TCP-controlled flows that back off in response to congestion. Evidently, the latter issue gains importance as the amount of streaming traffic in the Internet grows.

The above considerations have motivated an interest in *TCP-friendly* or *equation-based* rate control protocols for streaming applications [16, 25, 27]. The key goal is to eliminate severe fluctuations in the window size and adjust the transmission rate in a smoother manner. In order to ensure fairness with competing TCP-controlled flows, the specific aim is to set the transmission rate to the 'fair' bandwidth share, i.e., the throughput that a long-lived TCP flow would receive under similar conditions.

Various methods have been proposed for determining the fair bandwidth share in an accurate and robust manner. Typical methods involve measuring the packet loss rate and round-trip delay (e.g. by running a low-rate connection to probe the network conditions). The corresponding throughput may then be estimated from well-established equations that express the throughput of a TCP-controlled flow in terms of the packet loss rate and round-trip delay, see for instance [21, 24]. Obviously, the adaptation mechanism faces the usual trade-off between responsiveness and smoothness, which is exacerbated by the fact that the estimation procedure relies on intrinsically noisy measurements.

In the present paper we explore the performance of streaming applications under such TCP-friendly rate control protocols. We consider a fixed number of streaming sessions which share a bottleneck link with a dynamic population of elastic flows. The assumption of persistent streaming users is motivated by the separation of time scales between the typical duration of streaming sessions (minutes to hours) and that of the majority of elastic flows (seconds to minutes). We assume that the sizes of the elastic flows exhibit heavy-tailed characteristics. The latter assumption is based on extensive measurement studies which show that file sizes in the Internet, and hence the volumes of elastic transfers, commonly have heavy-tailed features, see for instance [14].

As mentioned above, the design and implementation of TCP-friendly mechanisms is a significant challenge. In the present paper we leave implementation issues aside though, and investigate the performance under idealizing assumptions. Specifically, we assume the rate control mechanism reacts instantly and perfectly accurately to changes in the population of elastic flows, and maintains a constant rate otherwise. This results – at the flow level – in a fair sharing of the link rate in a Processor-Sharing (PS) manner. The PS discipline has emerged as a useful paradigm for modeling the bandwidth sharing among dynamically competing TCP flows, see for instance [4, 20, 22]. Although the PS paradigm may not be entirely justified for short flows, inspection of the proofs suggests that this assumption is actually not that crucial for the asymptotic results to hold. The effect of oscillations, inaccuracies and delays in the estimation procedure on the performance remains as a subject for future research.

We consider the probability that a possible deficit in service received by the streaming sessions compared to a nominal service target exceeds a certain threshold. The latter

probability provides a measure for the degree of disruption in the quality experienced by the streaming users. We determine the asymptotic behavior of the service deficit (or *workload*) probability for a large value of the threshold. The results yield useful qualitative insight into the occurrence of persistent quality disruption for the streaming users. We furthermore examine the delay performance of the elastic flows.
In [19], the authors consider a mixture of elastic transfers and streaming users sharing the network bandwidth according to weighted α-fair rate algorithms. Weighted α-fair allocations include various common fairness notions, such as max-min fairness and proportional fairness, as special cases. They also provide a tractable theoretical abstraction of the throughput allocations under decentralized feedback-based congestion control mechanisms such as TCP, and in particular cover TCP-friendly rate control protocols. In a recent paper [7], the authors derive various performance bounds for a related model with a combination of elastic flows and streaming traffic sharing the link bandwidth in a fair manner. The latter papers however focus on different performance metrics.
The remainder of the paper is organized as follows. In Section 2 we present a detailed model description. In Section 3 we analyze the delay and workload performance of the elastic flows by exploiting a useful relationship with a M/G/1 PS model with permanent customers. The main result is presented in Section 4, where we consider the workload asymptotics of the streaming users for the case of constant-rate traffic. Besides a heuristic interpretation of the result, we also give some preliminaries and an outline of the proof, which involves lower and upper bounds that asymptotically coincide. The proofs of the lower and upper bounds may be found in Sections 5 and 6, respectively. We extend the results to the case of variable-rate streaming traffic in Section 7. In addition, we consider the tail asymptotics of the simultaneous workload distribution of the K individual streaming users. In Section 8 we make some concluding remarks.

2 Model description

We consider two traffic classes sharing a link of unit rate. Class 1 consists of a static population of $K \geq 1$ statistically identical streaming sessions. These sessions stay in the system indefinitely. Class 2 consists of a dynamic configuration of elastic flows. These users arrive according to a renewal process with mean interarrival time $1/\lambda$, and have service requirements with distribution $B(\cdot)$ and mean $\beta < \infty$.
The elastic flows are TCP-controlled, while the transmission rates of the streaming sessions are adapted in a TCP-friendly fashion. Abstracting from packet-level details, we assume that this results in a fair sharing of the link rate according to the PS discipline. Thus, when there are $N(u)$ elastic flows in the system at time u, the available service rate for each of the users – either elastic or streaming – is $1/(K + N(u))$. Denote by $C_1(u) := K/(K + N(u))$ the total available service rate for the streaming traffic at time u. Define $C_1(s,t) := \int_{u=s}^{t} C_1(u)\mathrm{d}u$ as the total amount of service available for the streaming sessions during the time interval $[s,t]$.
In the present paper, we will mainly be interested in the quantity $V_1(t) := \sup_{s\leq t}\{A_1(s,t) - C_1(s,t)\}$, where $A_1(s,t)$ denotes the amount of service which ideally should be available for the streaming traffic during the interval $[s,t]$. For example, $A_1(s,t)$ may be taken as the amount of streaming traffic that would nominally be generated during the interval

$[s,t]$ if there were ample bandwidth. Thus, $V_1(t)$ may be interpreted as the shortfall in service for the streaming traffic at time t compared to what should have been available in ideal circumstances. For conciseness, we will henceforth refer to $V_1(t)$ as the *workload* of the streaming traffic at time t. Throughout the paper, we also often refer to $A_1(s,t)$ as the amount of streaming traffic generated. It is worth emphasizing though that $A_1(s,t)$ represents just the amount of traffic which ideally should have been served, and not the amount of traffic that is actually generated, which is primarily governed by the fair service rates as described above. Thus, $V_1(t)$ provides just a virtual measure of a service deficit compared to an ideal environment, and by no means corresponds to the backlog or buffer content in an actual system.

In Sections 3–6 we will focus on the 'constant-rate' case $A_1(s,t) \equiv Kr(t-s)$, which amounts to a fixed target service rate r per streaming session. We will extend the analysis in Section 7 to the 'variable-rate' case where $A_1(s,t)$ is a general stochastic process with stationary increments.

We will also consider the quantity $V_2(t) := \sup_{s\leq t}\{A_2(s,t) - C_2(s,t)\}$, where $A_2(s,t)$ denotes the amount of elastic traffic generated during the time interval $[s,t]$, and $C_2(s,t)$ represents the amount of service available for the elastic flows during $[s,t]$. By definition, $C_2(s,t) := \int_{u=s}^{t} C_2(u)\mathrm{d}u$, with $C_2(u)$ denoting the total available service rate for the elastic traffic at time u. Evidently, $C_2(u) \geq 1 - C_1(u)$, with equality in case the streaming sessions always claim the full service rate available. For the elastic traffic, the latter case is equivalent to a G/G/1 PS queue with K permanent customers, accounting for the presence of the competing streaming sessions.

However, we allow for possible strict inequality in case the streaming sessions do not always consume the full service rate available, and the unused surplus is granted to the elastic class, i.e., $C_2(s,t) = t - s - B_1(s,t)$, with $B_i(s,t) \leq C_i(s,t)$ denoting the actual amount of service received by class i, $i = 1,2$, during the interval $[s,t]$. For example, when the 'workload' of the streaming sessions is zero, the actual service rate may be set to the minimum of the aggregate input rate and the total service rate available. In particular, in the 'constant-rate' case the actual service rate per streaming session at time u is then only $\min\{r, 1/(K+N(u))\}$ when $V_1(u) = 0$. Note that the total service rate is thus used at time u as long as $V_1(u)+V_2(u) > 0$, which implies that $V_1(t)+V_2(t) = \sup_{s\leq t}\{A_1(s,t)+A_2(s,t)-(t-s)\}$. Hence, the case $C_2(s,t) = t-s-B_1(s,t)$ will be termed the *work-conserving* scenario, whereas the case $C_2(u) = 1 - C_1(u) = N(u)/(K+N(u))$ will be referred to as the *permanent-customer* scenario. It may be checked that the work-conserving and permanent-customer scenarios provide lower and upper bounds for the general case with $t - s - C_1(s,t) \leq C_2(s,t) \leq t - s - B_1(s,t)$.

Define $\rho := \lambda\beta$ as the traffic intensity of class 2. Without proof, we claim that $\rho < 1$ is a necessary and sufficient condition for class 2 to be stable. While the former is obvious, the latter may be concluded from the comparison with the G/G/1 PS queue with K permanent customers mentioned above (see [5] for the case of Poisson arrivals). For class 1 to be stable as well, we need to assume that $\rho + Kr < 1$, with $\mathbb{E}\{A(0,1)\} = Kr$. Here class 1 is said to be stable if the 'workload' $V_1(t)$ converges to a finite random variable as $t \to \infty$. Denote by V_i a random variable with the steady-state distribution of $V_i(t)$, $i = 1,2$. In Sections 4–7, we additionally assume that $(K+1)r > 1 - \rho$, which implies that the system is critically loaded in the sense that one extra streaming session – or a 'persistent' elastic flow – would cause instability. Combined, the above two assumptions give $Kr < 1 - \rho < (K+1)r$. We finally introduce some additional notation. Let B be a random variable distributed as the generic service requirement of an elastic user, and let B^r be a random variable distributed as the residual lifetime of B, i.e., $B^r(x) = \mathbb{P}\{B^r < x\} = \frac{1}{\beta}\int_0^x (1 - B(y))\mathrm{d}y$. We assume that the service requirement distribution is regularly varying of index $-\nu$ (denoted as $B(\cdot) \in \mathcal{R}_{-\nu}$), i.e., $1 - B(x) \sim L(x)x^{-\nu}, \nu > 1$, with $L(x)$ some slowly varying function. Here, and throughout the paper, we use the notation $f(x) \sim g(x)$ to indicate that $f(x)/g(x) \to 1$ as $x \to \infty$. (A function $L(\cdot)$ is called slowly varying if $L(\eta x) \sim L(x)$ for all $\eta > 1$.) It follows from Karamata's Theorem [6, Theorem. 5.1.11] that $x\mathbb{P}\{B > x\} \sim (\nu - 1)\beta\,\mathbb{P}\{B^r > x\}$, so that $B^r(\cdot) \in \mathcal{R}_{1-\nu}$.

Remark 2.1. The analysis may be generalized to the case of *Discriminatory* Processor Sharing (DPS), that is, when the rate share per streaming session is $w/(wK + N(u))$ rather than $1/(K + N(u))$, for some positive weight factor w. In the 'constant-rate' case, it is in fact easily verified that the workload for K streaming sessions each with weight w and target rate r is equivalent to that in a model with $K' = wK$ streaming sessions each with unit weight and target rate $r' = r/w$ (with some abuse of terminology when wK is not integer). For notational transparency, we henceforth focus on the case $w = 1$. ◇

3 Delay performance of the elastic flows

As mentioned earlier, our model shows strong resemblance with a G/G/1 PS queue with K permanent customers [5]. The permanent customers play the role of the persistent streaming users in our model, while the regular (non-permanent) customers correspond to the elastic flows, inheriting the same arrival process and service requirement distribution $B(\cdot)$. It may be checked that the service rate available for the elastic class in our model is always at least that in the model with K permanent customers. Hence, the number of elastic flows, their individual residual service requirements, their respective delays (sojourn times), and the workload of the elastic class are stochastically dominated by the corresponding quantities in the model with permanent customers. This may be formally shown using similar arguments as in the proof of Lemma 4 in [8]. The stochastic ordering between the two models is particularly useful, since it provides upper bounds for several performance measures of interest in our model in terms of the model with permanent customers. In order for the bounds to be analytically tractable, we assume in the remainder of the section that the elastic flows arrive according to a Poisson process of rate λ.

Remark 3.1. As noted earlier, in the special case where the service rate of the elastic class is always $C_2(t) \equiv \frac{N(t)}{K+N(t)}$ (which we named the *permanent-customer* scenario), the two models are actually equivalent in terms of the number of elastic users and their respective residual service requirements. In that case, the inequalities in Equations (2)-(6) below hold with equality. ◇

The M/G/1 PS queue with permanent customers is a special case of the model studied in [13], where each customer receives service at rate $f(n)$, $0 \le f(n) < \infty$, when there are n customers. To obtain the model with K permanent customers, we take $f(n) = \frac{1}{K+n}$. Let $N_{(K)}$ be the number of regular customers in the model with K permanent customers and, given $N_{(K)} = n$, let $\hat{B}_1, \ldots, \hat{B}_n$ be their residual service requirements. Then, according to [13],

$$\mathbb{P}\left\{N_{(K)} = n; \hat{B}_1 > x_1; \ldots ; \hat{B}_n > x_n\right\} = (1-\rho)^{K+1}\rho^n \binom{n+K}{n} \prod_{m=1}^{n} \mathbb{P}\left\{B^r > x_m\right\}. \quad (1)$$

(When $w \neq 1$ and wK is not integer, the above formula remains valid upon substituting wK for K and replacing the factorial function in the binomial coefficients by the Gamma function.) We thus obtain an upper bound for the probability that the service rate of the streaming users is below a given desired rate s:

$$\mathbb{P}\left\{\frac{1}{K+N} < s\right\} \leq \mathbb{P}\left\{N_{(K)} > \lfloor 1/s - K \rfloor\right\} = \sum_{j=0}^{K} \binom{\lfloor 1/s \rfloor + 1}{j} (1-\rho)^j \rho^{\lfloor 1/s \rfloor + 1 - j}. \quad (2)$$

As mentioned above, the delay (sojourn time) of elastic users in our model (denoted by S_2) is stochastically dominated by the corresponding quantity in the model with permanent customers. In the M/G/1 PS queue with m permanent customers, let $S_{(m)}$ be the delay and $S_{(m)}(x)$ be the conditional sojourn time *given* that the service requirement of the customer is x. It is known that this random variable is the $(m+1)$-fold convolution of the distribution of $S_{\mathrm{PS}}(x)$, the conditional sojourn time in the standard M/G/1 PS queue [5]:

$$\mathbb{P}\left\{S_{(m)}(x) \leq t\right\} = \mathbb{P}\left\{\sum_{j=1}^{m+1} S_{\mathrm{PS},j}(x) \leq t\right\},$$

where $S_{\mathrm{PS},j}$, $j = 1, \ldots, m+1$, represent i.i.d. copies of S_{PS}. (It is worth emphasizing that the unconditional sojourn time does not allow for a similar decomposition.) In particular, using that $\mathbb{E}S_{\mathrm{PS}}(x) = \frac{x}{1-\rho}$, we obtain an upper bound for the conditional mean delay of elastic users in our model (denoted as $S_2(x)$):

$$\mathbb{E}S_2(x) \leq \mathbb{E}S_{(K)}(x) = (K+1)\frac{x}{1-\rho}, \quad (3)$$

and, hence, the (unconditional) mean delay satisfies

$$\mathbb{E}S_2 \leq (K+1)\frac{\beta}{1-\rho}. \quad (4)$$

We now turn to the tail asymptotics for the unconditional sojourn time. The next proposition shows that the exact asymptotics of S_2 depend on the assumptions on $C_2(s,t)$ in case $B_1(s,t) < C_1(s,t)$. As observed in Remark 3.1, in case $C_2(t) \equiv \frac{N(t)}{K+N(t)}$, the model is equivalent to the M/G/1 PS queue with K permanent customers. Asymptotically, the equivalence also continues to hold when the system is critically loaded, i.e., $(K+1)r > 1-\rho$, which implies that class-1 users will be rarely non-backlogged over the course of a long sojourn time. However, the sojourn time asymptotics change when the system is below critical load, i.e., $(K+1)r < 1-\rho$, and the elastic flows receive (part of) the capacity left over by the streaming users, i.e., $C_2(t) > \frac{N(t)}{K+N(t)}$.

Proposition 3.1. *If $B(\cdot) \in \mathcal{R}_{-\nu}$ and $(K+1)r > 1-\rho$ or $C_2(t) \equiv \frac{N(t)}{K+N(t)}$, or both, then*

$$\mathbb{P}\left\{S_2 > x\right\} \sim \mathbb{P}\left\{S_{(K)} > x\right\} \sim \mathbb{P}\left\{B > \frac{(1-\rho)x}{K+1}\right\}.$$

In contrast, if $(K+1)r < 1-\rho$ and $C_2(s,t) \equiv t-s-B_1(s,t)$, then

$$\mathbb{P}\left\{S_2 > x\right\} \sim \mathbb{P}\left\{B > (1-\rho-Kr)x\right\}.$$

Proof. The asymptotics for $S_{(K)}$ (and, thus, for S_2 in the permanent-customer scenario) follow from [17]. As noted above, the service rate of a customer is $f(n) = \frac{1}{K+n}$ when there are n non-permanent customers in the system. We can therefore apply [17, Theorem 3] to obtain $\gamma^f = \frac{1-\rho}{K+1}$ and the desired result follows.
For the remainder of the proof we only provide an intuitive sketch. (We refer to Appendix C for a detailed proof.) In both cases, a large delay of an elastic flow is due to a large service requirement of the flow itself, and the ratio between the two quantities is simply the average service rate received by the large flow. Over the duration of the large flow, the other elastic flows continue to produce traffic at an average rate ρ, and also receive service at an average rate ρ. The remaining service capacity is shared among the large elastic flow and the streaming users, each entitled to a fair share $(1-\rho)/(K+1)$. In case $(K+1)r > 1-\rho$, the fair share of the streaming users is below their average input rate r. Thus, the streaming users will be almost constantly backlogged, and the average service rate for the large elastic flow is just $(1-\rho)/(K+1)$. In case $(K+1)r < 1-\rho$, the fair share of the streaming users exceeds their average 'input rate' r. Hence, the streaming users will only claim an average service rate Kr, and the average service rate left for the large elastic flow is $1-\rho-Kr$ now. □

In case the system is not critically loaded and $t-s-C_1(s,t) < C_2(s,t) < t-s-B_1(s,t)$ for at least some s and t, we obtain the immediate bound

$$\mathbb{P}\left\{S_2 > x\right\} \leq (1+o(1))\mathbb{P}\left\{B > \frac{(1-\rho)x}{K+1}\right\}, \qquad \text{as } x \to \infty. \tag{5}$$

Remark 3.2. The result for the *permanent customer* scenario is formulated for regularly varying service requirements, but it may readily be extended (following the proof of [23, Theorem 4.1]) to the slightly larger class of *intermediately* regularly varying distributions. ◇

Finally, we turn to the workload of the elastic class which is also stochastically dominated by the corresponding quantity in the model with permanent customers. Again, we first state a result for the M/G/1 PS queue with permanent customers.

Proposition 3.2. *If $B(\cdot) \in \mathcal{R}_{-\nu}$, then $V_{(m)}$, the workload in the M/G/1 PS queue with m permanent customers, satisfies*

$$\mathbb{P}\left\{V_{(m)} > x\right\} \sim \mathbb{E}N_{(m)}\,\mathbb{P}\left\{B^r > x\right\} = \frac{(m+1)\rho}{1-\rho}\,\mathbb{P}\left\{B^r > x\right\}.$$

Proof. From (1) we observe that, given $N_{(m)} = n$, $\hat{B}_1, \ldots, \hat{B}_n$ are i.i.d. copies of B^r. Using [29] together with $V_{(m)} = \sum_{i=1}^{N_{(m)}} \hat{B}_i$, and the fact that $\mathbb{P}\left\{N > n\right\}$ decays geometrically fast when $n \to \infty$, we obtain the desired equivalence. □

As an immediate corollary, we derive

$$\mathbb{P}\{V_2 > x\} \le (1+o(1))\frac{(K+1)\rho}{1-\rho}\,\mathbb{P}\{B^r > x\}, \qquad \text{as } x \to \infty. \tag{6}$$

4 Workload asymptotics of the streaming traffic

In this section we turn the attention to the workload distribution of class 1. For convenience, we assume that each class-1 source generates traffic at a constant rate r. The latter assumption is however not essential for the asymptotic results to hold, and in Section 7 we extend the results to the case of variable-rate class-1 traffic. In the remainder of the paper, we assume that $\rho + Kr < 1$ to ensure stability. In addition, we impose the condition that $(K+1)r > 1-\rho$, i.e., the system is critically loaded. Thus, $Kr < 1-\rho < (K+1)r$.

The next theorem provides the main result of the paper.

Theorem 4.1. *If $B(\cdot) \in \mathcal{R}_{-\nu}$ and $Kr < 1-\rho < (K+1)r$, then*

$$\mathbb{P}\{V_1 > x\} \sim \frac{\rho}{1-\rho-Kr}\mathbb{P}\left\{B^r > \frac{x\frac{1-\rho}{K+1}}{K(r-\frac{1-\rho}{K+1})}\right\}. \tag{7}$$

The proof of the above theorem involves asymptotic lower and upper bounds which will be provided in Sections 5 and 6, respectively. In this section, we sketch a heuristic derivation of the result, which will also serve as an outline for the construction of the lower bound in Subsection 5.1. In addition, we give an intuitive interpretation, which provides the basis for the lower bound in Subsection 5.2 and the upper bound in Section 6. First, however, we give some basic relations between traffic processes, amounts of service and workloads, and state a few preliminary results.

Preliminary results
The amounts of service satisfy the following simple inequality

$$B_1(s,t) + B_2(s,t) \le t - s. \tag{8}$$

For the workloads, the following obvious identity relation holds for $i = 1, 2$ and $s < t$,

$$V_i(t) = V_i(s) + A_i(s,t) - B_i(s,t). \tag{9}$$

As mentioned in Section 2, in the work-conserving scenario, i.e., $C_2(s,t) \equiv t - s - B_1(s,t)$, the system is equivalent in terms of the total workload to a single queue of unit rate fed by the aggregate class-1 and class-2 traffic processes,

$$V_1(t) + V_2(t) = \sup_{s \le t}\{A_1(s,t) + A_2(s,t) - (t-s)\}. \tag{10}$$

In particular, in the constant-rate case,

$$\begin{aligned} V_1(t) + V_2(t) &= \sup_{s \le t}\{Kr(t-s) + A_2(s,t) - (t-s)\} \\ &= \sup_{s \le t}\{A_2(s,t) - (1-Kr)(t-s)\} \\ &= V_2^{1-Kr}(t), \end{aligned} \tag{11}$$

with $V_2^c(t)$ the workload at time t in an isolated queue with service rate c fed by class 2 only. For any $\rho < c$, let V_2^c be its steady-state version. The asymptotic tail distribution of the latter quantity is given by the next theorem, which is originally due to Cohen [12], and has been extended to subexponential distributions by Pakes [26].

Theorem 4.2. *Assume that $\rho < c$. Then, $B(\cdot) \in \mathcal{R}_{-\nu}$ iff $\mathbb{P}\{V_2^c < \cdot\} \in \mathcal{R}_{1-\nu}$, and then*

$$\mathbb{P}\{V_2^c > x\} \sim \frac{\rho}{c-\rho}\mathbb{P}\{B^r > x\}.$$

The same relation holds when V_2^c represents the workload distribution at arrival epochs of class 2.

Relation (11) plays a central role in the proof of Theorem 4.1. In the sequel we will consider several extensions of the basic model, allowing the system to be non-work-conserving (e.g., the *permanent-customer* scenario) and having variable-rate streaming traffic (with mean Kr). In those cases, (11) does not hold as a sample path identity, but (under some assumptions) $V_1 + V_2$ and V_2^{1-Kr} are *asymptotically* equivalent in the following sense (similar reduced-load type of equivalences may be found in, e.g., [1, 18, 30]):

$$\mathbb{P}\{V_1 + V_2 > x\} \sim \mathbb{P}\left\{V_2^{1-Kr} > x\right\}. \tag{12}$$

The main intuitive idea is that a large total workload is most likely due to the arrival of a large class-2 user. Since the system is critically loaded, the class-1 workload builds up in the presence of the large class-2 user, so that the full service capacity is used and the system behaves as if it were work-conserving. The detailed proof of (12) is deferred to Appendix A (Proposition A.1).

Heuristic arguments

In queueing systems with heavy-tailed characteristics, rare events tend to occur as a consequence of a single most-probable cause. We will specifically show that in the present context the most likely way for a large class-1 workload V_1 to occur arises from the arrival of a class-2 user with a large service requirement B_{tag}, while the system shows average behavior otherwise. We will refer to the class-2 user as the "tagged" user.

Define $B_{\text{tag}}(s,t)$ as the amount of service received by the tagged user in $(s,t]$. In addition, denote by $B_2^-(s,t)$ the amount of service received by class-2 users in the time interval $(s,t]$, except for the tagged user. Then (8) may be rewritten as follows

$$B_1(s,t) + B_{\text{tag}}(s,t) + B_2^-(s,t) \leq t - s. \tag{13}$$

Suppose that the tagged user arrives at time $-y - z_0$, with $z_0 = \frac{x}{K(r - \frac{1-\rho}{K+1})}$, $B_{\text{tag}} \geq x + (1 - \rho - Kr)(y + z_0)$, and $y \geq 0$. The amount of class-2 traffic generated during the time interval $[-y - z_0, 0]$ is close to average, i.e., $A_2(-y - z_0, 0) \approx \rho(y + z_0)$. Since class 2 is stable, regardless of the presence of the tagged user, the amount of service received roughly equals the amount of class-2 traffic generated during the time interval $[-y - z_0, 0]$, i.e., $B_2^-(-y - z_0, 0) \approx \rho(y + z_0)$. The cumulative amount of service received by the tagged user up to time 0 is either $B_1(-y - z_0, 0)/K$ or B_{tag}, depending on whether the user is still present at time 0 or not.

Using the inequality (13), the amount of service received by class 1 is approximately

$$\begin{aligned} B_1(-y-z_0,0) &\leq y+z_0-B_{\text{tag}}(-y-z_0,0)-B_2^-(-y-z_0,0) \\ &\approx (1-\rho)(y+z_0)-\min\{B_{\text{tag}},B_1(-y-z_0,0)/K\}. \end{aligned}$$

Thus,

$$\begin{aligned} B_1(-y-z_0,0) &\leq \max\{(1-\rho)(y+z_0)-B_{\text{tag}},\frac{K}{K+1}(1-\rho)(y+z_0)\} \\ &\leq \max\{Kr(y+z_0)-x,\frac{K}{K+1}(1-\rho)(y+z_0)\}. \end{aligned}$$

Using the above inequality and the identity relation (9), the class-1 workload at time 0 is

$$\begin{aligned} V_1(0) &\geq A_1(-y-z_0,0)-B_1(-y-z_0,0) \\ &\geq Kr(y+z_0)-\max\{Kr(y+z_0)-x,\frac{K}{K+1}(1-\rho)(y+z_0)\} \\ &= \min\{x,K(r-\frac{1-\rho}{K+1})(y+z_0)\}\geq\min\{x,K(r-\frac{1-\rho}{K+1})z_0\}=x. \end{aligned}$$

In the case of Poisson arrivals of class 2 we obtain (by integrating with respect to y and neglecting the probability of two or more "large" users)

$$\mathbb{P}\{V_1>x\}\geq\int_{y=0}^{\infty}\lambda\mathbb{P}\left\{B_{\text{tag}}>\frac{1-\rho}{K+1}z_0+(1-\rho-Kr)y\right\}\mathrm{d}y,$$

which agrees with the right-hand side of (7).
Of course, there are alternative scenarios that could potentially lead to a large class-1 workload. Theorem 4.1 thus indirectly indicates that these are extremely unlikely compared to the one described above, as will be rigorously shown in Section 6.
A formal proof based on the above heuristics (in case of renewal arrivals of class 2) may be found in Subsection 5.1. The arrival of a class-2 user with a large service requirement in fact also results in a large total amount of work in the system after its arrival. We will use this alternative interpretation of the dominant scenario in Subsection 5.2 to derive a lower bound in case of renewal class-2 arrivals and in Section 6 to obtain an upper bound. In particular, we will show that the event $V_1(-t_1)+V_2(-t_1)\geq x+(1-\rho-Kr)t_1$, with $t_1:=\frac{x}{K(r-\frac{1-\rho}{K+1})}$, corresponds to the dominant scenario described above. Using Proposition A.1 and Theorem 4.2, we then obtain that the probability of the latter event coincides with the right-hand side of (7).

Finally, note that the dominant scenario crucially depends on the critical load, i.e., $1-\rho<(K+1)r$. Section 8 briefly discusses the case of a non-critically loaded system.

5 Lower bound

In this section we derive asymptotic lower bounds for $\mathbb{P}\{V_1>x\}$ using two different approaches. In Subsection 5.1, we explicitly use the arrival of a class-2 user with a large service requirement (as described in the heuristics in Section 4) as the most likely way for a large class-1 workload to occur. We believe that this approach is especially insightful, as it brings out the typical cause of a large class-1 workload. In Subsection 5.2, we provide a proof based on the alternative characterization of the dominant scenario in Section 4. This approach is consistent with the derivation of the upper bound in Section 6. Moreover, it allows for modifications to include variable-rate class-2 traffic.

5.1 Approach 1

To obtain a lower bound for $\mathbb{P}\{V_1 > x\}$, we start by deriving a sufficient sample-path condition for the event $V_1(0) > x$ to occur (Lemma 5.1). Next, we convert the sample-path statement into a probabilistic lower bound which can be used to determine the asymptotic tail behavior of $\mathbb{P}\{V_1 > x\}$ (Proposition 5.2).

Consider the following three events.

1. $\exists y \geq 0$ such that at time $-t_0$, with $t_0 := \frac{x(1+K\epsilon+K\gamma)}{K(r-\frac{1-\rho+\delta}{K+1})} + y$, a tagged class-2 user arrives with service requirement

$$B_{\text{tag}} \geq \frac{x(1+K\epsilon+K\gamma)}{K(r-\frac{1-\rho+\delta}{K+1})} \frac{1-\rho+\delta}{K+1} + y(1-\rho+\delta-Kr) + (\epsilon+\gamma)x \tag{14}$$

2. For the amount of class-2 traffic arriving in the interval $(-t_0, 0]$ it holds that

$$A_2(-t_0, 0) \geq (\rho-\delta)t_0 - (K+1)\gamma x \tag{15}$$

3. The amount of class-2 work at time 0, except from the tagged user, satisfies

$$V_2^-(0) \leq (K+1)\epsilon x \tag{16}$$

We first prove the next sample-path relation.

Lemma 5.1. *If the events (14)-(16) occur simultaneously with $\delta \leq (K+1)r - (1-\rho)$, then $V_1(0) > x$.*

Proof. We distinguish between two cases: (i) the large tagged user is still present in the system at time 0; and (ii) the tagged user already left before time 0.
First consider case (i) and denote by $B_1^+(s,t)$ the amount of service received by the class-1 users and the large tagged class-2 user together in the interval $(s,t]$. Then, using (8) and (9),

$$\begin{aligned} B_1^+(-t_0, 0) &\leq t_0 - V_2(-t_0) - A_2(-t_0, 0) + V_2^-(0) \\ &\leq t_0 - A_2(-t_0, 0) + V_2^-(0) \\ &\leq (1-\rho+\delta)t_0 + (K+1)(\epsilon+\gamma)x, \end{aligned} \tag{17}$$

where we used (15) and (16) in the third inequality. Because of the PS discipline, we have

$B_1(-t_0,0) \leq \frac{K}{K+1} B_1^+(-t_0,0)$. Combining this with (17) and using (9) yields

$$
\begin{aligned}
V_1(0) &\geq A_1(-t_0,0) - B_1(-t_0,0) \\
&\geq Krt_0 - \frac{K}{K+1}\left[(1-\rho+\delta)t_0 + (K+1)(\epsilon+\gamma)x\right] \\
&= K\left(r - \frac{1-\rho+\delta}{K+1}\right) t_0 - K(\epsilon+\gamma)x \\
&\geq K\left(r - \frac{1-\rho+\delta}{K+1}\right) \frac{x(1+K\epsilon+K\gamma)}{K(r-\frac{1-\rho+\delta}{K+1})} - K(\epsilon+\gamma)x \\
&= x,
\end{aligned}
$$

where we used $\delta \leq (K+1)r - (1-\rho)$ in the fourth step.
Next, consider case (ii). From (8) and (9), we obtain

$$
\begin{aligned}
B_1(-t_0,0) &\leq t_0 - V_2(-t_0) - A_2(-t_0,0) + V_2^-(0) - B_{\mathrm{tag}} \\
&\leq t_0 - A_2(-t_0,0) + V_2^-(0) - B_{\mathrm{tag}} \\
&\leq (1-\rho+\delta)t_0 + (K+1)(\epsilon+\gamma)x - B_{\mathrm{tag}},
\end{aligned}
$$

where we used (15) and (16) in the final inequality. Applying similar arguments as in case (i) yields

$$
\begin{aligned}
V_1(0) &\geq A_1(-t_0,0) - B_1(-t_0,0) \\
&\geq (Kr-1+\rho-\delta)t_0 - (K+1)(\epsilon+\gamma)x + B_{\mathrm{tag}} \\
&\geq (Kr-1+\rho-\delta)\left[\frac{x(1+K\epsilon+K\gamma)}{K(r-\frac{1-\rho+\delta}{K+1})} + y\right] - (K+1)(\epsilon+\gamma)x \\
&\quad + \frac{x(1+K\epsilon+K\gamma)}{K(r-\frac{1-\rho+\delta}{K+1})}\frac{1-\rho+\delta}{K+1} + y(1-\rho+\delta-Kr) + (\epsilon+\gamma)x \\
&= x,
\end{aligned}
$$

where we used (14) in the third inequality. This completes the proof. □

We now use the sample-path relation of Lemma 5.1 to prove the next asymptotic lower bound for the class-1 workload distribution.

Proposition 5.2. *(lower bound) Assume the class-2 arrivals follow a renewal process with mean interarrival time $\alpha = 1/\lambda = \beta/\rho$. If $B(\cdot) \in \mathcal{R}_{-\nu}$ and $Kr < 1-\rho < (K+1)r$, then*

$$
\liminf_{x\to\infty} \frac{\mathbb{P}\left\{V_1 > x\right\}}{\frac{\rho}{1-\rho-Kr}\mathbb{P}\left\{B^r > \frac{x\frac{1-\rho}{K+1}}{K(r-\frac{1-\rho}{K+1})}\right\}} \geq 1.
$$

Proof. Let $-t_0 - \tau_{-m}$ be the arrival epoch of the $(m+1)$-th class-2 user before time $-t_0$ (counting backwards). In particular, τ_0 is the backward recurrence time of the class-2 arrival process at time $-t_0$. The corresponding service requirements are denoted by B_{-m}, $m \geq 0$.
In the following γ, δ, ϵ, κ and ζ are all small, but positive real numbers. Denote $g(\gamma,\delta,\epsilon,\kappa) := \frac{(1+K\epsilon+K\gamma)}{K(r-\frac{1-\rho+\delta}{K+1})}\frac{1-\rho+\delta}{K+1} + (\epsilon+\gamma) + (1-\rho+\delta-Kr)\kappa$, and rewrite (14) into $B_{-m} > g(\gamma,\delta,\epsilon,0)x + (1-\rho+\delta-Kr)\tau_{-m}$ for some $m \geq 0$. To bound the probability

of (16), we apply the model with $K+1$ permanent customers, giving $V_2^-(0) \le V_{(K+1)}(0)$. Now, using Lemma 5.1 yields

$$
\begin{aligned}
& \mathbb{P}\{V_1(0) > x\} \\
\ge\ & \mathbb{P}\{A_2(-t_0,0) \ge (\rho-\delta)t_0 - (K+1)\gamma x; V_2^-(0) \le (K+1)\epsilon x; \\
& \qquad \exists m \ge 0 : B_{-m} > g(\gamma,\delta,\epsilon,0)x + (1-\rho+\delta-Kr)\tau_{-m}; \\
& \qquad \forall k \ge 0 : \tau_{-k} \le k(\alpha+\zeta) + \kappa x\} \\
\ge\ & \mathbb{P}\{\exists m \ge 0 : B_{-m} > g(\gamma,\delta,\epsilon,\kappa)x + (1-\rho+\delta-Kr)m(\alpha+\zeta)\} \\
& \times \mathbb{P}\{A_2(-t_0,0) \ge (\rho-\delta)t_0 - (K+1)\gamma x; V_{(K+1)}(0) \le (K+1)\epsilon x; \\
& \qquad \forall k \ge 0 : \tau_{-k} \le k(\alpha+\zeta) + \kappa x\}.
\end{aligned} \tag{18}
$$

We study each of the two probabilities separately. First note that

$$
\begin{aligned}
& \mathbb{P}\{\exists m \ge 0 : B_{-m} > g(\gamma,\delta,\epsilon,\kappa)x + (1-\rho+\delta-Kr)m(\alpha+\zeta)\} \\
\ge\ & \sum_{m=0}^{\infty} \mathbb{P}\{B_{-m} > g(\gamma,\delta,\epsilon,\kappa)x + (1-\rho+\delta-Kr)m(\alpha+\zeta)\} \\
& - \sum_{m=0}^{\infty} \sum_{n=m+1}^{\infty} \mathbb{P}\{B_{-m} > g(\gamma,\delta,\epsilon,\kappa)x + (1-\rho+\delta-Kr)m(\alpha+\zeta), \\
& \qquad\qquad B_{-n} > g(\gamma,\delta,\epsilon,\kappa)x + (1-\rho+\delta-Kr)n(\alpha+\zeta)\} \\
\sim\ & (1+o(1))\frac{\beta/(\alpha+\zeta)}{1-\rho-Kr+\delta}\mathbb{P}\{B^r > g(\gamma,\delta,\epsilon,\kappa)x\},
\end{aligned} \tag{19}
$$

where we used similar arguments as in [10] in the final step. As for the second probability in (18), observe that the τ_{-k}, $A_2(-t_0,0)$, and $V_{(K+1)}(0)$ are not independent. However, we may write

$$
\begin{aligned}
& \mathbb{P}\left\{A_2(-t_0,0) \ge (\rho-\delta)t_0 - (K+1)\gamma x; V_{(K+1)}(0) \le (K+1)\epsilon x; \right. \\
& \qquad \left. \forall k \ge 0 : \tau_{-k} \le k(\alpha+\zeta) + \kappa x\right\} \\
\ge\ & \mathbb{P}\{A_2(-t_0,0) \ge (\rho-\delta)t_0 - (K+1)\gamma x\} - \mathbb{P}\left\{V_{(K+1)}(0) > (K+1)\epsilon x\right\} \\
& - \mathbb{P}\{\exists k \ge 0 : \tau_{-k} > k(\alpha+\zeta) + \kappa x\}.
\end{aligned}
$$

Now, $\mathbb{P}\{A_2(-t_0,0) \ge (\rho-\delta)t_0 - (K+1)\gamma x\} \to 1$ as $x \to \infty$ (and thus $t_0 \to \infty$). Moreover, since $V_{(K+1)}(0)$ has a proper distribution, we have

$$
\lim_{x\to\infty} \mathbb{P}\left\{V_{(K+1)}(0) > (K+1)\epsilon x\right\} = 0,
$$

and by the Strong Law of Large Numbers (the backward recurrence time at time $-t_0$ has a proper distribution because the renewal process has finite mean),

$$
\lim_{x\to\infty} \mathbb{P}\{\exists k \ge 0 : \tau_{-k} > k(\alpha+\zeta) + \kappa x\} = 0.
$$

Observing that the system is in steady state and using (19), we have

$$
\liminf_{x\to\infty} \frac{\mathbb{P}\{V_1 > x\}}{\frac{\beta/(\alpha+\zeta)}{1-\rho-Kr+\delta}\mathbb{P}\{B^r > g(\gamma,\delta,\epsilon,\kappa)x\}} \ge 1.
$$

Finally, use the fact that $B^r(\cdot) \in \mathcal{R}_{1-\nu}$ to obtain

$$\liminf_{x\to\infty} \frac{\mathbb{P}\{V_1 > x\}}{\frac{\rho}{1-\rho-Kr}\mathbb{P}\left\{B^r > \frac{x}{K(r-\frac{1-\rho}{K+1})}\frac{1-\rho}{K+1}\right\}} \tag{20}$$

$$\geq \liminf_{x\to\infty} \frac{\mathbb{P}\{V_1 > x\}}{\frac{\beta/(\alpha+\zeta)}{1-\rho-Kr+\delta}\mathbb{P}\{B^r > g(\gamma,\delta,\epsilon,\kappa)x\}} \frac{\frac{\beta/(\alpha+\zeta)}{1-\rho-Kr+\delta}\mathbb{P}\{B^r > g(\gamma,\delta,\epsilon,\kappa)x\}}{\frac{\rho}{1-\rho-Kr}\mathbb{P}\{B^r > g(0,0,0,0)x\}}$$

$$\geq \liminf_{x\to\infty} \frac{\frac{\beta/(\alpha+\zeta)}{1-\rho-Kr+\delta}\mathbb{P}\{B^r > g(\gamma,\delta,\epsilon,\kappa)x\}}{\frac{\rho}{1-\rho-Kr}\mathbb{P}\{B^r > g(0,0,0,0)x\}} \uparrow 1, \qquad \gamma,\delta,\epsilon,\kappa,\zeta \downarrow 0.$$

□

5.2 Approach 2

As in Subsection 5.1, we start by deriving a sufficient sample-path condition for the event $V_1(0) > x$ to occur, but now based on the alternative characterization of the dominant scenario in Section 4 (Lemma 5.3). Then, we translate the sample-path statement into a probabilistic lower bound which can be used to determine the asymptotic tail behavior of $\mathbb{P}\{V_1 > x\}$ (Proposition 5.4).

We first introduce some additional notation and terminology. In the proof we frequently use the notion of "small" users. A user is called "small" if its (initial) service requirement does not exceed κx, for some $\kappa > 0$ independent of x. Denote by $N^{(u,v]}(t)$ the number of class-2 users in the system at time t that arrived during $(u, v]$, and add the subscript $\leq \kappa x$ when only "small" class-2 users are considered. Define $t_0 := \frac{x(1+\gamma+M_0\kappa)}{K(r-\frac{1-\rho+\delta}{K+1})}$, and fix $L_0 \geq \frac{1+K\rho}{1-\rho}$ and $M_0 \geq \max\{L_0, \frac{\rho(K+L_0)}{1-\rho}\}$. In the proof, users arriving before time $-t_0$ are referred to as "old" users, while users arriving after time $-t_0$ are called "new". Let $-u_0$, $u_0 := \sup\{0 \leq t \leq t_0 : N^{(-\infty,-t_0]}(-t) \leq L_0\}$, be the first epoch after time $-t_0$ that there are less than L_0 "old" class-2 users. Similarly, let $-s_0$, $s_0 := \inf\{0 \leq t \leq t_0 : N^{(-t_0,-t]}_{\leq\kappa x}(-t) < M_0\}$, be the last epoch before time 0 that there are less than M_0 "new small" class-2 users in the system.

Now, for fixed $\delta, \epsilon, \kappa, L_0, M_0 > 0$, consider the following two events.

1. At time $-t_0$, the total amount of work in the system satisfies

$$V_1(-t_0) + V_2(-t_0) \geq x(1 + \gamma + M_0\kappa) - (Kr + \rho - 1 - \delta)t_0 \tag{21}$$

2. For the amount of "small" class-2 traffic arriving in $(-t_0, -s_0]$ it holds that

$$A_{2,\leq\kappa x}(-t_0, -s_0) \geq (\rho - \delta)(t_0 - s_0) - \gamma x \tag{22}$$

We first prove the next sample-path relation.

Lemma 5.3. *If the events (21) and (22) occur simultaneously, then* $V_1(0) > x$.

Proof. We distinguish between two cases, depending on whether $u_0 \leq s_0$ or $u_0 > s_0$. First, we consider the 'easy' case $u_0 \leq s_0$ (or alternatively $-u_0 \geq -s_0$). Observe that during the entire interval $(-t_0, 0]$ there are at least L_0 class-2 users in the system (either "old" or "new"). Thus, $B_2(-t_0,0) \geq \frac{L_0}{K} B_1(-t_0,0)$, so that $B_1(-t_0,0) \leq \frac{K}{K+L_0} t_0$. Using the above in addition to (9), we obtain

$$\begin{aligned} V_1(0) &\geq A_1(-t_0,0) - B_1(-t_0,0) \geq Krt_0 - \frac{K}{K+L_0} t_0 \\ &\geq K\Big(r - \frac{1}{K + \frac{1+K\rho}{1-\rho}}\Big) \frac{x(1+\gamma+M_0\kappa)}{K(r - \frac{1-\rho+\delta}{K+1})} > x, \end{aligned}$$

where we used the definition of t_0 and the fact that $L_0 \geq \frac{1+K\rho}{1-\rho}$ in the third step.

Now consider the 'hard' case $u_0 > s_0$ (or $-u_0 < -s_0$). Denote by $B_2^{(u,v]}(s,t)$ the amount of service received during $(s,t]$ by class-2 users arriving in the interval $(u,v]$ (again, add the subscript $\leq \kappa x$ when only "small" class-2 users are considered). Using (9), the amount of service received during $(-t_0,-s_0]$ by the "new" class-2 users is bounded from below by

$$\begin{aligned} B_2^{(-t_0,0]}(-t_0,-s_0) &\geq B_{2,\leq\kappa x}^{(-t_0,-s_0]}(-t_0,-s_0) \\ &\geq A_{2,\leq\kappa x}(-t_0,-s_0) - V_{2,\leq\kappa x}^{(-t_0,-s_0]}(-s_0) \\ &\geq (\rho-\delta)(t_0-s_0) - \gamma x - M_0\kappa x, \end{aligned}$$

where $V_{2,\leq\kappa x}^{(u,v]}(t)$ denotes the workload at time t associated with "small" class-2 users arriving in $(u,v]$. Note that the final step follows from (22) and the definition of s_0. Since $M_0 \geq \frac{\rho(K+L_0)}{1-\rho}$, we also have

$$B_2^{(-t_0,0]}(-s_0,0) \geq \frac{M_0}{M_0+K+L_0} s_0 \geq (\rho-\delta)s_0.$$

Hence,

$$B_2^{(-t_0,0]}(-t_0,0) \geq (\rho-\delta)t_0 - \gamma x - M_0\kappa x. \tag{23}$$

Next, denote by $n \geq 0$ the number of "old" class-2 users present at time 0. We distinguish between two cases: (i) $n=0$; and (ii) $n \geq 1$.

First, consider case (i). Note that $B_2^{(-\infty,-t_0]}(-t_0,0) = V_2(-t_0)$ and rewrite (8) into

$$B_1(-t_0,0) \leq t_0 - B_2^{(-\infty,-t_0]}(-t_0,0) - B_2^{(-t_0,0]}(-t_0,0). \tag{24}$$

Using (9), (21), (23), and (24), we deduce

$$\begin{aligned} V_1(0) &= V_1(-t_0) + A_1(-t_0,0) - B_1(-t_0,0) \\ &\geq V_1(-t_0) + V_2(-t_0) + Krt_0 - t_0 + (\rho-\delta)t_0 - (\gamma+M_0\kappa)x \\ &\geq x(1+\gamma+M_0\kappa) - (Kr+\rho-1-\delta)t_0 + Krt_0 - (1-\rho+\delta)t_0 - (\gamma+M_0\kappa)x \\ &= x. \end{aligned}$$

Second, consider case (ii). Because of the PS discipline, it follows from (8)

$$B_1(-t_0,0) \leq \frac{K}{K+1}[t_0 - B_2^{(-t_0,0]}(-t_0,0)]. \tag{25}$$

Now, combining (9), (23), and (25) yields

$$\begin{aligned} V_1(0) &\geq A_1(-t_0,0) - B_1(-t_0,0) \\ &\geq Krt_0 - \frac{K}{K+1}[(1-\rho+\delta)t_0 + (\gamma + M_0\kappa)x] \\ &= [Kr - \frac{K}{K+1}(1-\rho+\delta)]\frac{x(1+\gamma+M_0\kappa)}{K(r-\frac{1-\rho+\delta}{K+1})} - \frac{K}{K+1}(\gamma+M_0\kappa)x \\ &> x, \end{aligned}$$

where we used that $\gamma, \kappa, M_0 > 0$. This completes the proof. □

We now exploit the sample-path relation in Lemma 5.3 to establish the next asymptotic lower bound for the class-1 workload distribution.

Proposition 5.4. *(lower bound) If $B(\cdot) \in \mathcal{R}_{-\nu}$ and $Kr < 1-\rho < (K+1)r$, then*

$$\liminf_{x\to\infty} \frac{\mathbb{P}\{V_1 > x\}}{\frac{\rho}{1-\rho-Kr}\mathbb{P}\left\{B^r > \frac{x\frac{1-\rho}{K+1}}{K(r-\frac{1-\rho}{K+1})}\right\}} \geq 1.$$

Proof. First observe that the events (21) and (22) are not independent. However, $V_1(-T_0) + V_2(-T_0)$ and $A_{2,\leq\kappa x}(-t_0,-s_0)$ are independent, with $-T_0$ representing the last arrival epoch of class 2 before time $-t_0$. Note that

$$V_1(-t_0) + V_2(-t_0) \geq V_1(-T_0) + V_2(-T_0) - \tau_0,$$

where τ_0 represents the backward recurrence time of the class-2 arrival process at time $-t_0$ (see also Subsection 5.1), which is independent of $V_1(-T_0) + V_2(-T_0)$ as well. Using Lemma 5.3 and the above, we obtain

$$\begin{aligned} &\mathbb{P}\{V_1(0) > x\} \\ \geq\ &\mathbb{P}\{V_1(-T_0) + V_2(-T_0) > x(1+\gamma+M_0\kappa) - (Kr+\rho-1-\delta)t_0 + \tau_0; \\ &A_{2,\leq\kappa x}(-t_0,-s_0) \geq (\rho-\delta)(t_0-s_0) - \gamma x\} \\ \geq\ &\mathbb{P}\{V_1(-T_0) + V_2(-T_0) > x(1+\gamma+M_0\kappa+\epsilon) - (Kr+\rho-1-\delta)t_0\} \\ &\times\left[\mathbb{P}\left\{\sup_{0\leq t\leq t_0}\{(\rho-\delta)(t_0-t) - A_{2,\leq\kappa x}(-t_0,-t)\} \leq \gamma x\right\} - \mathbb{P}\{\tau_0 > \epsilon x\}\right]. \end{aligned}$$

Now, first invoking Proposition A.1 in Appendix A and then Theorem 4.2 yields

$$\begin{aligned} &\mathbb{P}\{V_1(-T_0) + V_2(-T_0) > x(1+\gamma+M_0\kappa+\epsilon) - (Kr+\rho-1-\delta)t_0\} \\ \sim\ &\frac{\rho}{1-\rho-Kr}\mathbb{P}\left\{B^r > \frac{x(1+\gamma+M_0\kappa)\frac{1-\rho+\delta}{K+1}}{K(r-\frac{1-\rho+\delta}{K+1})} + \epsilon x\right\}. \end{aligned} \tag{26}$$

Because τ_0 has a proper distribution, we have $\lim_{x\to\infty}\mathbb{P}\{\tau_0 > \epsilon x\} = 0$. Moreover, for $u > 0$ sufficiently large so that $\sup_{0\leq t}\{(\rho-\delta)t - A_{2,\leq u}(0,t)\}$ has a proper distribution, we have

$$\begin{aligned}
&\lim_{x\to\infty} \mathbb{P}\left\{ \sup_{0\le t\le t_0} \{(\rho-\delta)(t_0-t) - A_{2,\le\kappa x}(-t_0,-t)\} \le \gamma x \right\} \\
\ge\ &\lim_{x\to\infty} \mathbb{P}\left\{ \sup_{0\le t\le t_0} \{(\rho-\delta)(t_0-t) - A_{2,\le u}(-t_0,-t)\} \le \gamma x \right\} \\
\ge\ &\lim_{x\to\infty} \mathbb{P}\left\{ \sup_{t\ge 0} \{(\rho-\delta)t - A_{2,\le u}(0,t)\} \le \gamma x \right\} = 1.
\end{aligned}$$

Combining the above arguments and applying (26), we obtain

$$\liminf_{x\to\infty} \frac{\mathbb{P}\{V_1 > x\}}{\frac{\rho}{1-Kr-\rho}\mathbb{P}\left\{B^r > \frac{x(1+\gamma+M_0\kappa)\frac{1-\rho+\delta}{K+1}}{K(r-\frac{1-\rho+\delta}{K+1})} + \epsilon x\right\}} \ge 1.$$

The proof may then be readily completed along the lines of (20). □

6 Upper bound

In this section we derive an asymptotic upper bound for $\mathbb{P}\{V_1 > x\}$. In the proof we frequently use the notion of a "large" user. A user is called "large" if its (initial) service requirement exceeds the value κx, for some fixed $\kappa > 0$ independent of x. Also, let $N_{>b}(s,t)$ be the number of class-2 users arriving during the time interval $(s,t]$ whose service requirement exceeds the value b. In particular, let $N(s,t) := N_{>0}(s,t)$ be the total number of class-2 users arriving in the interval $(s,t]$.
To handle scenarios in which the system is not work-conserving, we introduce the epoch $s^* := \inf\{t \ge 0 : V_1(-t) = 0\}$, which represents the last epoch before time 0 that the class-1 workload was zero. Note that $V_1(t) > 0$ for $t \in (-s^*, 0]$, and the system thus uses the full service capacity during the given interval. For epochs at which $V_1(t) = 0$, we make the following observation.

Observation 6.1. If $V_1(t) = 0$, then the available service rate for class 1 at time t is at least Kr, hence $\frac{K}{K+N(t)} \ge Kr$. Rewriting the inequality gives that $N(t) \le M$, with $M := \lfloor \frac{1}{r} \rfloor - K$. ◇

We are now ready to prove the upper bound for $\mathbb{P}\{V_1 > x\}$.

Proposition 6.1. *(upper bound) If $B(\cdot) \in \mathcal{R}_{-\nu}$ and $Kr < 1-\rho < (K+1)r$, then*

$$\limsup_{x\to\infty} \frac{\mathbb{P}\{V_1 > x\}}{\frac{\rho}{1-\rho-Kr}\mathbb{P}\left\{B^r > \frac{x\frac{1-\rho}{K+1}}{K(r-\frac{1-\rho}{K+1})}\right\}} \le 1.$$

Proof. Let $t_1 := \frac{x(1-\epsilon)}{K(r-\frac{1-\rho-\delta}{K+1})}$. Then, for $\delta > 0, 0 < \epsilon < 1$,

$$\begin{aligned}
&\mathbb{P}\{V_1(0) > x\} \\
\le\ &\mathbb{P}\{V_1(-t_1) + V_2(-t_1) > x(1-\epsilon) - (Kr+\rho+\delta-1)t_1\} &(27)\\
&+\mathbb{P}\{V_1(-t_1) + V_2(-t_1) \le x(1-\epsilon) - (Kr+\rho+\delta-1)t_1; V_1(0) > x\}. &(28)
\end{aligned}$$

First, we determine the asymptotic behavior of (27). Then we show that the probability in (28) is negligible compared to the probability in (27) as $x \to \infty$. This way, we prove that the scenario described in Section 4 is indeed the dominant one.

Let us start with the former and note that the system at time $-t_1$ is in steady state. First, use Proposition A.1 and then Theorem 4.2 to obtain that (27) behaves as

$$\begin{aligned} & \mathbb{P}\left\{V_1(-t_1)+V_2(-t_1) > x(1-\epsilon) - (Kr+\rho+\delta-1)t_1\right\} \\ \sim\ & \frac{\rho}{1-\rho-Kr}\mathbb{P}\left\{B^r > \frac{x(1-\epsilon)\frac{1-\rho-\delta}{K+1}}{K(r-\frac{1-\rho-\delta}{K+1})}\right\}. \end{aligned}$$

Using the fact that $B^r(\cdot) \in \mathcal{R}_{1-\nu}$ (and letting $\delta, \epsilon \downarrow 0$), it easily follows that

$$\limsup_{x\to\infty} \frac{\mathbb{P}\left\{V_1(-t_1)+V_2(-t_1) > x(1-\epsilon) - (Kr+\rho+\delta-1)t_1\right\}}{\mathbb{P}\left\{B^r > \frac{x\frac{1-\rho}{K+1}}{K(r-\frac{1-\rho}{K+1})}\right\}} \le 1.$$

To prove that any alternative scenario is highly unlikely compared to the dominant one, we show that, for $0 < \delta < 1-\rho-Kr$ and $0 < \epsilon < 1$,

$$\limsup_{x\to\infty} \frac{\mathbb{P}\left\{V_1(-t_1)+V_2(-t_1) \le x(1-\epsilon) - (Kr+\rho+\delta-1)t_1; V_1(0) > x\right\}}{\mathbb{P}\left\{B^r > \frac{x\frac{1-\rho}{K+1}}{K(r-\frac{1-\rho}{K+1})}\right\}} = 0.$$

To do so, we split (28) by distinguishing between 0, 1, and 2 or more large user arrivals during $(-t_1, 0]$, respectively. More specifically, write

$$\begin{aligned} & \mathbb{P}\left\{V_1(-t_1)+V_2(-t_1) \le x(1-\epsilon) - (Kr+\rho+\delta-1)t_1; V_1(0) > x\right\} \qquad (29) \\ =\ & \mathbb{P}\left\{V_1(-t_1)+V_2(-t_1) \le x(1-\epsilon) - (Kr+\rho+\delta-1)t_1; N_{>\kappa x}(-t_1,0)=0; V_1(0) > x\right\} \\ & +\mathbb{P}\left\{V_1(-t_1)+V_2(-t_1) \le x(1-\epsilon) - (Kr+\rho+\delta-1)t_1; N_{>\kappa x}(-t_1,0)=1; V_1(0) > x\right\} \\ & +\mathbb{P}\left\{V_1(-t_1)+V_2(-t_1) \le x(1-\epsilon) - (Kr+\rho+\delta-1)t_1; N_{>\kappa x}(-t_1,0)\ge 2; V_1(0) > x\right\} \\ =:\ & I + II + III. \end{aligned}$$

In the remainder of the proof we show that each of the three terms is negligible compared to the dominant scenario.

Term I

To bound term I, we consider the total workload at time 0. Recall that s^* represents the last epoch before time 0 that the class-1 workload was zero, and define $s' := \min\{s^*, t_1\}$, so that $V_1(t) > 0$ for $t \in (-s', 0]$. Then, using (9) and the fact that the system is work-conserving during $(-s', 0]$, we have

$$\begin{aligned} V_1(0)+V_2(0) &= V_1(-s')+V_2(-s')+Krs'+A_2(-s',0)-s' \\ &= V_1(-s')+V_2(-s')-(1-Kr-\rho-\delta)s'+A_2(-s',0)-(\rho+\delta)s' \\ &\le \max\{V_1(-t_1)+V_2(-t_1)-(1-Kr-\rho-\delta)t_1, V_2(-s^*)\} \\ &\quad + \sup_{0\le s\le t_1}\{A_2(-s,0)-(\rho+\delta)s\}, \end{aligned}$$

where we choose $0 < \delta < 1 - Kr - \rho$. Moreover, take $\kappa > 0$ such that $M\kappa < 1$. Then, combining the above and using Observation 6.1 yields

$$\begin{aligned} I &\leq \mathbb{P}\{\max\{V_1(-t_1) + V_2(-t_1) - (1 - Kr - \rho - \delta)t_1, V_2(-s^*)\} \\ &\quad + \sup_{0\leq s\leq t_1} \{A_2(-s,0) - (\rho+\delta)s\} > x; \\ &\qquad V_1(-t_1) + V_2(-t_1) < x(1-\epsilon) - (Kr+\rho+\delta-1)t_1; N_{>\kappa x}(-t_1,0) = 0\} \\ &\leq \mathbb{P}\left\{\max\{(1-\epsilon)x, M\kappa x\} + \sup_{0\leq s\leq t_1}\{A_2(-s,0) - (\rho+\delta)s\} > x \mid N_{>\kappa x}(-t_1,0) = 0\right\} \\ &\leq \mathbb{P}\left\{\sup_{0\leq s\leq t_1}\{A_2(-s,0) - (\rho+\delta)s\} > \xi x \mid N_{>\kappa x}(-t_1,0) = 0\right\}, \end{aligned}$$

where $\xi := \min\{\epsilon, 1 - M\kappa\}$. Lemma B.4 in Appendix B implies that $I = o(\mathbb{P}\{B^r > x\})$, as $x \to \infty$.

Term II

By conditioning on $V_1(-t_1) + V_2(-t_1)$, we obtain that term II equals

$$\begin{aligned} &\mathbb{P}\{\eta x < V_1(-t_1) + V_2(-t_1) < x(1-\epsilon) - (Kr+\rho+\delta-1)t_1; N_{>\kappa x}(-t_1,0) = 1; V_1(0) > x\} \\ &+ \mathbb{P}\{V_1(-t_1) + V_2(-t_1) < \eta x; N_{>\kappa x}(-t_1,0) = 1; V_1(0) > x\}. \qquad (30) \end{aligned}$$

Again by Theorem 4.2 and Proposition A.1, in addition to Lemma B.3 with $t_1 = \gamma x$, we can control the first term of (30) as a "combination of two unlikely events". Specifically, as $x \to \infty$, the term is bounded by

$$\mathbb{P}\{V_1(-t_1) + V_2(-t_1) > \eta x\}\,\mathbb{P}\left\{I(B > \kappa x) + \tilde{N}_{>\kappa x}(-t_1,0) \geq 1\right\} = o(\mathbb{P}\{B^r > x\}),$$

with $I(\cdot)$ the indicator function, and $\tilde{N}_{>\kappa x}(-t_1,0)$ having the same distribution as $N_{>\kappa x}(-t_1,0)$, but independent of $V_1(-t_1) + V_2(-t_1)$.

For the second term, we use $s' = \min\{s^*, t_1\}$ (as in term I), so that $V_1(t) > 0$ for $t \in (-s', 0]$. Also, we tag the user with service requirement larger than κx, and let $V_2^-(t)$ be the class-2 workload at time t, excluding the tagged class-2 user. As in Section 4, denote by $B_2^-(s,t)$ the amount of service received by class 2 in the interval $(s,t]$, except for the tagged user. Then, using (9) in the first step and Observation 6.1 in the second, we find

$$B_2^-(-s',0) = V_2^-(-s') + A_2^-(-s',0) - V_2(0) \leq \zeta x + A_2^-(-s',0),$$

where $A_2^-(-s',0)$ denotes the amount of class-2 traffic generated during $(-s',0]$ excluding the tagged user, and $\zeta := \max\{\eta, M\kappa\}$. The large user together with the class-1 users receive the remaining amount of service: $B_1^+(-s',0) \geq s' - A_2^-(-s',0) - \zeta x$. Because of the PS discipline, $B_1(-s',0) \geq \frac{K}{K+1}B_1^+(-s',0)$. Thus, using the above and applying (9),

$$\begin{aligned} V_1(0) &= V_1(-s') + A_1(-s',0) - B_1(-s',0) \\ &\leq \max\{V_1(-t_1), V_1(-s^*)\} + Krs' - \frac{K(s' - A_2^-(-s',0) - \zeta x)}{K+1} \\ &\leq \zeta x + \sup_{0\leq s\leq t_1}\left\{Krs - \frac{K(s - A_2^-(-s,0) - \zeta x)}{K+1}\right\}. \end{aligned}$$

Thus,

$$II \leq \mathbb{P}\left\{\zeta x + \sup_{0\leq s\leq t_1}\left\{Krs - \frac{K(s - A_2^-(-s,0) - \zeta x)}{K+1}\right\} > x \mid N_{>\kappa x}(-t_1,0) = 1\right\} + o(\mathbb{P}\{B^r > x\}),$$

as $x \to \infty$. Choose η, κ such that $\max\{\eta, M\kappa\} \leq \frac{K+1}{K+3}\epsilon$. Then, using $r > \frac{1-\rho}{K+1}$ in the second inequality and substituting $x = \frac{t_1 K(r - \frac{1-\rho-\delta}{K+1})}{1-\epsilon}$ yields

$$\begin{aligned}
& \mathbb{P}\left\{\zeta x + \sup_{0\leq s\leq t_1}\left\{Krs - \frac{K(s - A_2^-(-s,0) - \zeta x)}{K+1}\right\} > x \mid N_{>\kappa x}(-t_1,0) = 1\right\} \\
= \;& \mathbb{P}\left\{\sup_{0\leq s\leq t_1}\left\{Krs - \frac{K(s - A_2^-(-s,0))}{K+1}\right\} > x(1 - \frac{3K+1}{K+1}\zeta) + \frac{K}{K+1}\zeta x \mid N_{>\kappa x}(-t_1,0) = 1\right\} \\
\leq \;& \mathbb{P}\left\{\sup_{0\leq s\leq t_1}\left\{Krs - \frac{K(s - A_2^-(-s,0))}{K+1}\right\} - t_1 K(r - \frac{1-\rho-\delta}{K+1}) > \frac{K}{K+1}\zeta x \mid N_{>\kappa x}(-t_1,0) = 1\right\} \\
\leq \;& \mathbb{P}\left\{\sup_{0\leq s\leq t_1}\left\{A_2^-(-s,0) - (\rho+\delta)s\right\} > \zeta x \mid N_{>\kappa x}(-t_1,0) = 1\right\} \\
\leq \;& \mathbb{P}\left\{\sup_{0\leq s\leq t_1}\left\{A_2(-s,0) - (\rho+\delta)s\right\} > \zeta x \mid N_{>\kappa x}(-t_1,0) = 0\right\},
\end{aligned}$$

which can be controlled using Lemma B.4. This completes the estimation of term II.

Term III
It follows directly from Lemma B.3 that $III = o(\mathbb{P}\{B^r > x\})$, as $x \to \infty$.

The proof is now completed by first letting $x \to \infty$, then $\eta, \kappa \downarrow 0$, and finally $\delta, \epsilon \downarrow 0$. □

7 Generalization to variable-rate streaming traffic

As mentioned earlier, the assumption that class 1 generates traffic at a constant rate Kr is actually not crucial. In this section, we show that our results remain valid in case class 1 generates traffic according to a general stationary process, provided that deviations from the mean are sufficiently unlikely. In such a scenario, the variations in class-1 traffic do not matter asymptotically, because they average out.
First, in Subsection 7.1 we consider the total workload of class 1 and extend Theorem 4.1 to the case of variable-rate streaming sources. Second, in Subsection 7.2 we consider the tail asymptotics of the joint workload distribution of individual class-1 users. Note that the individual class-1 workloads are not necessarily equal, since the traffic rates of the individual streaming sources also vary.

7.1 Total workload

In this subsection, we show that our results remain valid in case class 1 generates traffic according to a general stationary process with mean rate $\mathbb{E}[A_1(t,t+1)] = Kr$, provided that significant deviations from the mean are sufficiently unlikely. More specifically, we assume that the class-1 traffic satisfies the following assumption:

Assumption 7.1. *For all $\phi > 0$ and $\psi > 0$,*

$$\mathbb{P}\left\{\sup_{t\geq 0}\{A_1(-t,0) - K(r+\psi)t\} > \phi x\right\} = o(\mathbb{P}\{B^r > x\}), \qquad \text{as } x \to \infty.$$

Note that Assumption 7.1 holds for all $\phi > 0$ whenever it holds for one such value. Assumption 7.1 serves to ensure that the likelihood that rate variations in class-1 traffic cause a large workload is asymptotically negligible compared to scenarios with a large class-2 user described earlier. Also, observe that it may be equivalently expressed as

$$\mathbb{P}\left\{V_1^{K(r+\psi)} > \phi x\right\} = o(\mathbb{P}\{B^r > x\}), \qquad \text{as } x \to \infty, \tag{31}$$

where V_1^c denotes the steady-state workload in a system with service capacity c fed by class 1 only. Assumption 7.1 is satisfied by a wide range of traffic processes, as illustrated by the next two examples.

Example 7.1. *(Instantaneous bursts)* Let each class-1 user generate instantaneous bursts according to a renewal process, and let the burst sizes have distribution $F_1(\cdot)$, with mean σ_1. Let the interarrival times between bursts also be generally distributed with mean σ_1/r. Assume that $1 - F_1(x) = o(\mathbb{P}\{B > x\})$ as $x \to \infty$. Then, it follows from [3, Theorem 4.1] that Assumption 7.1 is satisfied.

Example 7.2. *(On-Off source)* Let each class-1 user generate traffic according to an On-Off process, alternating between On- and Off-periods. The On-periods have general distribution $F_1(\cdot)$ with mean σ_1, and the Off-periods also follow a general distribution with mean $1/\lambda_1$. A class-1 user produces traffic at a constant rate r_{on} while On, and generates traffic at rate r_{off} while Off, $r_{\text{off}} < r < r_{\text{on}}$ (including the important special case in which $r_{\text{off}} = 0$), with $r(1 + \lambda_1\sigma_1) = r_{\text{off}} + r_{\text{on}}\lambda_1\sigma_1$.
Moreover, assume that $1 - F_1(x) = o(\mathbb{P}\{B > x\})$ as $x \to \infty$. Now, asymptotic results for a fluid queue fed by multiple homogeneous On-Off sources (in particular [15], [30, Corollary 3.1] with $N^* = 1$), imply that Assumption 7.1 is satisfied.

In the remainder of the section, we show that our results remain valid under Assumption 7.1. In particular, we prove that Theorem 4.1 still holds. We add the superscript 'var' to indicate quantities corresponding to the scenario with variable-rate streaming sources.

Theorem 7.1. *Suppose that the process $\{A_1(-t,0), t \geq 0\}$ satisfies Assumption 7.1. If $B(\cdot) \in \mathcal{R}_{-\nu}$ and $Kr < 1 - \rho < (K+1)r$, then*

$$\mathbb{P}\{V_1^{\text{var}} > x\} \sim \frac{\rho}{1-\rho-Kr}\mathbb{P}\left\{B^r > \frac{x\frac{1-\rho}{K+1}}{K(r - \frac{1-\rho}{K+1})}\right\}.$$

As before, the proof of Theorem 7.1 involves lower and upper bounds. In fact, the lower bound largely follows the lines of the proof of Proposition 5.2 (in Section 5), and is hardly affected by the variable rate of class 1. Informally speaking, the idea is to replace $A_1(s,t)$ by $K(r-\psi)(t-s) - \phi x$, and then use $\mathbb{E}[A_1(t,t+1)] = Kr$ to show that the correction terms $K\psi(t-s)$ and ϕx can be asymptotically neglected. More specifically, because the process $\{K(r-\psi)t - A_1(-t,0), t \geq 0\}$ has negative drift, for all $\phi, \psi > 0$,

$$\mathbb{P}\left\{\sup_{t\geq 0}\{K(r-\psi)t - A_1(-t,0)\} > \phi x\right\} \to 0, \qquad \text{as } x \to \infty. \tag{32}$$

Note that the above expression relates to long periods with less than average class-1 input, as opposed to Assumption 7.1 where periods with more than average class-1 traffic are considered.

Before describing the modifications of Subsection 5.2 required to handle variable-rate class-1 traffic, we note that a slightly more substantial modification is needed, to obtain an equivalence between $V_1^{\rm var}+V_2^{\rm var}$ and V_2^{1-Kr}. Moreover, in the lower bound we encounter the difficulty that $V_1^{\rm var}(-t_0) + V_2^{\rm var}(-t_0)$ and $A_1(-t_0,0)$ may no longer be independent. These issues are addressed in the proof of Proposition D.1 in Appendix D, where we extend relation (12) to the case of variable-rate class-1 traffic. Proposition D.1 may also be of independent interest. In addition, we show in the proposition that relation (12) remains valid for a non-critically loaded *work-conserving* system.

For the upper bound, we provide a proof based on a comparison with a leaky-bucket system and use results of Section 4 (in particular Theorem 4.1).

We now give the proofs of the lower and upper bounds, together yielding Theorem 7.1.

Proposition 7.2. *(lower bound) If $B(\cdot) \in \mathcal{R}_{-\nu}$ and $Kr < 1-\rho < (K+1)r$, then*

$$\liminf_{x\to\infty} \frac{\mathbb{P}\{V_1^{\rm var} > x\}}{\frac{\rho}{1-\rho-Kr}\mathbb{P}\left\{B^r > \frac{x\frac{1-\rho}{K+1}}{K(r-\frac{1-\rho}{K+1})}\right\}} \geq 1.$$

Proof. In view of the similarities with Subsection 5.2, we only give an outline of the proof. As in Subsection 5.2, consider the following three events:

- At time $-t_0$, with $t_0 := \frac{x(1+\gamma+M_0\kappa+\phi)}{K(r-\psi-\frac{1-\rho+\delta}{K+1})}$, the total amount of work in the system satisfies

$$V_1^{\rm var}(-t_0) + V_2^{\rm var}(-t_0) \geq x(1+\gamma+M_0\kappa+\phi) - (K(r-\psi)+\rho-1-\delta)t_0 \tag{33}$$

- The event (22)

- For the amount of class-1 traffic arriving in the interval $(-t_0, 0]$ it holds that

$$A_1(-t_0,0) \geq K(r-\psi)t_0 - \phi x \tag{34}$$

Some calculations similar to the proof of Lemma 5.3 show that, if the events (33), (22), and (34) occur simultaneously, then $V_1^{\rm var}(0) > x$. As in Subsection 5.2, let $-T_0$ be the last class-2 arrival epoch before time $-t_0$. We may write

$$
\begin{aligned}
& \mathbb{P}\{V_1^{\mathrm{var}}(0) > x\} \\
\geq\ & \mathbb{P}\{V_1^{\mathrm{var}}(-T_0) + V_2^{\mathrm{var}}(-T_0) > x(1+\gamma+M_0\kappa+\phi+\epsilon) - (K(r-\psi)+\rho-1-\delta)T_0; \\
& A_1(-T_0,0) \geq K(r-\psi)T_0 - \phi x; A_{2,\leq\kappa x}(-T_0,-s_0) \geq (\rho-\delta)(T_0-s_0) - \gamma x; \tau_0 \leq \epsilon x\} \\
\geq\ & \mathbb{P}\{V_1^{\mathrm{var}}(-T_0) + V_2^{\mathrm{var}}(-T_0) > x(1+\gamma+M_0\kappa+\phi+\epsilon) - (K(r-\psi)+\rho-1-\delta)T_0; \\
& \overrightarrow{U}_1^{K(r-\psi)}(-T_0) \leq \phi x\} \\
& \times \left[\mathbb{P}\left\{\sup_{0\leq t\leq T_0}\{(\rho-\delta)(T_0-t) - A_{2,\leq\kappa x}(-T_0,-t)\} \leq \gamma x\right\} - \mathbb{P}\{\tau_0 > \epsilon x\}\right],
\end{aligned}
$$

where $\overrightarrow{U}_1^c(-T_0) := \sup_{0\leq t\leq T_0}\{c(T_0-t) - A_1(-T_0,-t)\}$. The second and third probabilities can be treated as in Subsection 5.2. For the first probability, apply (45) (see the proof of Proposition D.1 in the Appendix) and then Theorem 4.2, to obtain

$$
\liminf_{x\to\infty} \frac{\mathbb{P}\{V_1^{\mathrm{var}} > x\}}{\frac{\rho}{1-\rho-Kr}\mathbb{P}\left\{B^r > \frac{x(1+\gamma+M_0\kappa+\phi)\frac{1-\rho+\delta}{K+1}}{K(r-\psi-\frac{1-\rho+\delta}{K+1})} + \epsilon x\right\}} \geq 1.
$$

Proposition 7.2 then follows from the fact that $B^r(\cdot) \in \mathcal{R}_{1-\nu}$ (let $\gamma,\delta,\epsilon,\kappa,\phi,\psi \downarrow 0$). □

For the proof of the upper bound we compare the class-1 workload in the scenario with variable-rate streaming traffic to that in a scenario with constant-rate streaming traffic. Suppose we feed the variable-rate streaming traffic into a system (the leaky bucket) that drains at constant rate $K(r+\psi)$ into a second resource that is shared with the elastic class according to $C_2(t) = N_{(K)}(t)/(N_{(K)}(t)+K)$ (see Section 3). Because the drain rate of the first resource never exceeds $K(r+\psi)$, the second resource is closely related to the class-1 workload in the case of constant-rate traffic (in fact, the *permanent-customer* scenario provides an upper bound). The total class-1 workload at the first and second resources at time t is an upper bound for $V_1^{\mathrm{var}}(t)$ (see Equation (35) below). The proof is then established by using Theorem 4.1.

Proposition 7.3. *(upper bound) Suppose that the process $\{A_1(-t,0), t \geq 0\}$ satisfies Assumption 7.1. If $B(\cdot) \in \mathcal{R}_{-\nu}$ and $Kr < 1-\rho < (K+1)r$, then*

$$
\limsup_{x\to\infty} \frac{\mathbb{P}\{V_1^{\mathrm{var}} > x\}}{\frac{\rho}{1-\rho-Kr}\mathbb{P}\left\{B^r > \frac{x\frac{1-\rho}{K+1}}{K(r-\frac{1-\rho}{K+1})}\right\}} \leq 1.
$$

Proof. Let $\psi > 0$. Using the definition of $V_1(t)$ in Section 2, we obtain the following representation

$$
V_1^{\mathrm{var}}(t) = \sup_{s\leq t}\{A_1(s,t) - C_1(s,t)\} = \sup_{s\leq t}\{A_1(s,t) - \int_s^t \frac{K}{K+N^{\mathrm{var}}(u)}\mathrm{d}u\},
$$

where the integral represents the amount of service available for class 1. Then,

$$\begin{aligned} V_1^{\mathrm{var}}(t) &= \sup_{s\le t}\{A_1(s,t) - K(r+\psi)(t-s) + K(r+\psi)(t-s) - \int_s^t \frac{K}{K+N^{\mathrm{var}}(u)}\mathrm{d}u\} \\ &\le \sup_{s\le t}\{A_1(s,t) - K(r+\psi)(t-s)\} \\ &\quad + \sup_{s\le t}\{K(r+\psi)(t-s) - \int_s^t \frac{K}{K+N^{\mathrm{var}}(u)}\mathrm{d}u\}. \end{aligned}$$

Let $V_1^{\mathrm{cst},\psi}(t) = \sup_{s\le t}\{K(r+\psi)(t-s) - \int_s^t \frac{K}{K+N_{(K)}(u)}\mathrm{d}u\}$ be the class-1 workload in a scenario with constant rate $r+\psi$ per streaming user and $C_2(t) \equiv N_{(K)}(t)/(N_{(K)}(t)+K)$ (independent of the class-1 workload; this corresponds to the *permanent-customer* scenario discussed in Section 3). Similar to the constant-rate model, $N^{\mathrm{var}}(t) \le N_{(K)}(t)$. Thus,

$$\int_s^t \frac{K}{K+N^{\mathrm{var}}(u)}\mathrm{d}u \ge \int_s^t \frac{K}{K+N_{(K)}(u)}\mathrm{d}u,$$

so that

$$V_1^{\mathrm{var}}(t) \le V_1^{K(r+\psi)}(t) + V_1^{\mathrm{cst},\psi}(t). \tag{35}$$

For any $\xi > 0$, this sample-path relation implies

$$\mathbb{P}\{V_1^{\mathrm{var}} > x\} \le \mathbb{P}\left\{V_1^{K(r+\psi)} > \xi x\right\} + \mathbb{P}\left\{V_1^{\mathrm{cst},\psi} > (1-\xi)x\right\},$$

where $V_1^{K(r+\psi)}$ and $V_1^{\mathrm{cst},\psi}$ have the limiting distributions of $V_1^{K(r+\psi)}(t)$ and $V_1^{\mathrm{cst},\psi}(t)$ for $t\to\infty$. The first term can be controlled by (31). For the second term, apply Theorem 4.1, use the fact that $B^r(\cdot) \in \mathcal{R}_{1-\nu}$, and let $\xi,\psi \downarrow 0$. This gives the desired result. □

7.2 Individual workloads

In this subsection we consider the asymptotics of the simultaneous workload distribution of the individual streaming users. In Subsection 7.1, we showed that a large service deficit for the K class-1 users together is most likely due to the arrival of a large class-2 user. Using similar arguments, we now also argue that the service deficits of the individual class-1 users are approximately equal after the arrival of a large class-2 user.

Denote by $A_{1,k}(s,t)$, $k = 1,\dots,K$, the total traffic of streaming user k during the interval $[s,t]$ with mean rate $\mathbb{E}[A_{1,k}(t,t+1)] = r$. We make a similar assumption for the individual class-1 traffic processes as for the total traffic process in Subsection 7.1 (Assumption 7.1):

Assumption 7.2. *For all $\phi > 0$ and $\psi > 0$, $k = 1,\dots K$,*

$$\mathbb{P}\left\{\sup_{t\ge 0}\{A_{1,k}(-t,0) - (r+\psi)t\} > \phi x\right\} = o(\mathbb{P}\{B^r > x\}), \qquad \text{as } x\to\infty.$$

Assumption 7.2 serves to ensure that the likelihood that rate variations in traffic of individual class-1 users cause a large workload is asymptotically negligible compared to scenarios with a large class-2 user as described earlier.

Similar to $V_1(t)$, define $V_{1,k}^{\text{var}}(t) := \sup_{s\leq t}\{A_{1,k}(s,t) - C_{1,k}(s,t)\}$, where $C_{1,k}(s,t)$ denotes the total available service rate for streaming user k during the time interval $[s,t]$. Again, we added the superscript 'var' to indicate that the quantity corresponds to the scenario with variable-rate streaming sources. Note that $C_{1,k}(s,t) \geq \int_s^t 1/(K+N(u))\mathrm{d}u$ and also $\sum_{k=1}^K C_{1,k}(s,t) = C_1(s,t)$. The first relation holds with equality in case the streaming users always claim the full service rate available. However, we may allow for strict inequality in case several class-1 users do not always consume the service rate available and the unused surplus is redistributed among the other class-1 and class-2 users. Observe that the exact definition of $C_{1,k}(s,t)$ is not crucial in case $Kr < 1-\rho < (K+1)r$, because the workload of each class-1 user builds up in the presence of the large class-2 user, and each class-1 user will thus use its full service capacity.

Finally, denote the vectors $\boldsymbol{V_1}^{\text{var}} = (V_{1,1}^{\text{var}}, \cdots, V_{1,K}^{\text{var}})$, with $V_{1,k}^{\text{var}}$ the steady-state version of $V_{1,k}^{\text{var}}(t)$, and $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_K)$. Moreover, let $\alpha^* := \max \alpha_k$. Then, we may derive a similar upper bound as in Propositions 6.1 and 7.3.

Proposition 7.4. *(upper bound) Suppose that the processes $\{A_{1,k}(-t,0), t \geq 0\}$, $k = 1,\ldots,K$, satisfy Assumption 7.2. If $B(\cdot) \in \mathcal{R}_{-\nu}$ and $Kr < 1-\rho < (K+1)r$, then*

$$\limsup_{x\to\infty} \frac{\mathbb{P}\{\boldsymbol{V_1}^{\text{var}} > \boldsymbol{\alpha}x\}}{\mathbb{P}\{V_1^{\text{var}} > K\alpha^* x\}} = \limsup_{x\to\infty} \frac{\mathbb{P}\{\boldsymbol{V_1}^{\text{var}} > \boldsymbol{\alpha}x\}}{\frac{\rho}{1-\rho-Kr}\mathbb{P}\left\{B^r > \frac{K\alpha^* x\frac{1-\rho}{K+1}}{K(r-\frac{1-\rho}{K+1})}\right\}} \leq 1.$$

Proof. Let $k^* := \arg\max \alpha_k$, and note that

$$\mathbb{P}\{\boldsymbol{V_1}^{\text{var}} > \boldsymbol{\alpha}x\} \leq \mathbb{P}\left\{V_{1,k^*}^{\text{var}} > \alpha^* x\right\}.$$

Using a similar construction for streaming user k^* (i.e., the leaky bucket) as in the proof of Proposition 7.3, we obtain the following sample path relation

$$\begin{aligned}
V_{1,k^*}^{\text{var}}(t) &\leq \sup_{s\leq t}\{A_{1,k^*}(s,t) - \int_s^t \frac{1}{K+N^{\text{var}}(u)}\mathrm{d}u\} \\
&\leq \sup_{s\leq t}\{A_{1,k^*}(s,t) - (r+\psi)(t-s)\} + \sup_{s\leq t}\{(r+\psi)(t-s) - \int_s^t \frac{1}{K+N^{\text{var}}(u)}\mathrm{d}u\} \\
&\leq \sup_{s\leq t}\{A_{1,k^*}(s,t) - (r+\psi)(t-s)\} + V_1^{\text{cst},\psi}/K,
\end{aligned}$$

where we used the *permanent-customer* scenario as an upper bound in the final step. Combining the arguments above, the proof may be finished along similar lines as the proof of Proposition 7.3. □

For the lower bound, modifications to one of the proofs in Section 5 would imply that we have to keep track of all individual workloads and received amounts of services. In view of the exceedingly large amount of details and notational complexity, we present the next result as a conjecture:

Conjecture 7.5. *Suppose that the processes $\{A_{1,k}(-t,0), t \geq 0\}$, $k = 1,\ldots,K$, satisfy Assumption 7.2. If $B(\cdot) \in \mathcal{R}_{-\nu}$ and $Kr < 1-\rho < (K+1)r$, then*

$$\begin{aligned}
\mathbb{P}\{\boldsymbol{V_1}^{\text{var}} > \boldsymbol{\alpha}x\} &\sim \mathbb{P}\{V_1^{\text{var}} > K\alpha^* x\} \\
&\sim \frac{\rho}{1-\rho-Kr}\mathbb{P}\left\{B^r > \frac{x\frac{1-\rho}{K+1}}{K(r-\frac{1-\rho}{K+1})}\right\}.
\end{aligned}$$

Conjecture 7.5 implies that the asymptotic tail probability of the K-dimensional random vector $\boldsymbol{V}_{\mathbf{1}}^{\mathrm{var}}$ can be reduced to the tail probability of a 1-dimensional random variable B^r. In other words, we conclude that the workloads of the K individual class-1 users can only simultaneously grow large, requiring the presence of a large class-2 user.

8 Concluding remarks

We considered a bottleneck link shared by heavy-tailed TCP-controlled elastic flows and streaming sessions regulated by a TCP-friendly rate control protocol. We determined the asymptotic tail distribution of a possible shortfall in service received by the streaming users compared to a nominal service target. We showed that the distribution inherits the heavy-tailed behavior of the residual service requirement of an elastic flow.
In the case that the elastic flows arrive according to a Poisson process, we further derived bounds for performance measures for both classes of traffic by exploiting a relationship with the M/G/1 PS queue with permanent customers. In particular, we obtained bounds for the probability that the rate of the streaming applications falls below a given target rate, as well as for the delay and workload distributions of the elastic flows.
Besides the bounds provided by the M/G/1 PS queue with permanent customers, we also determined the exact delay asymptotics of the elastic flows, suggesting a certain dichotomy in the tail asymptotics, depending on whether the system is critically loaded or not.
The service deficit distribution of the streaming users was derived for critical load, i.e., an additional 'persistent' elastic flow would cause instability of the streaming class. In general, the most likely scenario for the class-1 workload to grow large involves the simultaneous presence of $l \geq 1$ large class-2 users, where $l := \min\left\{a \in \mathbb{N} : \frac{1-\rho}{K+a} < r\right\}$ is the number of 'persistent' elastic flows required to cause instability of the streaming class (class 1). This gives rise to the following conjecture:

Conjecture 8.1. *If* $B(\cdot) \in \mathcal{R}_{-\nu}$ *and* $\rho + Kr < 1$, *then*

$$\mathbb{P}\{V_1 > x\} = O(\mathbb{P}\{B^r > x\}^l), \qquad \text{as } x \to \infty.$$

Guillemin *et al.* [17] obtained similar asymptotics for the distribution of the available amount of service during an interval of length x in PS queues. However, obtaining exact asymptotics is a difficult task in this case as witnessed by [30].
Several other interesting issues remain for further research, e.g., transient performance measures, scenarios with finite buffers and/or dynamic populations of streaming sessions, and the performance impact of oscillations, inaccuracies, and delays in the estimation of the fair bandwidth share.

A Proof of (12) for constant-rate streaming traffic

As mentioned previously, the asymptotic relation (12) plays a key role in our proofs, and is valid for several model extensions. To keep the presentation transparent, we first prove this relation in the next proposition for the case of constant-rate streaming traffic (assuming critical load). Appendix D extends this result to variable-rate streaming traffic (as well as work-conserving, but possibly non-critically loaded, scenarios).

Proposition A.1. *If $B(\cdot) \in \mathcal{R}_{-\nu}$ and $Kr < 1 - \rho < (K+1)r$, then,*

$$\mathbb{P}\left\{V_1 + V_2 > x\right\} \sim \mathbb{P}\left\{V_2^{1-Kr} > x\right\}.$$

This asymptotic relation also holds when $V_1 + V_2$ and V_2^{1-Kr} represent the workloads embedded at class-2 arrival epochs rather than at arbitrary instants.

Proof. First observe that

$$\begin{aligned}
\mathbb{P}\left\{V_1(0) + V_2(0) > x\right\} &\geq \mathbb{P}\left\{\sup_{t \geq 0}\{A_1(-t,0) + A_2(-t,0) - t\} > x\right\} \\
&= \mathbb{P}\left\{V_2^{1-Kr} > x\right\}.
\end{aligned}$$

It remains to be shown that

$$\limsup_{x\to\infty} \frac{\mathbb{P}\left\{V_1 + V_2 > x\right\}}{\mathbb{P}\left\{V_2^{1-Kr} > x\right\}} \leq 1. \tag{36}$$

As defined in Section 6, $s^* := \inf\{t > 0 : V_1(-t) = 0\}$ is the last epoch before time 0 that the class-1 workload was zero. Hence, $V_1(t) > 0$ for $t \in (-s^*, 0]$, implying that the system operates at the full service rate during that interval. Now, as described in Section 4, the idea of the proof is that a large total workload is most likely caused by the arrival of a large class-2 user. In particular, the class-1 workload starts to build in the presence of a persistent class-2 user, and it may be shown that time s^* is close to the arrival epoch of the large user.

More formally, we split the class-2 workload at time t into workloads contributed by users with initial service requirements smaller than (or equal to) ϵx ($V_{2,\leq\epsilon x}(t)$), and those with initial service requirements larger than ϵx ($V_{2,>\epsilon x}(t)$). Moreover, let $V^c_{2,\leq\epsilon x}(t)$, $V^c_{2,>\epsilon x}(t)$ be the workloads in an isolated queue fed by class-2 traffic of users with service requirements smaller than, larger than ϵx, respectively. Then, use (9), apply Observation 6.1 to bound $V_{2,\leq\epsilon x}(-s^*)$ and Lemma B.1 (stated below) to bound $V_{2,>\epsilon x}(-s^*)$:

$$\begin{aligned}
& V_1(0) + V_2(0) \\
= \;& V_1(-s^*) + V_{2,\leq\epsilon x}(-s^*) + V_{2,>\epsilon x}(-s^*) + A_1(-s^*,0) + A_{2,\leq\epsilon x}(-s^*,0) + A_{2,>\epsilon x}(-s^*,0) - s^* \\
\leq \;& 0 + M\epsilon x + A_{2,\leq\epsilon x}(-s^*,0) - (\rho+\delta)s^* + V_{2,>\epsilon x}^{1-Kr-\rho-\delta}(-s^*) + A_{2,>\epsilon x}(-s^*,0) \\
& -(1 - Kr - \rho - \delta)s^* \\
\leq \;& M\epsilon x + V_{2,\leq\epsilon x}^{\rho+\delta}(0) + V_{2,>\epsilon x}^{1-Kr-\rho-\delta}(0).
\end{aligned}$$

Converting this sample-path relation into a probabilistic upper bound gives (take $\epsilon < 1/M$)

$$\begin{aligned}
\mathbb{P}\left\{V_1 + V_2 > x\right\} &\leq \mathbb{P}\left\{V_{2,\leq\epsilon x}^{\rho+\delta}(0) + V_{2,>\epsilon x}^{1-Kr-\rho-\delta}(0) > (1 - M\epsilon)x\right\} \\
&\leq \mathbb{P}\left\{V_{2,\leq\epsilon x}^{\rho+\delta}(0) > \xi(1 - M\epsilon)x\right\} + \mathbb{P}\left\{V_{2,>\epsilon x}^{1-Kr-\rho-\delta}(0) > (1-\xi)(1 - M\epsilon)x\right\}.
\end{aligned}$$

The first term can be made sufficiently small for any fixed δ, ϵ, ξ, using similar arguments as in [9]. For the second term, we first apply Lemma B.2 (given below) and Theorem 4.2, and then use the fact that $B^r(\cdot) \in \mathcal{R}_{1-\nu}$, and let δ, ξ, $\epsilon \downarrow 0$.

Note that the above proof applies regardless of whether 0 is an arbitrary instant or a class-2 arrival epoch. □

B Technical lemmas

Lemma B.1. *For $1-\rho < (K+1)r$, $\epsilon > 0$, and $\delta > 0$,*

$$V_{2,>\epsilon x}(-s^*) \leq V^r_{2,>\epsilon x}(-s^*) \leq V^{1-Kr-\rho-\delta}_{2,>\epsilon x}(-s^*).$$

Proof. Denote by $u^* := \inf\{u \geq s^* : V_{2,>\epsilon x}(-u) = 0\}$ the last epoch before time $-s^*$ that no large class-2 user was present. Hence, $N_{>\epsilon x}(t) \geq 1$ for $t \in (-u^*, -s^*]$. Observe that the amount of service received by the large users during $(-u^*, -s^*]$ then satisfies

$$B_{2,>\epsilon x}(-u^*, -s^*) \geq \int_{-u^*}^{-s^*} N_{>\epsilon x}(t) c_1(t) \mathrm{d}t \geq \int_{-u^*}^{-s^*} c_1(t)\mathrm{d}t \geq r(u^* - s^*),$$

where $c_1(t)$ is the service rate of an individual streaming user at time t. Here, the final step follows from the fact that $V_1(-s^*) = 0$ and the service received during $(-u^*, -s^*]$ exceeds the amount of traffic generated. Using the above in the second step and (9) in the first and final one, gives

$$\begin{aligned}
V_{2,>\epsilon x}(-s^*) &= V_{2,>\epsilon x}(-u^*) + A_{2,>\epsilon x}(-u^*, -s^*) - B_{2,>\epsilon x}(-u^*, -s^*) \\
&\leq A_{2,>\epsilon x}(-u^*, -s^*) - r(u^* - s^*) \\
&\leq V^r_{2,>\epsilon x}(-u^*) + A_{2,>\epsilon x}(-u^*, -s^*) - r(u^* - s^*) \\
&\leq V^r_{2,>\epsilon x}(-s^*).
\end{aligned}$$

Finally, $V^r_{2,>\epsilon x}(-s^*) \leq V^{1-Kr-\rho-\delta}_{2,>\epsilon x}(-s^*)$ follows directly from $\delta \geq 0$ and $1-\rho < (K+1)r$. □

Lemma B.2. *For all $c, \epsilon > 0$,*

$$\mathbb{P}\left\{V^c_{2,>\epsilon x} > x\right\} \leq (1+o(1))\frac{\rho}{c}\mathbb{P}\{B^r > x\} \sim \mathbb{P}\left\{V_2^{c+\rho} > x\right\}, \qquad \text{as } x \to \infty.$$

Proof. Fix L, $0 < L < \infty$, and consider an isolated system of capacity c, where only class-2 users with service requirements larger than L are admitted. The system load then equals $\rho_L := \lambda \mathbb{P}\{B > L\} \mathbb{E}[B|B > L]$. Moreover, using Theorem 4.2 (take L large enough, such that $\rho_L < c$), yields

$$\mathbb{P}\left\{V^c_{2,>L} > x\right\} \sim \frac{\rho_L}{c-\rho_L}\mathbb{P}\{B^r_{>L} > x\}.$$

For $x > L$, the probability on the right-hand side may be rewritten as follows

$$\begin{aligned}
\mathbb{P}\{B^r_{>L} > x\} &= \frac{1}{\mathbb{E}[B|B>L]}\int_x^\infty \mathbb{P}\{B > y | B > L\}\mathrm{d}y \\
&= \frac{1}{\mathbb{E}[B|B>L]}\int_x^\infty \frac{\mathbb{P}\{B > y\}}{\mathbb{P}\{B > L\}}\mathrm{d}y \\
&= \frac{\mathbb{P}\{B^r > x\}\mathbb{E}B}{\mathbb{P}\{B > L\}\mathbb{E}[B|B>L]} = \frac{\rho}{\rho_L}\mathbb{P}\{B^r > x\}.
\end{aligned}$$

Combining the above gives

$$\mathbb{P}\left\{V^c_{2,>L} > x\right\} \sim \frac{\rho}{c-\rho_L}\mathbb{P}\{B^r > x\}. \tag{37}$$

Now, observe that for $x \geq L/\epsilon$, we have $V^c_{2,>\epsilon x}(t) \leq V^c_{2,>L}(t)$, so that the first part of the result may be obtained from (37), letting $L \to \infty$, and observing that $\rho_L \to 0$ as $L \to \infty$. The second part follows from Theorem 4.2. □

Lemma B.3. *For all $k \in \mathbb{N}$, $\kappa > 0$ (fixed), and $\gamma > 0$,*

$$\mathbb{P}\left\{N_{>\kappa x}(-\gamma x, 0) \geq k\right\} = O(\mathbb{P}\left\{B^r > x\right\}^k), \qquad \text{as } x \to \infty.$$

Proof. Consider the time interval $(-t, 0)$ and denote by $T_{>\kappa x}(n)$ the interarrival time between the $(n-1)$-th and n-th user arrival after time $-t$ with service requirement larger than κx (with the natural amendment that the 0-th arrival represents the last arrival before time $-t$ with service requirement larger than κx). Also, let $T^r_{>\kappa x}(n)$ denote its residual interarrival time and let τ be an arbitrary class-2 arrival epoch. We first prove the lemma for $k = 1$. Note that

$$\mathbb{P}\left\{N_{>\kappa x}(-t, 0) \geq 1\right\} \leq \mathbb{E}[N_{>\kappa x}(-t, 0)] = \mathbb{E}[N(-t, 0)]\mathbb{P}\left\{B > \kappa x\right\} = \lambda t \mathbb{P}\left\{B > \kappa x\right\}.$$

In the final step we use that $-t$ is an arbitrary time instant, so that $N(-t, 0)$ is a stationary renewal process [**?**]. The statement of the lemma now follows for $k = 1$ by taking $t = \gamma x$ and using the fact that $B(\cdot)$ is regularly varying.
To extend this result to $k \geq 2$, note that, for all n,

$$\mathbb{P}\left\{T_{>\kappa x}(n) \leq t\right\} = \mathbb{P}\left\{N_{>\kappa x}(\tau, \tau + t) \geq 1\right\} \leq \mathbb{E}[N_{>\kappa x}(\tau, \tau+t)] = \mathbb{E}[N(\tau, \tau+t)]\mathbb{P}\left\{B > \kappa x\right\}.$$

By the Elementary Renewal Theorem [**?**], $\frac{1}{t}\mathbb{E}[N(\tau, \tau + t)] \to \lambda$ as $t \to \infty$, so that for any $\delta > 0$ there exists a $\bar{t}$ such that $\mathbb{E}[N(\tau, \tau + t)] \leq (\lambda + \delta)t$ for all $t \geq \bar{t}$.
Note that the following two events are equivalent for $k \geq 1$ (where we define the empty sum equal to 0 in case $k = 1$).

$$\{N_{>\kappa x}(-t, 0) \geq k\} = \{T^r_{>\kappa x}(1) + \sum_{n=2}^{k} T_{>\kappa x}(n) \leq t\}.$$

Thus, for $k \geq 2$ and $t \geq \bar{t}$,

$$\begin{aligned}\mathbb{P}\left\{N_{>\kappa x}(-t, 0) \geq k\right\} &\leq \mathbb{P}\left\{T^r_{>\kappa x}(1) + \sum_{n=2}^{k-1} T_{>\kappa x}(n) \leq t\right\}\mathbb{P}\left\{T_{>\kappa x}(k) \leq t\right\} \\ &\leq \mathbb{P}\left\{N_{>\kappa x}(-t, 0) \geq k - 1\right\}(\lambda + \delta)t\mathbb{P}\left\{B > \kappa x\right\}.\end{aligned}$$

By induction on k we obtain, for $k \geq 2$ and $t \geq \bar{t}$,

$$\mathbb{P}\left\{N_{>\kappa x}(-t, 0) \geq k\right\} \leq ((\lambda + \delta)t\mathbb{P}\left\{B > \kappa x\right\})^k.$$

Again, by taking $t = \gamma x$ (for large enough x) and using the fact that $B(\cdot)$ is regularly varying, the lemma follows. □

In case the class-2 users arrive according to a Poisson process, Lemma B.3 can also be shown more directly. The crucial observation is that the number of class-2 arrivals with a service requirement larger than κx also follows a Poisson process, however with parameter $\lambda\mathbb{P}\left\{B > \kappa x\right\}$. Using the Poisson distribution function and taking the sum of a geometric series then completes the proof.

Lemma B.4. *There exists a $\kappa^* > 0$ such that for all $\kappa \in (0, \kappa^*]$,*

$$\mathbb{P}\left\{\sup_{0\le s\le \gamma x} \{A_2(-s,0) - (\rho+\delta)s\} > \epsilon x \mid N_{>\kappa x}(-\gamma x, 0) = 0\right\} = o(\mathbb{P}\{B^r > x\}).$$

Proof. Denote the interarrival time between the $(n-1)$-th and n-th user by T_n, and the service requirement of the n-th user by B_n. Let $S_n := X_1 + \ldots + X_n$ be a random walk with step sizes $X_m := B_m - (\rho+\delta)T_m$, with $\delta > 0$. Since $\rho = \mathbb{E}B_m/\mathbb{E}T_m$, we have $\mathbb{E}X_m < 0$, i.e., the random walk has negative drift. Observe that by the saw-tooth nature of the process $\{A_2(-s,0) - (\rho+\delta)s\}$, the process attains a local maximum at epochs right after a jump, thus,

$$\sup_{0\le s\le \gamma x} \{A_2(-s,0) - (\rho+\delta)s\} \le B_1 + \sup_{1\le n\le N(-\gamma x,0)} S_n.$$

Then, conditioning on the total number of class-2 arrivals in $(-\gamma x, 0)$ yields

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{0\le s\le \gamma x} \{A_2(-s,0) - (\rho+\delta)s\} > \epsilon x \mid N_{>\kappa x}(-\gamma x, 0) = 0\right\} \\
= \;& \sum_{n=0}^{\infty} \mathbb{P}\left\{\sup_{0\le s\le \gamma x} \{A_2(-s,0) - (\rho+\delta)s\} > \epsilon x \mid N_{>\kappa x}(-\gamma x, 0) = 0; N(-\gamma x,0) = n\right\} \\
&\times \mathbb{P}\{N(-\gamma x, 0) = n\} \\
\le \;& \sum_{n=0}^{\bar{M}x} \mathbb{P}\left\{B_1 + \sup_{0\le m\le n}\left\{\sum_{i=1}^{m} X_i\right\} > \epsilon x \mid X_i < \kappa x, i = 1,\ldots,n\right\} \mathbb{P}\{N(-\gamma x,0) = n\} \\
&+ \sum_{n=\bar{M}x+1}^{\infty} \mathbb{P}\{N(-\gamma x,0) = n\} \\
\le \;& \max_{0\le n\le \bar{M}x} \mathbb{P}\left\{\sup_{0\le m\le n} S_m > (\epsilon-\kappa)x \mid X_i < \kappa x, i = 1,\ldots,n\right\} + \mathbb{P}\left\{N(-\gamma x,0) > \bar{M}x\right\} \\
\le \;& \mathbb{P}\left\{\sup_{0\le m\le \bar{M}x} S_m > (\epsilon-\kappa)x \mid X_i < \kappa x, i = 1,\ldots,n\right\} + \mathbb{P}\left\{N(-\gamma x,0) > \bar{M}x\right\}, \quad (38)
\end{aligned}$$

where the third inequality follows from the fact that $B_1 \le \epsilon x$. The second term of (38) decays exponentially fast in x when $\bar{M} > \lambda\gamma$. The first term may be rewritten as follows:

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{0\le m\le \bar{M}x} S_m > (\epsilon-\kappa)x \mid X_i < \kappa x, i = 1,\ldots,n\right\} \\
\le \;& \sum_{m=0}^{\bar{M}x} \mathbb{P}\{S_m > (\epsilon-\kappa)x \mid X_i < \kappa x, i = 1,\ldots,n\}.
\end{aligned}$$

This can be made sufficiently small by employing a powerful lemma of Resnick & Samorodnitsky [28]. According to this lemma, there exists a $\kappa^* > 0$ and a function $\phi(\cdot) \in \mathcal{R}_{-\alpha}$, with $\alpha > \nu$, such that for all $\kappa \in (0, \kappa^*]$ the first term of (38) can be bounded by $\bar{M}x\phi(x)$. Take $\phi(x) = x^{-1-\zeta}\mathbb{P}\{B^r > x\}$, with $\zeta = \alpha - \nu$, and note that $\bar{M}x\phi(x) = o(\mathbb{P}\{B^r > x\})$ to complete the proof. ☐

C Proof of Proposition 3.1

Proposition 3.1 *If $B(\cdot) \in \mathcal{R}_{-\nu}$ and either $(K+1)r > 1-\rho$ or $C_2(t) \equiv \frac{N(t)}{K+N(t)}$ or both, then*

$$\mathbb{P}\{S_2 > x\} \sim \mathbb{P}\left\{B > \frac{(1-\rho)x}{K+1}\right\}. \tag{39}$$

In contrast, if $(K+1)r < 1-\rho$ and $C_2(s,t) \equiv t - s - B_1(s,t)$, then

$$\mathbb{P}\{S_2 > x\} \sim \mathbb{P}\{B > (1-\rho-Kr)x\}. \tag{40}$$

Proof. First, the case $C_2(t) \equiv \frac{N(t)}{K+N(t)}$ follows directly from [17]. This result also directly provides the desired upper bound in case the system is critically loaded, i.e., $(K+1)r > 1-\rho$. The lower bound for (39) and the proof of (40) are somewhat similar to proofs of delay asymptotics in [8, 11, 17].

Let B_0 be the service requirement of a class-2 user arriving at time 0 and denote by S_0 its sojourn time. Also, let $B_0(0,t)$ be the amount of service received during $(0,t]$ if it had an infinite service requirement. Now, observe that an actual user arriving at time 0 would receive the same amount of service $B_0(0,t)$ if it is still present at time t. Thus, assume that at time 0 a persistent class-2 user arrives. Then,

$$\mathbb{P}\{S_0 > t\} = \mathbb{P}\{B_0 > B_0(0,t)\}. \tag{41}$$

For conciseness, we now first give the proof of (40) and then provide the lower bound for (39).

Proof of (40). We apply the framework developed in [11, 17]. In particular, we show that Assumptions (A-2) and (A-3) in [17] are satisfied. For Assumption (A-2), use (8) and (9):

$$B_0(0,t) = t + V_1(t) + V_2(t) - Krt - A_2(0,t) - V_1(0) - V_2(0).$$

Because the system is stable, both $(V_1(t)+V_2(t))/t \to 0$ and $(V_1(0)+V_2(0))/t \to 0$ when $t \to \infty$. Moreover, since $A_2(0,t)/t \to \rho$ for $t \to \infty$, we have

$$\lim_{t\to\infty} \frac{B_0(0,t)}{t} = 1-\rho-Kr,$$

giving Assumption (A-2). For Assumption (A-3), observe that $B_0(0,t) \geq \int_0^t 1/(K+1+N_{(K+1)}(u))\mathrm{d}u$. Thus, from the proof of [17, Theorem 3] (take $f(n) = \frac{1}{K+1+n}$), it follows that there exists a finite constant $D > 0$, such that

$$\mathbb{P}\{B_0(0,t) \leq Dt\} \leq \mathbb{P}\left\{\int_0^t \frac{1}{K+1+N_{(K+1)}(u)}\mathrm{d}u \leq Dt\right\} = o(\mathbb{P}\{B > x\}).$$

Since Assumptions (A-1)-(A-3) are satisfied, we may apply [17, Theorem 1] to obtain (40).

Lower bound for (39). Let $B^-_{2,\leq\kappa t}(s,t)$ ($B^-_{2,>\kappa t}(s,t)$) be the amount of service received by class-2 users with initial service requirements smaller than (larger than) κt, excluding the persistent class-2 user. Also, let $s_t := \sup\{0 \leq u \leq t : V_1(u) = 0\}$ be the last epoch before time t that the class-1 workload was zero. Recall that $V_2^c(t) = \sup_{0\leq s\leq t}\{A_2(s,t) - c(t-s)\}$. Using (8) and (9) in addition to Observation 6.1, we deduce

$$\begin{aligned} B_0(s_t,t) + B_1(s_t,t) + B^-_{2,>\kappa t}(s_t,t) &= t - s_t - B^-_{2,\leq\kappa t}(s_t,t) \\ &\geq t - s_t - A_{2,\leq\kappa t}(s_t,t) - V^-_{2,\leq\kappa t}(s_t) \\ &\geq (1-\rho-\epsilon)(t-s_t) + (\rho+\epsilon)(t-s_t) - A_2(s_t,t) - V^-_{2,\leq\kappa t}(s_t) \\ &\geq (1-\rho-\epsilon)(t-s_t) - V_2^{\rho+\epsilon}(t) - M\kappa t, \end{aligned}$$

where $A_{2,\leq\kappa t}(s,t)$ is the amount of "small" class-2 traffic generated during $(s,t]$ (see also Subsection 5.2), and $V^-_{2,\leq\kappa t}(s)$ is the workload at time s associated with "small" class-2 users, excluding the persistent user. Because class 1 uses the total available capacity during $(s_t,t]$, we have $B_1(s_t,t) = KB_0(s_t,t)$. Also, $V_1(s_t) = 0$ implies $B_1(0,s_t) \geq Krs_t$. Combining the above, and taking $\epsilon > 0$ sufficiently small, yields

$$\begin{aligned} B_1(0,t) + B^-_{2,>\kappa t}(0,t) &\geq Krs_t + \frac{K}{K+1}[(1-\rho-\epsilon)(t-s_t) - V_2^{\rho+\epsilon}(t) - M\kappa t] \\ &\geq \frac{K}{K+1}[(1-\rho-\epsilon)t - V_2^{\rho+\epsilon}(t) - M\kappa t]. \end{aligned} \tag{42}$$

Now, applying (8) and (9), we obtain

$$\begin{aligned} B_0(0,t) &\leq t - B_1(0,t) - B^-_{2,>\kappa t}(0,t) - B^-_{2,\leq\kappa t}(0,t) \\ &\leq (1-\rho+\epsilon)t - \frac{K}{K+1}[(1-\rho-\epsilon)t - V_2^{\rho+\epsilon}(t) - M\kappa t] + V^-_{2,\leq\kappa t}(t) + (\rho-\epsilon)t - A_{2,\leq\kappa t}(0,t). \end{aligned}$$

Moreover, observe that $V^-_{2,\leq\kappa t}(t) \leq V_{(K+1)}(t)$. Using these sample-path arguments and (41) yields

$$\begin{aligned} &\mathbb{P}\{S_0 > t\} \\ \geq\ &\mathbb{P}\left\{B_0 > \frac{1-\rho+(2K+1)\epsilon}{K+1}t + \frac{K}{K+1}[V_2^{\rho+\epsilon}(t) + M\kappa t] + V^-_{2,\leq\kappa t}(t) + (\rho-\epsilon)t - A_{2,\leq\kappa t}(0,t) \leq \epsilon t\right\} \\ \geq\ &\mathbb{P}\left\{B_0 > \frac{1-\rho+(4K+2)\epsilon+KM\kappa}{K+1}t\right\} \\ &\mathbb{P}\left\{\frac{K}{K+1}V_2^{\rho+\epsilon}(t) + V_{(K+1)}(t) + (\rho-\epsilon)t - A_{2,\leq\kappa t}(0,t) \leq \frac{2K+1}{K+1}\epsilon t\right\}. \end{aligned} \tag{43}$$

Note that $V_2^{\rho+\epsilon}(t)$, $V_{(K+1)}(t)$, and $A_{2,\leq\kappa t}(0,t)$ are not independent. However, the probability in (43) can be bounded from below by

$$\mathbb{P}\{A_{2,\leq\kappa t}(0,t) \geq (\rho-\epsilon)t\} - \mathbb{P}\left\{V_2^{\rho+\epsilon} \geq \epsilon t\right\} - \mathbb{P}\{V_{(K+1)}(t) \geq \epsilon t\}. \tag{44}$$

The first term of (44) may be treated as in Subsection 5.2, giving $\mathbb{P}\{A_{2,\leq\kappa t}(0,t) \geq (\rho-\epsilon)t\} \to 1$ as $t \to \infty$. For the second term, we note that $V_2^{\rho+\epsilon}$ has a non-defective distribution. Moreover, we use the fact that a system with $K+1$ permanent customers also has a proper limiting distribution to handle the third probability in (44) (see also Subsection 5.1).
Now, use the fact that $B(\cdot) \in \mathcal{R}_{-\nu}$ and let $\epsilon, \kappa \downarrow 0$ to obtain the lower bound for (40), which completes the proof. □

D Proof of (12) for variable-rate streaming traffic

We now extend Proposition A.1 (relation (12)) to variable-rate streaming traffic. In fact, a slightly stronger result is needed in the proof of the lower bound of Theorem 7.1. However, $\mathbb{P}\{V_1^{\text{var}} + V_2^{\text{var}} > x\} \geq (1+o(1))\mathbb{P}\left\{V_2^{1-Kr} > x\right\}$ is a direct consequence of the proof, and the following proposition may be of independent interest. It also shows that the asymptotic equivalence holds under non-critical load if the system is work-conserving.

Proposition D.1. *Suppose that the process $\{A_1(-t,0), t \geq 0\}$ satisfies Assumption 7.1 and $\rho + Kr < 1$. If $B(\cdot) \in \mathcal{R}_{-\nu}$ and one of the two following conditions is satisfied*

(i) the system is critically loaded, i.e., $1 - \rho < (K+1)r$;

(ii) the system is work-conserving, i.e., $C_2(s,t) \equiv t - s - B_1(s,t)$;

then

$$\mathbb{P}\{V_1^{\text{var}} + V_2^{\text{var}} > x\} \sim \mathbb{P}\left\{V_2^{1-Kr} > x\right\}.$$

This asymptotic relation also holds when $V_1 + V_2$ and V_2^{1-Kr} represent the workloads embedded at class-2 arrival epochs rather than at arbitrary instants.

Proof. The proofs again involve lower and upper bounds which asymptotically coincide. The lower bound is the same for both cases (i) and (ii).
(Lower bound) In fact, we will prove a slightly stronger result. Define $\overleftarrow{U}_1^c(0) := \sup_{t\geq 0}\{ct - A_1(-t,0)\}$ and recall that $\overrightarrow{U}_1^c(0) = \sup_{t\geq 0}\{ct - A_1(0,t)\}$. We show that

$$\liminf_{x\to\infty} \frac{\mathbb{P}\left\{V_1^{\text{var}}(0) + V_2^{\text{var}}(0) > x; \overrightarrow{U}_1^{K(r-\psi)}(0) \leq \phi x\right\}}{\mathbb{P}\left\{V_2^{1-Kr} > x\right\}} \geq 1. \tag{45}$$

Using the work-conserving scenario as a lower bound in addition to (10), we have for any $\xi > 0$,

$$\begin{aligned} &\mathbb{P}\left\{V_1^{\text{var}}(0) + V_2^{\text{var}}(0) > x; \overrightarrow{U}_1^{K(r-\psi)}(0) \leq \phi x\right\} \\ \geq\ &\mathbb{P}\left\{\sup_{t\geq 0}\{A_1(-t,0) - K(r-\psi)t + A_2(-t,0) - (1-K(r-\psi))t\} > x; \overrightarrow{U}_1^{K(r-\psi)}(0) \leq \phi x\right\} \\ \geq\ &\mathbb{P}\left\{V_2^{1-K(r-\psi)}(0) - \overleftarrow{U}_1^{K(r-\psi)}(0) > x; \overrightarrow{U}_1^{K(r-\psi)}(0) \leq \phi x\right\} \\ \geq\ &\mathbb{P}\left\{V_2^{1-K(r-\psi)}(0) \geq (1+\xi)x\right\}\mathbb{P}\left\{\overleftarrow{U}_1^{K(r-\psi)}(0) \leq \xi x; \overrightarrow{U}_1^{K(r-\psi)}(0) \leq \phi x\right\}. \end{aligned}$$

Note that

$$\mathbb{P}\left\{\overleftarrow{U}_1^{K(r-\psi)}(0) \leq \xi x; \overrightarrow{U}_1^{K(r-\psi)}(0) \leq \phi x\right\} \geq \mathbb{P}\left\{\overleftarrow{U}_1^{K(r-\psi)}(0) \leq \xi x\right\} - \mathbb{P}\left\{\overrightarrow{U}_1^{K(r-\psi)}(0) \geq \phi x\right\}.$$

Because both $\overleftarrow{U}_1^{K(r-\psi)}(0)$ and $\overrightarrow{U}_1^{K(r-\psi)}(0)$ have a proper distribution, $\mathbb{P}\left\{\overleftarrow{U}_1^{K(r-\psi)}(0) \leq \xi x\right\} \to 1$ and $\mathbb{P}\left\{\overrightarrow{U}_1^{K(r-\psi)}(0) \geq \phi x\right\} \to 0$ as $x \to \infty$ (see also (32)). Hence, it follows that

$$\liminf_{x\to\infty} \frac{\mathbb{P}\left\{V_1^{\text{var}}(0) + V_2^{\text{var}}(0) > x; \overrightarrow{U}_1^{K(r-\psi)}(0) \leq \phi x\right\}}{\mathbb{P}\left\{V_2^{1-K(r-\psi)} > (1+\xi)x\right\}} \geq 1.$$

Finally, let $\xi, \psi, \phi \downarrow 0$ and use Theorem 4.2 and the fact that $B^r(\cdot) \in \mathcal{R}_{1-\nu}$ to obtain (45). The lower bound is a direct consequence.
(Upper bound for part (i)) We now show that for a critically loaded system

$$\limsup_{x\to\infty} \frac{\mathbb{P}\left\{V_1^{\text{var}} + V_2^{\text{var}} > x\right\}}{\mathbb{P}\left\{V_2^{1-Kr} > x\right\}} \leq 1. \tag{46}$$

To do so, we apply the leaky-bucket idea of Section 7. Recall that in the reference system, class 1 generates traffic at constant rate $K(r+\psi)$, and class 2 receives service at rate $N_{(K)}(t)/(K + N_{(K)}(t))$, independently of class 1. Note that $V_2^{\text{var}}(t) \leq V_{(K)}(t) = V_2^{\text{cst},\psi}(t)$, with $V_{(K)}(t)$ the workload at time t in an isolated queue fed by class 2 with K permanent customers, and $V_2^{\text{cst},\psi}(t)$ the class-2 workload at time t in the reference system. Thus, combining the above with (35) yields

$$V_1^{\text{var}}(t) + V_2^{\text{var}}(t) \leq V_1^{K(r+\psi)}(t) + V_1^{\text{cst},\psi}(t) + V_2^{\text{cst},\psi}(t).$$

Converting this sample-path relation into a probabilistic upper bound gives

$$\mathbb{P}\left\{V_1^{\text{var}} + V_2^{\text{var}} > x\right\} \leq \mathbb{P}\left\{V_1^{K(r+\psi)} > \xi x\right\} + \mathbb{P}\left\{V_1^{\text{cst},\psi} + V_2^{\text{cst},\psi} > (1-\xi)x\right\}.$$

Again, the first term can be controlled by Assumption 7.1. For the second term, apply Proposition A.1 and Theorem 4.2, use the fact that $B^r(\cdot) \in \mathcal{R}_{1-\nu}$, and then let $\psi, \xi \downarrow 0$.
(Upper bound for part (ii)) It remains to be shown that (46) holds if the system is work-conserving. Using sample-path arguments, we have that $V_1^{\text{var}}(t) + V_2^{\text{var}}(t) \leq V_1^{K(r+\psi)}(t) + V_2^{1-K(r+\psi)}(t)$, so that, for any $\phi \in (0,1)$,

$$\mathbb{P}\left\{V_1^{\text{var}} + V_2^{\text{var}} > x\right\} \leq \mathbb{P}\left\{V_1^{K(r+\psi)} > \phi x\right\} + \mathbb{P}\left\{V_2^{1-K(r+\psi)} > (1-\phi)x\right\}.$$

It follows from Assumption 7.1, Theorem 4.2, and the fact that $B^r(\cdot) \in \mathcal{R}_{1-\nu}$ that

$$\mathbb{P}\left\{V_1^{K(r+\psi)} > \phi x\right\} = o(\mathbb{P}\left\{V_2^{1-K(r+\psi)} > (1-\phi)x\right\}),$$

as $x \to \infty$. Thus,

$$\limsup_{x\to\infty} \frac{\mathbb{P}\left\{V_1^{\text{var}} + V_2^{\text{var}} > x\right\}}{\mathbb{P}\left\{V_2^{1-K(r+\psi)} > (1-\phi)x\right\}} \leq 1.$$

Finally, let ψ, $\phi \downarrow 0$ and use Theorem 4.2 and the fact that $B^r(\cdot) \in \mathcal{R}_{1-\nu}$ to obtain (46). Note that the above proof applies regardless of whether 0 is an arbitrary instant or a class-2 arrival epoch. □

References

[1] Agrawal, R., Makowski, A.M., Nain, Ph. (1999). On a reduced load equivalence for fluid queues under subexponentiality. *Queueing Systems* **33**, 5–41.

[2] Asmussen, S. (2003). *Applied Probability and Queues.* Springer, New York, USA.

[3] Asmussen, S., Schmidli, H., Schmidt, V. (1999). Tail probabilities for non-standard risk and queueing processes with subexponential jumps. *Adv. Appl. Prob.* **31**, 422–447.

[4] Ben Fredj, S., Bonald, T., Proutière, A., Régnié, G., Roberts, J.W. (2001). Statistical bandwidth sharing: a study of congestion at the flow level. In: *Proc. ACM SIGCOMM 2001*, 111–122.

[5] Van den Berg, J.L., Boxma, O.J. (1991). The M/G/1 queue with processor sharing and its relation to a feedback queue. *Queueing Systems* **9**, 365–401.

[6] Bingham, N.H., Goldie, C., Teugels, J. (1987). *Regular Variation.* Cambridge University Press, Cambridge, UK.

[7] Bonald, T., Proutière, A. (2004). On performance bounds for the integration of elastic and adaptive streaming flows. In: *Proc. ACM Sigmetrics/Performance 2004*, 235–245.

[8] Borst, S.C., Núñez-Queija, R., Van Uitert, M.J.G. (2002). User-level performance of elastic traffic in a differentiated-services environment. *Perf. Eval.* **49**, Special Issue – Proc. Performance 2002 (Rome), 507–519.

[9] Borst, S.C., Zwart, A.P. (2001). Fluid queues with heavy-tailed $M/G/\infty$ input. *Math. Oper. Res.*, to appear

[10] Boxma, O.J., Foss, S.G., Lasgouttes, J.-M., Núñez-Queija, R. (2003). Waiting time asymptotics in the single server queue with service in random order. *Queueing Systems* **46**, 35–73.

[11] Boyer, J., Guillemin F., Robert, Ph., Zwart, A.P. (2003). Heavy tailed M/G/1-PS queues with impatience and admission control in packet networks. *Proc. Infocom 2003.*

[12] Cohen, J.W. (1973). Some results on regular variation for distributions in queueing and fluctuation theory. *J. Appl. Prob.* **10**, 343–353.

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

[14] Crovella, M., Bestavros, A. (1996). Self-similarity in World Wide Web traffic: evidence and possible causes. In: *Proc. ACM Sigmetrics '96*, 160–169.

[15] Dumas, V., Simonian, A. (2000). Asymptotic bounds for the fluid queue fed by subexponential on/off sources. *Adv. Appl. Prob.* **32**, 244–255.

[16] Floyd, S., Handley, M., Padhye, J., Widmer, J. (2000). Equation-based congestion control for unicast applications. In: *Proc. ACM SIGCOMM 2000*, 43–54.

[17] Guillemin, F., Robert, Ph., Zwart, A.P. (2004). Tail asymptotics for processor sharing queues. *Adv. Appl. Prob.* **36**, 525–543.

[18] Jelenković, P.R., Momčilović, P., Zwart, A.P. (2002). Reduced load equivalence under subexponentiality. *Queueing Systems* **46**, 97–112.

[19] Key, P.B., Massoulié, L., Bain, A., Kelly, F.P. (2003). A network flow model for mixtures of file transfers and streaming traffic. In: *Providing QoS in Heterogeneous Environments, Proc. ITC-18*, 1021–1030.

[20] Massoulié, L., Roberts, J.W. (1999). Bandwidth sharing: Objectives and algorithms. In: *Proc. IEEE Infocom '99*, 1395–1403.

[21] Mathis, M., Semke, J., Mahdavi, J., Ott, T.J. (1997). The macroscopic behavior of the TCP congestion avoidance algorithm. *Comp. Commun. Rev.* **27**, 67–82.

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

[23] Núñez-Queija, R. (2002). Queues with equally heavy sojourn time and service requirement distributions. *Ann. Oper. Res.* **113**, 101–117.

[24] Padhye, J., Firoiu, V., Towsley, D., Kurose, J. (2000). Modeling TCP Reno performance: a simple model and its empirical validation. *IEEE/ACM Trans. Netw.* **8**, 133–145.

[25] Padhye, J., Kurose, J., Towsley, D., Koodli, R. (1999). A model-based TCP-friendly rate control protocol. In: *Proc. IEEE NOSSDAV '99.*

[26] Pakes, A.G. (1975). On the tails of waiting-time distributions. *J. Appl. Prob.* **12**, 555–564.

[27] Rejaie, R., Handley, M., Estrin, D. (1999). RAP: an end-to-end rate-based congestion control mechanism for real-time streams in the Internet. In: *Proc. IEEE Infocom '99*, 1337–1346.

[28] Resnick, S., Samorodnitsky, G. (1999). Activity periods of an infinite server queue and performance of certain heavy tailed fluid queues. *Queueing Systems* **33**, 43–71.

[29] Stam, A.J. (1973). Regular variation of the tail of a subordinated probability distribution. *Adv. Appl. Prob.* **5**, 308–327.

[30] Zwart, A.P., Borst, S.C., Mandjes, M. (2001). Exact asymptotics for fluid queues fed by multiple heavy-tailed on-off flows. *Ann. Appl. Prob.* **14**, 903–957.