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 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. Since UDP does not

*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)

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, priorization 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 [14, 23, 25]. 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 [19, 22].

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 [12].

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 [3, 18, 20]. 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 most of the asymptotic results to hold.

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 furthermore examine the delay performance of the elastic flows.

In [17], 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 [6], 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 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. 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 [4] 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., $1 - B(x) \sim L(x)x^{-\nu}, \nu > 0$, 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 [5, Thm. 5.1.11] that $x\mathbb{P}\{B > x\} \sim (\nu - 1)\beta\,\mathbb{P}\{B^r > x\}$, so that $B^r(\cdot) \in \mathcal{R}_{1-\nu}$.

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 [4]. 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 $B(\cdot)$. 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. It may be checked that, in general, 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 [7]. 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.

The M/G/1 PS queue with permanent customers is a special case of the model studied in [11]. To obtain the model with K permanent customers, we take the service rate of each customer to be $f(n) = \frac{1}{K+n}$, when there are n customers. 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 [11],

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

We thus obtain an upper bound for the probability that the service rate of the streaming users is below a given desired rate s:

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

In the M/G/1 PS queue with m permanent customers, let $S_{(m)}$ be the steady-state sojourn time of a non-permanent elastic flow. As mentioned above, the delay (sojourn time) of elastic flows in our model (denoted by S_2) is stochastically dominated by $S_{(K)}$. The next proposition shows that the exact sojourn time asymptotics of S_2 depend on the assumptions on $C_2(s,t)$ in case $B_1(s,t) < C_1(s,t)$. Similar delay asymptotics were obtained in [7, 9, 15, 21].

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*

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

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

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

Proof. The asymptotics for $S_{(K)}$ (and, thus, for S_2 in the permanent-customer scenario) follow from [15]. 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 [15, Theorem 3] to obtain $\gamma^f = \frac{1-\rho}{m+1}$ and the desired result follows.

For the remainder of the proof we only provide an intuitive sketch. (We refer to [2] for a detailed proof.) 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 receive service roughly equal to their average input 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 streaming users will only claim an average service rate Kr, leaving and average service rate of $1-\rho-Kr$ for the large elastic flow. Otherwise, the large elastic flow is provided exactly with its fair share. □

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. We state a result for the M/G/1 PS queue with permanent customers and refer to [2] for a proof.

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*

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

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.

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*

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

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 may also serve as an outline for the construction of the lower bound, see [2] for details. In addition, we give an alternative interpretation, which provides the basis for the the lower bound in Section 5 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

$$B_1(s,t) + B_2(s,t) \leq t-s. \tag{2}$$

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

$$V_i(t) = V_i(s) + A_i(s,t) - B_i(s,t). \tag{3}$$

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,

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

In particular, in the constant-rate case,

$$\begin{aligned} V_1(t) + V_2(t) &= \sup_{s \leq t}\{Kr(t-s) + A_2(s,t) - (t-s)\} \\ &= \sup_{s \leq t}\{A_2(s,t) - (1-Kr)(t-s)\} \\ &= V_2^{1-Kr}(t), \end{aligned} \tag{4}$$

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 [10], and has been extended to subexponential distributions by Pakes [24].

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 (4) 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, (4) 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, 16, 26]):

$$\mathbb{P}\{V_1 + V_2 > x\} \sim \mathbb{P}\left\{V_2^{1-Kr} > x\right\}. \tag{5}$$

The 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 (5) 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 (2) may be rewritten as follows

$$B_1(s,t) + B_{\text{tag}}(s,t) + B_2^-(s,t) \leq t-s. \tag{6}$$

Suppose that the tagged user arrives at time $-y-z_0$, with $z_0 = \frac{x}{K(r-\frac{1-\rho}{K+1})}$, $B_{\rm tag} \geq x + (1-\rho-Kr)(y+z_0)$, 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 $B_{\rm tag}$, depending on whether the user is still present at time 0 or not. Using the inequality (6), the amount of service received by class 1 is approximately

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

Thus,

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

Using the above inequality and the identity relation (3), 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 asymptotically small probability of two or more "large" users)

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

which agrees with the right-hand side of (1).
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 [2]. 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 Section 5 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 (1).

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 obtain an asymptotic 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, based on the alternative characterization of the dominant scenario in Section 4 (Lemma 5.1). Next, 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 5.2).

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 $-s_0$, $s_0 := \inf\{0 \leq t \leq t_0 : N_{\leq \kappa x}^{(-t_0,-t]}(-t) \leq M_0\}$, be the last epoch before time 0 that there are less than M_0 "new small" class-2 users.

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

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

$$V_1(-t_0) + V_2(-t_0) \geq x(1+\gamma+M_0\kappa) - (Kr+\rho-1-\delta)t_0 \tag{7}$$

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

$$A_{2,\leq\kappa x}(-t_0, -s_0) \geq (\rho-\delta)(t_0 - s_0) - \gamma x \tag{8}$$

We first prove the next sample-path relation.

Lemma 5.1. *If the events (7) and (8) 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 equivalently $-u_0 \geq -s_0$). 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 (3), we obtain

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

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 (3), 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 (8) 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,

$$B_2^{(-t_0,0]}(-t_0,0) \geq (\rho-\delta)t_0 - \gamma x - M_0\kappa x. \tag{9}$$

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 (2) into

$$B_1(-t_0,0) \leq t_0 - B_2^{(-\infty,-t_0]}(-t_0,0) - B_2^{(-t_0,0]}(-t_0,0). \tag{10}$$

Using (3), (7), (9), and (10), 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 (2)

$$B_1(-t_0,0) \leq \frac{K}{K+1}[t_0 - B_2^{(-t_0,0]}(-t_0,0)]. \tag{11}$$

Now, combining (3), (9), and (11) 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 5.1 to establish the next asymptotic lower bound for the class-1 workload distribution.

Proposition 5.2. *(lower bound) If $B(\cdot) \in \mathcal{R}_{-\nu}$ and $Kr < 1-\rho < (K+1)r$, then*

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

Proof. First observe that the events (7) and (8) 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 $-t_0$, which is independent of $V_1(-T_0)+V_2(-T_0)$ as well. Using Lemma 5.1 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

$$\begin{aligned}
& \mathbb{P}\{V_1(-T_0) + V_2(-T_0) > x(1+\gamma+M_0\kappa+\epsilon) - (Kr+\rho-1-\delta)t_0\} \\
\sim\ & \frac{\rho}{1-\rho-Kr}\mathbb{P}\left\{B^r > \frac{x(1+\gamma+M_0\kappa)\frac{1-\rho+\delta}{K+1}}{K(r-\frac{1-\rho+\delta}{K+1})} + \epsilon x\right\}. \qquad (12)
\end{aligned}$$

Because τ_0 has a proper distribution, we have $\lim_{x\to\infty}\mathbb{P}\{\tau_0 > \epsilon x\} = 0$. For x sufficiently large, $\sup_{t\geq 0}\{(\rho-\delta)t - A_{2,\leq\kappa x}(0,t)\}$ also has a non-defective distribution yielding

$$\lim_{x\to\infty}\mathbb{P}\left\{\sup_{0\leq t\leq t_0}\{(\rho-\delta)(t_0-t) - A_{2,\leq\kappa x}(-t_0,-t)\} \leq \gamma x\right\} = 1.$$

Combining the above arguments and applying (12), we obtain

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

Use the fact that $B^r(\cdot) \in \mathcal{R}_{1-\nu}$ and let $\gamma,\delta,\epsilon,\kappa \downarrow 0$ to complete the proof. □

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 $s^* := \inf\{t \geq 0 : V_1(-t) = 0\}$, 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 rate 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 $\frac{K}{K+N(t)} \geq Kr$. Rewriting the inequality gives that $N(t) \leq M$, with $M := \lfloor \frac{1}{r} \rfloor - K$. $\diamond$

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*

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

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

$$\begin{aligned} & \mathbb{P}\{V_1(0) > x\} \\ \leq\ & \mathbb{P}\{V_1(-t_1) + V_2(-t_1) > x(1-\epsilon) - (Kr + \rho + \delta - 1)t_1\} && (13) \\ & + \mathbb{P}\{V_1(-t_1) + V_2(-t_1) \leq x(1-\epsilon) - (Kr + \rho + \delta - 1)t_1; V_1(0) > x\}\,. && (14) \end{aligned}$$

First, we determine the asymptotic behavior of (13). Then we show that (14) is negligible compared to (13) 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. First use Proposition A.1 and then Theorem 4.2 to obtain that (13) behaves as

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

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

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

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

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

To do so, we split (14) by distinguishing between 0, 1, and 2 or more large user arrivals during $(-t_1, 0]$, respectively. More specifically, write

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

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 (3) and the fact that the system is work-conserving during $(-s', 0]$, we have

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

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

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

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

Term II

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

$$\begin{aligned}
& \mathbb{P}\{\eta x < V_1(-t_1) + V_2(-t_1) < x(1-\epsilon) - (Kr+\rho+\delta-1)t_1; N_{>\kappa x}(-t_1, 0) = 1; V_1(0) > x\} \\
& + \mathbb{P}\{V_1(-t_1) + V_2(-t_1) < \eta x; N_{>\kappa x}(-t_1, 0) = 1; V_1(0) > x\}. \qquad (15)
\end{aligned}$$

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 (15) as a "combination of two unlikely events". Specifically, 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 (3) 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 (3),

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

Thus,

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

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 $x=\frac{t_1K(r-\frac{1-\rho-\delta}{K+1})}{1-\epsilon}$ yields

$$\begin{aligned}
&\mathbb{P}\left\{\zeta x+\sup_{0\leq s\leq t_1}\left\{Krs-\frac{K(s-A_2^-(-s,0)-\zeta x)}{K+1}\right\}+\zeta x>x\mid N_{>\kappa x}(-t_1,0)=1\right\}\\
=\;&\mathbb{P}\left\{\sup_{0\leq s\leq t_1}\left\{Krs-\frac{K(s-A_2^-(-s,0))}{K+1}\right\}>x(1-\frac{3K+1}{K+1}\zeta)+\frac{K}{K+1}\zeta x\mid N_{>\kappa x}(-t_1,0)=1\right\}\\
\leq\;&\mathbb{P}\left\{\sup_{0\leq s\leq t_1}\left\{Krs-\frac{K(s-A_2^-(-s,0))}{K+1}\right\}-t_1K(r-\frac{1-\rho-\delta}{K+1})>\frac{K}{K+1}\zeta x\mid N_{>\kappa x}(-t_1,0)=1\right\}\\
\leq\;&\mathbb{P}\left\{\sup_{0\leq s\leq t_1}\left\{A_2^-(-s,0)-(\rho+\delta)s\right\}>\zeta x\mid N_{>\kappa x}(-t_1,0)=1\right\}\\
\leq\;&\mathbb{P}\left\{\sup_{0\leq s\leq t_1}\left\{A_2(-s,0)-(\rho+\delta)s\right\}>\zeta x\mid N_{>\kappa x}(-t_1,0)=0\right\},
\end{aligned}$$

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 $III=o(\mathbb{P}\{B^r>x\})$.

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 with mean rate $\mathbb{E}[A_1(t,t+1)] = Kr$, provided that significant 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. More specifically, we assume that the class-1 traffic satisfies the following assumption:

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

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

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

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

where V_1^c denotes the steady-state workload in a system with service rate c fed by class 1 only. Assumption 7.1 is satisfied by a wide range of traffic processes, such as instantaneous bursts and On-Off sources (see [2] for details).
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 that the class-1 workload corresponds 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 only concerns modifications 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. A slightly more substantial modification is needed to show the asymptotic equivalence between $V_1^{\text{var}} + V_2^{\text{var}}$ and V_2^{1-Kr}. This is done in [2, Proposition D.1], where we extend relation (5) to the case of variable-rate class-1 traffic.
For the upper bound, the proof is based on a comparison with a leaky-bucket system. More specifically, we make the following comparison between the class-1 workload in the variable-rate scenario and that in the constant-rate scenario. 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 class-1 workload at the second resource is then bounded by $V_1^{\text{cst},\psi}(t) = \sup_{s\leq t}\{K(r+\psi)(t-s) - \int_s^t \frac{K}{K+N_{(K)}(u)}\text{d}u\}$, which corresponds to the *permanent-customer* scenario when the 'target rate'

per streaming user is constant and equal to $r+\psi$. $V_1^{\text{var}}(t)$ is then bounded by $V_1^{K(r+\psi)}(t) + V_1^{\text{cst},\psi}(t)$. This sample-path relation combined with Assumption 7.1, Theorem 4.1, and the fact that $B^r(\cdot) \in \mathcal{R}_{1-\nu}$, serves as the basis for a rigorous proof (see [2]).

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 the 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 requirements of the elastic flows. 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 $l := \min\left\{a : \frac{1-\rho}{K+a} < r\right\}$ 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*

$$\mathbb{P}\{V_1 > x\} = O(\mathbb{P}\{B^r > x\}^l).$$

Guillemin *et al.* [15] obtained similar asymptotics for the available amount of service during $(0,x)$ in PS queues. However, obtaining exact asymptotics is a difficult task in this case as witnessed by [26].
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 (5)

The asymptotic relation (5) plays a key role in our proofs, and is valid for several model extensions. To keep the presentation transparent, we only prove this relation here for the case of constant-rate streaming traffic, assuming the system is critically loaded, i.e., $1-\rho < (K+1)r$. Proposition D.1 in [2] extends this result to variable-rate streaming traffic satisfying Assumption 7.1 when either (i) the system is critically loaded, or (ii) the system is work-conserving.

Proposition A.1. *Suppose that $B(\cdot) \in \mathcal{R}_{-\nu}$ and $Kr < 1-\rho$. If $A_1(0,t) \equiv Krt$ and $1-\rho < (K+1)r$, then*

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

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. First observe that

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

It remains to be shown that

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

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 (3), 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 $\epsilon<1/M$)

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

The first term can be made sufficiently small for any fixed δ, ϵ, ξ, using similar arguments as in [8]. 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 $\delta>0$*

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

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

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

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 (3) in the first and final one, gives

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

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

Due to space limitations, we refer to [2] for the proofs of the following three lemmas.

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

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

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

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

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

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

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] Bekker, R., Borst, S.C., Núñez-Queija, R. (2004). Performance of TCP-friendly streaming sessions in the presence of heavy-tailed elastic flows. In preparation.

[3] 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. SIGCOMM 2001*, 111–122.

[4] 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.

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

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

[7] 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.

[8] Borst, S.C., Zwart, A.P. (2001). Fluid queues with heavy-tailed M/G/∞ input. SPOR-Report 2001-02, Eindhoven University of Technology.

[9] 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.*

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

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

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

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

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

[15] Guillemin, F., Robert, Ph., Zwart, A.P. (2003). Tail asymptotics for processor sharing queues. *Adv. Appl. Prob.*, to appear.

[16] Jelenković, P.R., Momčilović, P., Zwart, A.P. (2004). Reduced load equivalence under subexponentiality. *Queueing Systems* **46**, 97–112.

[17] 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.

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

[19] 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.

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

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

[22] 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.

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

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

[25] 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. Infocom '99*, 1337–1346.

[26] Zwart, A.P., Borst, S.C., Mandjes, M. (2001). Exact asymptotics for fluid queues fed by multiple heavy-tailed on-off flows. SPOR-Report 2000-14, Eindhoven University of Technology. Shortened version in: *Proc. Infocom 2001*, 279–288.