Generalized Processor Sharing With Light-Tailed and Heavy-Tailed Input

Sem Borst, Michel Mandjes, and Miranda van Uitert

***Abstract*—We consider a queue fed by a mixture of light-tailed and heavy-tailed traffic. The two traffic flows are served in accordance with the generalized processor sharing (GPS) discipline. GPS-based scheduling algorithms, such as weighted fair queueing, have emerged as an important mechanism for achieving service differentiation in integrated networks. We derive the asymptotic workload behavior of the light-tailed traffic flow under the assumption that its GPS weight is larger than its traffic intensity. The GPS mechanism ensures that the workload is bounded above by that in an isolated system with the light-tailed flow served in isolation at a constant rate equal to its GPS weight. We show that the workload distribution is in fact asymptotically equivalent to that in the isolated system, multiplied with a certain pre-factor, which accounts for the interaction with the heavy-tailed flow. Specifically, the pre-factor represents the probability that the heavy-tailed flow is backlogged long enough for the light-tailed flow to reach overflow. The results provide crucial qualitative insight in the typical overflow scenario.**

***Index Terms*—Generalized processor sharing (GPS), heavy-tailed traffic, large deviations, light-tailed traffic, Markov fluid, regular variation, weighted fair queueing, workload asymptotics.**

I. Introduction

THE next-generation Internet is expected to support a wide variety of services, such as voice, video, and data applications. Voice and video communications induce far more stringent quality-of-service (QoS) requirements than the typical sort of data applications which currently account for the bulk of the Internet traffic. The integration of heterogeneous services thus raises the need for differentiated QoS, catering to the specific requirements of the various traffic flows.

One potential approach to achieve service differentiation is through the use of discriminatory scheduling algorithms, which distinguish between packets of various traffic streams. Because of scalability issues, it is practically infeasible, though, to manipulate packets at the granularity level of individual traffic flows in the core of any large-scale high-speed network. To avoid these complexity problems, traffic flows may instead be aggregated into a small number of classes with roughly similar features, with scheduling mechanisms acting at the coarser level of aggregate streams. With a little simplification, the majority of applications may, for example, be broadly categorized into just two classes, one containing *streaming* traffic (e.g., audio and video communications), the other one comprising *elastic* traffic (e.g., file transfers). This is a crucial element of the DiffServ proposal [5], which defines the expedited forwarding (EF) class for delay-sensitive traffic and the assured forwarding (AF) class for traffic with some degree of delay tolerance.

In view of the delay requirements, it is desirable that streaming applications receive some sort of priority over elastic traffic, at least over short time scales. Strict priority scheduling may, however, not be ideal, since it may lead to starvation of the best-effort traffic. Even temporary starvation effects may cause end-to-end flow control mechanisms such as TCP to suffer a severe degradation in throughput performance. The generalized processor sharing (GPS) discipline provides a potential mechanism for implementing priority scheduling in a tunable way, with strict priority scheduling as an extreme option [28]. In GPS-based scheduling algorithms, such as weighted fair queueing (WFQ), the link capacity is shared in proportion to certain class-defined weight factors. By setting the weight factor for the best-effort class relatively low, one can still provide some degree of priority to the streaming applications, while avoiding starvation of the elastic traffic.

Besides achieving service differentiation, scheduling mechanisms also play a role in controlling the performance impact of bursty traffic. Extensive measurements have shown that bursty traffic behavior may extend over a wide range of time scales, and may manifest itself in long-range dependence and self-similarity [23], [30]. The occurrence of these phenomena is commonly attributed to extreme variability and heavy-tailed characteristics in the traffic patterns [3], [13]. These observations have triggered a strong interest in queueing models with heavy-tailed traffic processes (see, for instance, [29] and [33]).

Although the presence of heavy-tailed traffic characteristics is widely acknowledged, the practical implications for network performance and traffic engineering remain controversial. For small buffer sizes, the effect of heavy-tailed traffic characteristics is not as dramatic as indicated by theoretical studies for infinite buffer sizes, especially at high levels of multiplexing [12], [17], [24], [32]. For large buffer sizes, flow control mechanisms such as TCP prevent heavy-tailed activity patterns from overwhelming the buffers [2].

In this paper, we specifically examine the potential role of GPS-based scheduling mechanisms in protecting light-tailed

Manuscript received April 12, 2002; revised January 25, 2003; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor D. Lee.

S. Borst is with the Center for Mathematics and Computer Science (CWI), 1090 GB Amsterdam, The Netherlands. He is also with Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974 USA and the Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands (e-mail: sem@cwi.nl).

M. Mandjes is with the Center for Mathematics and Computer Science (CWI), 1090 GB Amsterdam, The Netherlands. He is also with the Faculty of Mathematical Sciences, University of Twente, Enschede, The Netherlands (e-mail: michel@cwi.nl).

M. van Uitert is with the Center for Mathematics and Computer Science (CWI), 1090 GB Amsterdam, The Netherlands (e-mail: miranda@cwi.nl).

Digital Object Identifier 10.1109/TNET.2003.818195

traffic flows from the impact of heavy-tailed traffic processes. Large-deviations results for GPS models with light-tailed traffic may be found in [26] and [34]. Workload asymptotics for GPS queues with heavy-tailed traffic flows were obtained in [6] and [21]. The latter results show a sharp dichotomy in qualitative behavior, depending on the traffic intensities and the relative values of the weight parameters. For certain weight combinations, an individual flow with heavy-tailed characteristics is effectively served at a *constant* rate, which is only influenced by the average rates of the other flows. In particular, the flow is essentially immune from excessive activity of flows with heavier-tailed characteristics. For other weight combinations, however, a flow may be strongly affected by the activity of heavier-tailed flows and may inherit their traffic characteristics. The latter result, in fact, also applies for light-tailed flows when their traffic intensity exceeds their GPS weight. In this paper, we derive the asymptotic workload behavior of the light-tailed flow for the more plausible situation where its GPS weight is larger than its traffic intensity.

The remainder of this paper is organized as follows. In Section II, we present a detailed model description and state some important preliminary results. In Section III, we provide an overview of the main results of the paper, which characterize the exact asymptotic behavior of the workload distribution of the light-tailed flow. The subsequent sections give a sketch of the proofs. We start in Section IV with deriving lower and upper bounds for the workload distribution of the light-tailed flow. In Section V, we proceed to prove some auxiliary results for the light-tailed flow in isolation. Although the bounds seem quite crude by themselves, we show in Section VI that they asymptotically coincide, yielding the exact asymptotic behavior. One of the asymptotic terms involves the probability that the heavy-tailed flow is backlogged long enough for overflow to occur, which is computed in Sections VII and VIII.

II. Model Description

We now present a detailed model description (see Fig. 1). We consider two traffic flows sharing a link of unit rate. Traffic from the flows is served in accordance with the GPS discipline, which operates as follows. Flow i is assigned a weight ϕ_i, $i = 1, 2$, with $\phi_1 + \phi_2 = 1$. As long as both flows are backlogged, flow i is served at rate ϕ_i, $i = 1, 2$. If one of the flows is not backlogged, however, then the service rate is reallocated to the other flow, which is then served at the full link rate (if backlogged). (It may occur that one of the flows is not backlogged, while generating traffic at some rate $r_i < \phi_i$. In that case, only the *excess* service rate $\phi_i - r_i$, is reallocated to the other flow.) Denote by $V_i(t)$ the workload of flow i at time t and by V_i a random variable whose distribution is the limit distribution of $V_i(t)$ for $t \to \infty$ (assuming it exists). The goal of this paper is to derive the exact asymptotic behavior of the workload distribution of flow 1, i.e., we calculate $\mathbb{P}(V_1 > x)$ for $x \to \infty$.

We introduce the following notation. For $t > 0$, we denote by $A_i(t)$ the amount of traffic generated by flow i during the time interval $(0, t]$. For $t \leq 0$, $A_i(t)$ denotes the negative counterpart of the amount of traffic generated by flow i in $(t, 0]$. Assuming that $A_i(\cdot)$ is reversible and has stationary increments, we define

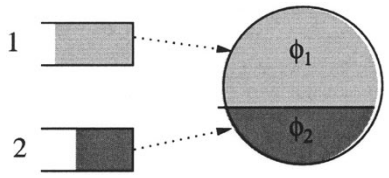

Fig. 1. Schematic representation of our model.

$A_i(s, t) := A_i(t) - A_i(s)$ to be the amount of traffic generated in $(s, t]$, $s < t$, $s, t \in \mathbb{R}$. Define $B_i(s, t)$ as the amount of service received by flow i during $(s, t]$. Then the following identity relation holds:

$$V_i(t) = V_i(s) + A_i(s, t) - B_i(s, t), \quad \text{for all } s < t. \tag{1}$$

For any $c \geq 0$, denote by $V_i^c(t) := \sup_{s \leq t}\{A_i(s, t) - c(t - s)\}$ the workload at time t in a fictitious queue with service rate c fed by flow i (viewed in isolation). Denote by ρ_i the traffic intensity of flow i (as will be defined in detail below). For $c > \rho_i$, V_i^c is a random variable whose distribution is the limit distribution of $V_i^c(t)$ for $t \to \infty$ (assuming it exists). Then a similar identity relation as above holds:

$$V_i^c(t) = V_i^c(s) + A_i(s, t) - B_i^c(s, t), \quad \text{for all } s < t. \tag{2}$$

In the next two sections, we describe the traffic model that we consider for both flows. We first introduce some additional notation. For any two real functions $g(\cdot)$ and $h(\cdot)$, we use the notational convention $g(x) \sim h(x)$ to denote $\lim_{x\to\infty} g(x)/h(x) = 1$ (or $g(x) = h(x)(1 + o(1))$ as $x \to \infty$). Also, $g(x) \lesssim h(x)$ denotes $\limsup_{x\to\infty} g(x)/h(x) \leq 1$, and $g(x) \gtrsim h(x)$ denotes $\liminf_{x\to\infty} g(x)/h(x) \geq 1$. For any two random variables X and Y, we write $X \stackrel{d}{=} Y$ to denote that they have the same distribution function. For any positive real-valued random variable X with distribution function $F(\cdot)$, $\mathbb{E}X < \infty$, denote by $F^r(\cdot)$ the distribution function of the residual lifetime of X, i.e., $F^r(x) = (1/\mathbb{E}X)\int_0^x (1 - F(y))\mathrm{d}y$, and by X^r a random variable with that distribution. The classes of *subexponential*, *regularly varying*, and *intermediately regularly varying* distributions are denoted with the symbols $\mathcal{S}$, $\mathcal{R}$, and $\mathcal{IR}$, respectively. The definitions of these classes may be found in [4].

A. Traffic Model Flow 1

We assume that flow 1 is light-tailed. Specifically, we make the assumption that the input process $A_1(s, t)$ is a *Markov-modulated fluid*. Such a process can be described as follows. There is an irreducible Markov chain with a finite state space $\{1, 2, \ldots, d\}$. The corresponding transition rate matrix is denoted by $\Lambda := (\lambda_{ij})_{i,j=1,\ldots,d}$, where we follow the convention that $\lambda_{ii} := -\sum_{j\neq i} \lambda_{ij}$. Since the Markov chain is irreducible, there is a unique stationary distribution, which we denote by the (column) vector π. When the source is in state i, traffic is generated (as fluid) at constant rate $R_i < \infty$. Let R be the diagonal matrix with the coefficients R_i on the diagonal. Denote the mean rate by $\rho_1 := \sum_{i=1}^d \pi_i R_i$. Denote the peak rate by $R_P := \max_{i=1,\ldots,d} R_i$. It is important to observe that the class of Markov fluid input is closed under superposition, i.e., the superposition of Markov fluid sources can again be modeled as a Markov fluid source. The following standard result

for Markov-modulated fluid sources follows directly from [15], [20], and [22].

Proposition 2.1: For any $c \in (\rho_1, R_P)$

- The moment generating function of $A_1(0,t)$ reads

$$\mathbb{E}\left[\mathrm{e}^{sA_1(0,t)}\right] = \pi^T \mathrm{e}^{(\Lambda+sR)t}\mathbf{1}$$

with $\mathbf{1}$ the all one (column) vector of dimension d.

- For continuous and differentiable $M_c(\cdot)$

$$\lim_{t\to\infty} \frac{1}{t} \log \mathbb{E}\left[\mathrm{e}^{s(A_1(0,t)-ct)}\right] = M_c(s).$$

For finite positive C

$$\mathbb{E}\left[\mathrm{e}^{s(A_1(0,t)-ct)}\right] \leq C\mathrm{e}^{M_c(s)t}.$$

- Denoting by $s^*(c)$ the unique positive root of $M_c(s) = 0$, we have $M_c'(s^*(c)) > 0$, and

$$\liminf_{x\to\infty} \frac{1}{x} \log \mathbb{P}\left(V_1^c > x\right) = -s^*(c).$$

Although we restrict ourselves to Markov fluid input, we believe that our results are valid for a more general class of light-tailed input. This will be discussed in greater detail in Remark 6.1.

B. Traffic Model Flow 2

We assume that flow 2 is heavy-tailed. The input process $A_2(s,t)$ is either instantaneous or on–off, with heavy-tailed burst sizes or on periods, respectively.

1) Instantaneous Input: Here, flow 2 generates instantaneous traffic bursts according to a renewal process. The interarrival times between bursts U_2 have distribution function $U_2(\cdot)$ with mean $1/\lambda_2$. The burst sizes B_2 have distribution function $B_2(\cdot)$ with mean $\beta_2 < \infty$. Thus, the traffic intensity is $\rho_2 := \lambda_2\beta_2$. We assume that $B_2(\cdot)$ is regularly varying of index $-\nu_2$, i.e., $B_2(\cdot) \in \mathcal{R}_{-\nu_2}$ for some $\nu_2 > 1$. The next result, which is due to Pakes [27], then yields the tail behavior of the workload distribution of flow 2 in isolation.

Theorem 2.1: If $B_2^r(\cdot) \in \mathcal{S}$ and $\rho_2 < c$, then

$$\mathbb{P}\left(V_2^c > x\right) \sim \frac{\rho_2}{c-\rho_2} \mathbb{P}\left(B_2^r > x\right).$$

2) Fluid Input: Here, flow 2 generates traffic according to an on–off process, alternating between on and off periods. The off periods U_2 have distribution function $U_2(\cdot)$ with mean $1/\lambda_2$. The on periods A_2 have distribution function $A_2(\cdot)$ with mean $\alpha_2 < \infty$. While on, flow i produces traffic at constant rate r_2, so the mean burst size is $\alpha_2 r_2$. The fraction of time that flow 2 is off is $p_2 = 1/(1+\lambda_2\alpha_2)$. The traffic intensity is $\rho_2 = (1-p_2)r_2 = (\lambda_2\alpha_2 r_2)/(1+\lambda_2\alpha_2)$. We assume that $A_2(\cdot)$ is regularly varying of index $-\nu_2$, i.e., $A_2(\cdot) \in \mathcal{R}_{-\nu_2}$ for some $\nu_2 > 1$. The next result, which is due to Jelenković and Lazar [18], then yields the tail behavior of the workload distribution of flow 2 in isolation.

Theorem 2.2: If $A_2^r(\cdot) \in \mathcal{S}$ and $\rho_2 < c < r_2$, then

$$\mathbb{P}\left(V_2^c > x\right) \sim p_2 \frac{\rho_2}{c-\rho_2} \mathbb{P}\left(A_2^r > \frac{x}{r_2 - c}\right).$$

III. Overview of the Results

Throughout this paper, we assume $\rho_i < \phi_i$, $i = 1, 2$, which ensures stability of both flows. We first briefly discuss in Section III-A what happens if this condition fails to hold. In addition, we make the assumption that $r_2 > \phi_2$ in case of fluid input of flow 2. Otherwise, the workload of flow 2 would be zero, so the workload of flow 1 would be equal to the total workload V. The tail distribution of the latter quantity has been obtained in [10]. In Section III-B, we provide a heuristic explanation of the main results of this paper. The main result is then given in Section III-C, where we also present an example.

A. Case $\rho_1 > \phi_1$

To put things in perspective, we first briefly review the case that $\rho_1 > \phi_1$, while $\rho_1 + \rho_2 < 1$. If either: 1) $B_2^r(\cdot) \in \mathcal{IR}$ (instantaneous input); or 2) $A_2^r(\cdot) \in \mathcal{IR}$ with $r_2 > \phi_2$ (fluid input), then from [6]

$$\mathbb{P}(V_1 > x) \sim \frac{\phi_2 - \rho_2}{\phi_2} \frac{\rho_2}{1-\rho_1-\rho_2} \mathbb{P}\left(P_2^r > \frac{x}{\rho_1 - \phi_1}\right)$$

with P_2 a random variable whose distribution is the busy-period distribution in a queue with constant service rate ϕ_2 fed by flow 2.

The above result may be interpreted as follows. Large-deviations arguments suggest that the most likely way for flow 1 to build a large queue is that flow 2 generates a large burst, or experiences a long on period, while flow 1 itself shows roughly average behavior. Note that when flow 2 produces a large amount of traffic, so that it becomes backlogged for a long period of time, it receives service at rate ϕ_2. Thus, it will experience a busy period as if it were served at constant rate ϕ_2. During that period, flow 1 receives service at rate ϕ_1, while it generates traffic roughly at rate ρ_1, so its queue will grow approximately at rate $\rho_1 - \phi_1$. When flow 2 is not backlogged, the corresponding queue will drain approximately at rate $1 - \rho_1$.

Thus, the backlog of flow 1 behaves as that in a queue with constant service rate $1 - \rho_1$ fed by an on–off source with peak rate ϕ_2. The on and off periods correspond to the busy and idle periods of flow 2 when served at constant rate ϕ_2, respectively. Recall that the workload asymptotics of such an on–off source are given by Theorem 2.2. Setting $c = 1 - \rho_1$, $r_2 = \phi_2$, $p_2 = 1 - \rho_2/\phi_2$, and identifying A_2^r with P_2^r, we obtain the above result for the workload asymptotics of flow 1.

B. Heuristic Explanation of Main Results

We now focus on the case $\rho_1 < \phi_1$. Before presenting the main result, we first provide a heuristic derivation of the asymptotic behavior of $\mathbb{P}(V_1 > x)$ based on large-deviations arguments as in [1]. The overflow scenario described above for the case $\rho_1 > \phi_1$ cannot occur, and now flow 1 also must deviate from its "normal" behavior in order for the queue to grow. Specifically, large-deviations results suggest that flow 1 must behave as if its traffic intensity is temporarily increased from ρ_1 to some larger value $\hat{\rho}_1 > \phi_1$ (as will be specified below). During that time period, flow 2 is continuously backlogged, consuming service rate ϕ_2, thus leaving service rate ϕ_1 for flow 1. (Notice that if flow 2 were not permanently backlogged, then flow 1 would have to show even greater anomalous activity in

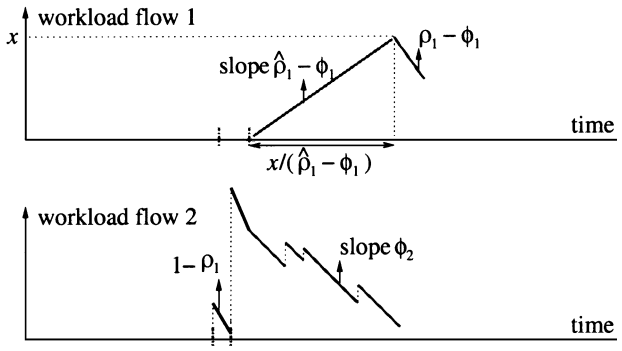

Fig. 2. Overflow scenario—instantaneous input flow 2.

order for a given backlog level to occur.) Prior to that period, flow 1 shows normal behavior, leaving an average service rate of $1-\rho_1$ for flow 2.

To summarize, the intuitive argument is as follows (see Fig. 2). A large backlog of level x of flow 1 occurs as a consequence of two rare events: 1) flow 1 shows similar "abnormal" behavior as is the typical cause of overflow when served in isolation, thus behaving as if its traffic intensity is increased from ρ_1 to $\hat{\rho}_1$ for a period of time $x/(\hat{\rho}_1 - \phi_1)$; and 2) during that period, flow 2 is constantly backlogged, demanding capacity ϕ_2, with ϕ_1 remaining for flow 1. As we will see later, the persistent backlog is most likely caused by flow 2 generating a large burst or initiating a long on period prior to that period.

These considerations lead to the following characterization of the asymptotic behavior of $\mathbb{P}(V_1 > x)$:

$$\mathbb{P}(V_1 > x) \sim \mathbb{P}\left(V_1^{\phi_1} > x\right)\mathbb{P}\left(T_2^{1-\rho_1} > \frac{x}{\hat{\rho}_1 - \phi_1}\right). \quad (3)$$

The second term represents the probability that flow 2 is continuously backlogged during a period of time $x/(\hat{\rho}_1 - \phi_1)$, receiving a service rate ϕ_2 starting from some time t on, and having received a service rate $1-\rho_1$ prior to time t. Here $T_2^{1-\rho_1}$ is a random variable whose distribution is the limit distribution of $T_2^{1-\rho_1,t}$ for $t \to \infty$, with

$$T_2^{c,t} := \inf\{u \geq 0 : V_2^c(t) + A_2(t, t+u) - \phi_2 u = 0\}$$

representing the drain time in a queue with service rate ϕ_2 fed by flow 2 with initial workload $V_2^c(t)$. The service rate $1-\rho_1$ reflects the fact that flow 1 has shown normal behavior prior to time t.

Thus, the workload distribution is asymptotically equivalent to that in an isolated system, multiplied with a certain pre-factor. The isolated system consists of flow 1 served in isolation at constant rate ϕ_1. The pre-factor represents the probability that flow 2 is backlogged long enough for flow 1 to reach overflow. The combination of light-tailed and heavy-tailed large deviations is similar to that in the *reduced-peak equivalence* result derived in [10] as well as that for an M/G/2 queue with heterogeneous servers studied in [11].

Note that the general decompositional form of (3) holds irrespective of the detailed traffic characteristics of the two flows. (In fact, the above intuitive arguments suggest that (3) may be true under somewhat milder assumptions than those made in Sections II-A and B. This will be discussed in greater detail in Remark 6.1.) However, the specific form of the two individual terms in (3) *does* depend on the detailed properties of the traffic processes. In particular, we need to distinguish whether flow 2 generates instantaneous or fluid input. In the latter case, it also depends on whether the peak rate r_2 exceeds $1-\rho_1$ or not.

C. Main Results

We now state the main theorem of the paper.

Theorem 3.1: Defining $\hat{\rho}_1 := M'_{\phi_1}(s^*(\phi_1)) + \phi_1$

$$\mathbb{P}(V_1 > x) \sim \mathbb{P}\left(V_1^{\phi_1} > x\right)\mathbb{P}\left(T_2^{1-\rho_1} > \frac{x}{\hat{\rho}_1 - \phi_1}\right).$$

Case I (instantaneous input):

$$\mathbb{P}\left(T_2^{1-\rho_1} > x\right) \sim \frac{\rho_2}{1-\rho_1-\rho_2}\mathbb{P}\left(B_2^r > x(\phi_2 - \rho_2)\right). \quad (4)$$

Case II-A (fluid input with $r_2 < 1-\rho_1$):

$$\mathbb{P}\left(T_2^{1-\rho_1} > x\right) \sim (1-p_2)\mathbb{P}\left(A_2^r > \frac{x(\phi_2-\rho_2)}{r_2-\rho_2}\right). \quad (5)$$

Case II-B (fluid input with $r_2 > 1-\rho_1$):

$$\mathbb{P}\left(T_2^{1-\rho_1} > x\right) \sim \frac{p_2\rho_2}{1-\rho_1-\rho_2}\mathbb{P}\left(A_2^r > \frac{x(\phi_2-\rho_2)}{r_2-\rho_2}\right). \quad (6)$$

Noting that $p_2\rho_2 = (1-p_2)(r_2-\rho_2)$, we can observe that in the limiting regime $r_2 \to 1-\rho_1$ cases II-A and II-B coincide. Also, case I can be seen as the limiting case of II-B if we use $r_2 A_2 = B_2$ and let $r_2 \to \infty$ so that $p_2 \downarrow 1$. In [7], a qualitatively similar result as in case I is derived for a system with two coupled queues, one having heavy-tailed input, the other one exhibiting light-tailed properties.

To illustrate Theorem 3.1, we give an example. Assume flow 1 to behave according to an on–off process with exponentially distributed on and off periods with means $1/\mu_1$ and $1/\mu_2$, respectively. When the flow is in the on state, it generates traffic at rate R_1. We assume flow 2 to generate instantaneous input with regularly varying burst sizes of index $-\nu_2$, i.e., $\mathbb{P}(B_2 > x) \sim C_2 x^{-\nu_2} l_2(x)$, with $l_2(\cdot)$ some slowly varying function. First, we determine the deviant traffic intensity $\hat{\rho}_1$ using [25]

$$\hat{\rho}_1 = \frac{\left(\frac{R_1\phi_1^2}{\mu_2}\right)}{\left(\frac{\phi_1^2}{\mu_2} + \frac{(R_1-\phi_1)^2}{\mu_1}\right)}.$$

Using [14], we obtain for flow 1

$$\mathbb{P}\left(V_1^{\phi_1} > x\right) \sim \frac{R_1}{\phi_1}\frac{\mu_2}{\mu_1+\mu_2}\exp\left\{-\left(\frac{\mu_1}{R_1-\phi_1} + \frac{\mu_2}{\phi_1}\right)x\right\}.$$

For flow 2, from (4)

$$\mathbb{P}(T_2^{1-\rho_1} > x)$$
$$\sim \frac{\rho_2}{1-\rho_1-\rho_2}\frac{C_2}{\beta_2(\nu_2-1)}\left(x(\phi_2-\rho_2)\right)^{1-\nu_2} l_2\left(x(\phi_2-\rho_2)\right).$$

This provides all the ingredients for $\mathbb{P}(V_1 > x)$ as required in Theorem 3.1.

The next sections are devoted to the formal proof of Theorem 3.1. We start in Section IV by deriving lower and upper bounds for the workload distribution of flow 1. We then proceed in Section V to prove some auxiliary results for flow 1 in isolation. Although the bounds derived in Section IV seem

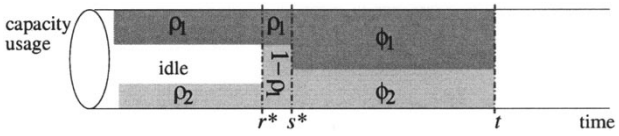

Fig. 3. Intuitive idea lower bound.

quite crude by themselves, we show in Section VI that they asymptotically coincide, yielding the exact asymptotic behavior of $\mathbb{P}(V_1 > x)$.

In order to determine the drain time distribution of flow 2 as specified in Theorem 3.1, we first establish in Section VII some preliminary results for flow 2 in isolation. Note that the specific form of the drain time distribution depends on whether flow 2 generates instantaneous or fluid input. In the latter case, we also need to distinguish whether the peak rate r_2 exceeds $1-\rho_1$ or not. We calculate the drain time distribution for the case of an instantaneous input process in Section VIII. In view of space constraints, we omit the corresponding analysis for fluid input processes (see [9] for details).

IV. Bounds

In this section, we derive lower and upper bounds for the workload distribution of flow 1. Refer to [9] for detailed proofs of the lemmas in this section.

A. Lower Bound

We start with a lower bound for the workload distribution of flow 1. The main idea (see Fig. 3) is that the following scenario is sufficient for the event $V_1(t) > x$ to occur (in fact, it is the only plausible one, as we will see later). Flow 1 starts to build up at some time s^* and, hence, is constantly backlogged throughout the time interval $[s^*, t]$. Flow 2 is also continuously backlogged during $[s^*, t]$. Thus, during that time period, flows 1 and 2 both receive service at rates ϕ_1 and ϕ_2, respectively. Flow 2 already becomes backlogged at time $r^* \le s^*$ and receives service approximately at rate $1-\rho_1$ during $[r^*, s^*]$, while flow 1 then shows roughly average behavior.

Lemma 4.1: Suppose $r^* \le s^* \le t$ and y exist such that

$$A_1(s^*,t) - \phi_1(t-s^*) > x$$
$$A_1(r^*,s^*) - (\rho_1-\epsilon)(s^*-r^*) \ge -y$$
$$\inf_{s^* \le u \le t}\{A_2(r^*,u) - (1-\rho_1+\epsilon)(s^*-r^*) - \phi_2(u-s^*)\} \ge y.$$

Then $V_1(t) > x$.

Proof: Using (1), GPS implies that

$$B_2(s,t) \ge \min\left\{\phi_2(t-s), V_2(s) + \inf_{s\le u\le t}\{A_2(s,u) + \phi_2(t-u)\}\right\}.$$

Combining this in (1) with $B_1(s,t) \le t - s - B_2(s,t)$, which holds by definition for all $s < t$, gives

$$V_1(t) \ge A_1(s,t) - \phi_1(t-s) + \min\left\{0, V_1(s) + V_2(s) + \inf_{s\le u\le t}\{A_2(s,u) - \phi_2(u-s)\}\right\}.$$

Using $B_1(r,s) + B_2(r,s) \le s - r$ together with (1) implies

$$V_1(s) + V_2(s) \ge A_1(r,s) + A_2(r,s) - (s-r) \quad \forall\, r \le s.$$

Substituting, we find that $V_1(t)$ is bounded from below by

$$A_1(s,t) - \phi_1(t-s) + \min\Big\{0, A_1(r,s) - (\rho_1-\epsilon)(s-r) + \inf_{s\le u\le t}\{A_2(r,u) - (1-\rho_1+\epsilon)(s-r) - \phi_2(u-s)\}\Big\}$$

for all $r \le s \le t$. □

The next step is to translate the above sample-path result into a probabilistic lower bound. We first introduce some additional notation. For any c and $w \ge 0$, define $V_i^c[w] := \sup_{0\le s\le w}\{A_i(-s,0) - cs\}$. Note that, for $c > \rho_i$, $V_i^c[\infty] \stackrel{d}{=} V_i^c$ as defined earlier. For any c, $v \ge 0$, and y, define

$$T_2^c(v,y) := \inf\left\{u \ge 0 : \sup_{0\le r\le v}\{A_2(-r,0) - cr\} + A_2(0,u) - \phi_2 u \le y\right\}$$

representing the drain time in a queue of capacity ϕ_2 fed by flow 2 with initial workload $\sup_{0\le r\le v}\{A_2(-r,0) - cr\} - y$. Define, for $c > \rho_2$

$$T_2^c(y) := T_2^c(\infty, y) = \inf\{u \ge 0 : V_2^c(0) + A_2(0,u) - \phi_2 u \le y\}$$

and note that $T_2^c(0) \stackrel{d}{=} T_2^c$ as defined earlier. Also, define

$$T_2(y) := T_2^c(0,y) = \inf\{u \ge 0 : A_2(0,u) - \phi_2 u \le y\}$$

(note that the latter quantity does not depend on the value of c) and $T_2 := T_2(0)$. Denote

$$P^{\rho_1-\epsilon}(s^*,v,x,y) := \mathbb{P}\left(\sup_{s^*-v\le r\le s^*}\{(\rho_1-\epsilon)(s^*-r) - A_1(r,s^*)\} \le y \,|\, A_1(s^*,0) + \phi_1 s^* > x\right).$$

The following corollary gives the probabilistic lower bound. The proof uses the sample-path relation as given in Lemma 4.1 and can be found in Appendix A.

Corollary 4.1: For any $v \ge 0$ and y

$$\mathbb{P}(V_1 > x) \ge \mathbb{P}\left(V_1^{\phi_1}\left[\frac{(1+\alpha)x}{\hat\rho_1 - \phi_1}\right] > x\right) \times P^{\rho_1-\epsilon}(s^*,v,x,y)\mathbb{P}\left(T_2^{1-\rho_1+\epsilon}(v,y) > \frac{(1+\alpha)x}{\hat\rho_1 - \phi_1}\right).$$

B. Upper Bound

We proceed with the upper bound. The idea is that the lower-bound scenario described above is basically also necessary for the event $V_1(t) > x$ to occur.

Lemma 4.2: Suppose $V_1(t) > x$. Then for all y there exist $r^* \le s^* \le t$ such that

$$A_1(s^*,t) - \phi_1(t-s^*) > x \tag{7}$$

and at least one of the three following events occurs:

$$A_1(r^*,s^*) - (\rho_1+\epsilon)(s^*-r^*) > y \tag{8}$$

or

$$V_1^{\phi_1}(t) > x + y \tag{9}$$

or

$$\inf_{s^* \le u \le t} \{A_2(r^*,u) - (1-\rho_1-\epsilon)(s^*-r^*) - \phi_2(u-s^*)\} > -2y. \quad (10)$$

Proof: Because of the GPS discipline, (7) is implied by $V_1(t) > x$, i.e.,

$$V_1(t) \le V_1^{\phi_1}(t) = \sup_{s \le t} \{A_1(s,t) - \phi_1(t-s)\}.$$

Hence, $s \le t$ exists such that $A_1(s,t) - \phi_1(t-s) > x$. Define $s^* := \inf\{s : A_1(u,t) - \phi_1(t-u) \le x \ \forall u > s\}$. Using this definition, it can be shown that flow 1 must be continuously backlogged during $[s^*,t]$. We now show by contradiction that $V_1(t) > x$ implies either (9) or

$$\forall u \in [s^*,t] : B_2(s^*,u) - \phi_2(u-s^*) > -y. \quad (11)$$

Suppose

$$\exists\, u^* \in [s^*,t] :\ B_2(s^*,u^*) - \phi_2(u^*-s^*) \le -y \quad (12)$$

and

$$\forall\, q \le s^* \le t : A_1(q,t) - \phi_1(t-q) \le x+y \quad (13)$$

hold. Since flow 1 is continuously backlogged during $[s^*,t]$

$$V_1(t) = V_1(s^*) + A_1(s^*,t) - (t-s^*) + B_2(s^*,u^*) + B_2(u^*,t)$$

and $B_2(u^*,t) \le \phi_2(t-u^*)$. Because of GPS

$$V_1(s^*) \le \sup_{r \le s^*} \{A_1(r,s^*) - \phi_1(s^*-r)\}.$$

Hence, using (12) and (13) gives

$$\begin{aligned} V_1(t) &\le \sup_{r \le s^*}\{A_1(r,s^*) - \phi_1(s^*-r)\} \\ &\quad + A_1(s^*,t) - (t-s^*) + \phi_2(t-s^*) - y \\ &= \sup_{r \le s^*}\{A_1(r,t) - \phi_1(t-r)\} - y \le x+y-y \\ &= x \end{aligned}$$

which is in contradiction with $V_1(t) > x$. Finally, we show that (11) implies (8) or (10). By definition

$$\begin{aligned} B_2(s^*,u) &\le V_2(s^*) + A_2(s^*,u) \\ &\le V(s^*) + A_2(s^*,u) \\ &= \sup_{r \le s^*} \{A_1(r,s^*) + A_2(r,s^*) - (s^*-r)\} + A_2(s^*,u). \end{aligned}$$

Hence

$$\begin{aligned} &\inf_{s^* \le u \le t} \{B_2(s^*,u) - \phi_2(u-s^*)\} \\ &\le \sup_{r \le s^*} \inf_{s^* \le u \le t} \{A_2(r,u) - (1-\rho_1-\epsilon)(s^*-r) - \phi_2(u-s^*)\} \\ &\qquad + \sup_{r \le s^*} \{A_1(r,s^*) - (\rho_1+\epsilon)(s^*-r)\}. \quad \square \end{aligned}$$

In the next corollary, the above sample-path relation is translated into a probabilistic upper bound. Denote

$$Q^{\rho_1+\epsilon}(s^*,x,y) := \mathbb{P}\left(\sup_{r \le s^*} \{A_1(r,s^*) - (\rho_1+\epsilon) \times (s^*-r)\} > y | A_1(s^*,0) + \phi_1 s^* > x\right).$$

Corollary 4.2: For any y

$$\begin{aligned} \mathbb{P}(V_1 > x) \le{}& \mathbb{P}\left(V_1^{\phi_1} > x\right) Q^{\rho_1+\epsilon}(s^*,x,y) \\ &+ \mathbb{P}\left(V_1^{\phi_1} > x\right) \mathbb{P}\left(T_2^{1-\rho_1-\epsilon}(-2y) > \frac{(1-\alpha)x}{\hat{\rho}_1 - \phi_1}\right) \\ &+ \mathbb{P}\left(V_1^{\phi_1} > x+y\right) + \mathbb{P}\left(V_1^{\phi_1}\left[\frac{(1-\alpha)x}{\hat{\rho}_1-\phi_1}\right] > x\right). \end{aligned}$$

The proof can be found in Appendix A.

V. Preliminary Results Light-Tailed Flow

In this section, we prove some auxiliary results for flow 1 in isolation. The results will be crucial in obtaining the asymptotic behavior of $\mathbb{P}(V_1 > x)$ in the GPS model as given in Theorem 3.1.

The following result is proven in [10], for a more general class of input processes than just Markov fluid sources.

Lemma 5.1: For any $\alpha > 0$

$$\liminf_{x \to \infty} \frac{\mathbb{P}\left(V_1^{\phi_1}\left[\frac{(1+\alpha)x}{\hat{\rho}_1-\phi_1}\right] > x\right)}{\mathbb{P}\left(V_1^{\phi_1} > x\right)} = 1 \quad (14)$$

where $\hat{\rho}_1 := M'_{\phi_1}(s^*(\phi_1)) + \phi_1$.

Lemma 5.2: For any $\gamma > 0$, $\epsilon > 0$, $t^* < 0$,

$$\lim_{x \to \infty} \mathbb{P}\left(\sup_{r \le t^*} \{(\rho_1-\epsilon)(t^*-r) - A_1(r,t^*)\} \le \gamma x | A_1(t^*,0) + \phi_1 t^* > x\right) = 1.$$

Proof: Recall that flow 1 is a Markov fluid source. We condition on the state of the underlying Markov chain at time t^*. Let $E_j(t^*)$ be the event that the state at time t^* is j, and $\pi_j(t^*) := \mathbb{P}(E_j(t^*) | A_1(t^*,0) + \phi_1 t^* > x)$, $j = 1,\ldots,d$. Then, the probability of interest equals

$$\sum_{j=1}^{d} \mathbb{P}\left(\sup_{r \le t^*}\{(\rho_1-\epsilon)(t^*-r) - A_1(r,t^*)\} \le \gamma x | E_j(t^*)\right) \pi_j(t^*).$$

The stated then follows by observing that

$$\lim_{x \to \infty} \mathbb{P}\left(\sup_{r \le t^*}\{(\rho_1-\epsilon)(t^*-r) - A_1(r,t^*)\} \le \gamma x | E_j(t^*)\right) = 1$$

for all $j = 1,\ldots,d$, since $\mathbb{E}[A_1(-t,0)] = \rho_1 t$. $\square$

Lemma 5.3: For any $\gamma > 0$, $\epsilon > 0$, $\mu > 0$, $t^* < 0$,

$$\lim_{x \to \infty} x^{\mu} \mathbb{P}\left(\sup_{r \le t^*} \{A_1(r,t^*) - (\rho_1+\epsilon)(t^*-r)\} > \gamma x | A_1(t^*,0) + \phi_1 t^* > x = 0\right).$$

Proof: Again, condition on the state of the underlying Markov chain at time t^*. Under this condition, the event $\{A_1(t^*,0) + \phi_1 t^* > x\}$ does not provide any extra information. The fact that there exist constants C, a (independent of j) such that [25, sec. 4]

$$\mathbb{P}\left(\sup_{r \le t^*}\{A_1(r,t^*) - (\rho_1+\epsilon)(t^*-r)\} > \gamma x | E_j(t^*)\right) \le C\mathrm{e}^{-ax}$$

proves the stated. $\square$

Lemma 5.4: For any $\gamma > 0$, $\mu > 0$,

$$\limsup_{x\to\infty} \frac{x^\mu \mathbb{P}\left(V_1^{\phi_1} > (1+\gamma)x\right)}{\mathbb{P}\left(V_1^{\phi_1} > x\right)} = 0.$$

Proof: The proof follows immediately from the fact that $\mathbb{P}(V_1^{\phi_1} > x)$ decays exponentially at rate s^*, where $s^* > 0$ is the solution of $M_{\phi_1}(s) = 0$ [22]. □

Lemma 5.5: For any $\alpha > 0$, $\mu > 0$,

$$\limsup_{x\to\infty} \frac{x^\mu \mathbb{P}\left(V_1^{\phi_1}\left[\frac{(1-\alpha)x}{\hat{\rho}_1-\phi_1}\right] > x\right)}{\mathbb{P}\left(V_1^{\phi_1} > x\right)} = 0.$$

Proof: The proof consists of three steps. First, we give a sufficient condition for the lemma to hold, explicitly using the fact that the Markov fluid source has a bounded peak rate R_P. Then, we estimate the decay rate of the event that a queue of capacity ϕ_1 fed by a Markov fluid source reaches overflow at time t. Finally, we identify the most likely epoch of overflow and show that this implies the required property.

1) Obviously

$$\begin{aligned}
&\mathbb{P}(V_1^{\phi_1}[((1-\alpha)x)/(\hat{\rho}_1-\phi_1)] > x)\\
&\quad\le \mathbb{P}\left(\exists t \le T_x(\alpha) : A_1(0,t) - \phi_1 t > x\right)\\
&\quad\le \sum_{t=0}^{T_x(\alpha)} \mathbb{P}\left(A_1(0,t) - \phi_1 t > x - (R_P - \phi_1)\right)
\end{aligned}$$

with $T_x(\alpha) := \lceil((1-\alpha)x)/(\hat{\rho}_1 - \phi_1)\rceil$. From

$$\begin{aligned}
&\max_{t=0,\ldots,T_x(\alpha)} \mathbb{P}(A_1(0,t) - \phi_1 t > x - (R_P - \phi_1))\\
&\quad\le \sum_{t=0}^{T_x(\alpha)} \mathbb{P}(A_1(0,t) - \phi_1 t > x - (R_P - \phi_1))\\
&\quad\le (T_x(\alpha)+1)\\
&\qquad\times \max_{t=0,\ldots,T_x(\alpha)} \mathbb{P}(A_1(0,t) - \phi_1 t > x - (R_P - \phi_1))
\end{aligned}$$

and $\lim_{x\to\infty} 1/x \log(T_x(\alpha)+1) = 0$, we find that

$$\begin{aligned}
&\limsup_{x\to\infty} \frac{1}{x}\log \sum_{t=0}^{T_x(\alpha)} \mathbb{P}(A_1(0,t) - \phi_1 t > x - (R_P - \phi_1))\\
&= \limsup_{x\to\infty} \frac{1}{x}\log \max_{t=0,\ldots,T_x(\alpha)} \mathbb{P}(A_1(0,t) - \phi_1 t > x - (R_P - \phi_1))\\
&\le \limsup_{x\to\infty} \frac{1}{x}\log \sup_{t\in[0,T_x(\alpha)]} \mathbb{P}(A_1(0,t) - \phi_1 t > x - (R_P - \phi_1))\\
&\le \limsup_{x\to\infty} \frac{1}{x}\log \sup_{t\in[S_x,T_x(\alpha)]} \mathbb{P}(A_1(0,t) - \phi_1 t > x - (R_P - \phi_1))
\end{aligned} \tag{15}$$

with $S_x := (x - R_P)/(R_P - \phi_1)$. Notice that we can indeed exclude all t smaller than S_x from the optimization, because in that range no overflow is possible. Clearly, we have proven the stated if we show that the latter decay rate is strictly smaller than s^* for all $\alpha > 0$.

2) For x large enough, and all t between S_x and $T_x(\alpha)$, due to Chebychev's inequality, and Property 2.1

$$\begin{aligned}
\mathbb{P}\left(A_1(0,t) - \phi_1 t > x - (r_P - \phi_1)\right) &\le \inf_{s>0} \frac{\mathbb{E}\mathrm{e}^{s(A_1(0,t)-\phi_1 t)}}{\mathrm{e}^{s(x-(r_P-\phi_1))}}\\
&\le C \inf_{s>0} \frac{\mathrm{e}^{M_{\phi_1}(s)t}}{\mathrm{e}^{s(x-(r_P-\phi_1))}}.
\end{aligned}$$

Now, replace t in (15) by $t_x(\beta) = ((1-\beta)x)/(\hat{\rho}_1 - \phi_1)$, then the supremum is over $\beta \in [\alpha, 1]$. The infimum over $s > 0$ is calculated by differentiation. We get the first-order condition

$$M'_{\phi_1}(s) = \frac{(x - (R_P - \phi_1))(\hat{\rho}_1 - \phi_1)}{(1-\beta)x}.$$

It is easily verified that the right-hand side of the previous equation equals $(\hat{\rho}_1 - \phi_1)(1+\beta)$ for x large and β small. Call the solution $s^*(\beta)$. Recall that s^* solves $M_{\phi_1}(s) = 0$, and that $M'_{\phi_1}(s^*) = \hat{\rho}_1 - \phi_1 > 0$ (see Property 2.1). Using

$$\begin{aligned}
M'_{\phi_1}(s) &= M'_{\phi_1}(s^*) + M''_{\phi_1}(s^*)(s - s^*) + \mathrm{O}\left((s-s^*)^2\right)\\
&= \hat{\rho}_1 - \phi_1 + M''_{\phi_1}(s^*)(s - s^*) + \mathrm{O}\left((s-s^*)^2\right)
\end{aligned}$$

we immediately obtain $s^*(\beta) = s^* + \delta\beta + \mathrm{O}(\beta^2)$, where $\delta := (\hat{\rho}_1 - \phi_1)/(M''_{\phi_1}(s^*)) > 0$, due to the convexity of $M_{\phi_1}(\cdot)$. Also

$$\begin{aligned}
M_{\phi_1}(s^*(\beta)) &= M_{\phi_1}(s^*) + M'_{\phi_1}(s^*)\delta\beta + \mathrm{O}(\beta^2)\\
&= M'_{\phi_1}(s^*)\delta\beta + \mathrm{O}(\beta^2)
\end{aligned}$$

and

$$\begin{aligned}
&\lim_{x\to\infty} (1/x)\log \inf_{s>0} \frac{\mathrm{e}^{t_x(\beta)M_{\phi_1}(s)}}{\mathrm{e}^{s(x-(r_P-\phi_1))}}\\
&\qquad= \lim_{x\to\infty} \frac{1}{x}\left(t_x(\beta)M_{\phi_1}(s^*(\beta)) - s^*(\beta)x\right)\\
&\qquad= \left(\frac{1-\beta}{\hat{\rho}_1-\phi_1}M'_{\phi_1}(s^*) - 1\right)\delta\beta - s^*\\
&\qquad= -\delta\left(\frac{M'_{\phi_1}(s^*)}{\hat{\rho}_1-\phi_1}\right)\beta^2 - s^*\\
&\qquad= -\delta\beta^2 - s^*.
\end{aligned}$$

3) Recall that we have to perform the optimization over $\beta \in [\alpha, 1]$. The supremum over β is clearly attained at $\beta = \alpha > 0$. Now the stated follows from the fact that $\mathbb{P}(V_1^{\phi_1} > x)$ decays at rate s^*, as explained in step 1). □

VI. Asymptotic Analysis

We now use the results from the previous section to show that the lower and upper bounds for $\mathbb{P}(V_1 > x)$ of Section IV asymptotically coincide, resulting in the decompositional form of (3). For the proof, we need to make certain assumptions on the behavior of the drain time distribution $\mathbb{P}(T_2^{1-\rho_1} > x/(\hat{\rho}_1 - \phi_1))$. Later, we will determine the specific form of the drain time distribution, and find that flow 2 indeed satisfies these assumptions.

For notational convenience, we frequently switch to a variable $\hat{x}$, which should be thought of as playing the role of $x/(\hat{\rho}_1-\phi_1)$.

Lemma 6.1: If flow 2 satisfies Assumptions 6.1–6.3 listed below with $c = 1-\rho_1$, then

$$\mathbb{P}(V_1 > x) \sim \mathbb{P}\left(V_1^{\phi_1} > x\right)\mathbb{P}\left(T_2^{1-\rho_1} > \frac{x}{\hat{\rho}_1-\phi_1}\right).$$

Assumption 6.1: For any $\alpha > 0$, $\gamma > 0$, $\epsilon > 0$, either

$$\text{(a)} \qquad \liminf_{\hat{x}\to\infty} \frac{\mathbb{P}\left(T_2^{c+\epsilon}(\gamma\hat{x}) > (1+\alpha)\hat{x}\right)}{\mathbb{P}(T_2^c > \hat{x})} = F^c(\alpha,\gamma,\epsilon)$$

with $\lim_{\alpha,\gamma,\epsilon\downarrow 0} F^c(\alpha,\gamma,\epsilon) = 1$, or

$$\text{(b)} \qquad \liminf_{\hat{x}\to\infty} \frac{\mathbb{P}\left(T_2 > (1+\alpha)\hat{x}\right)}{\mathbb{P}\left(T_2^c > \hat{x}\right)} = F(\alpha)$$

with $\lim_{\alpha\downarrow 0} F(\alpha) = 1$.

Assumption 6.2: For any $\alpha > 0$, $\gamma > 0$, $\epsilon > 0$,

$$\limsup_{\hat{x}\to\infty} \frac{\mathbb{P}\left(T_2^{c-\epsilon}(-\gamma\hat{x}) > (1-\alpha)\hat{x}\right)}{\mathbb{P}\left(T_2^c > \hat{x}\right)} = G^c(\alpha,\gamma,\epsilon)$$

with $\lim_{\alpha,\gamma,\epsilon\downarrow 0} G^c(\alpha,\gamma,\epsilon) = 1$.

Assumption 6.3: For some $\mu > 0$,

$$\liminf_{x\to\infty} \hat{x}^{\mu}\mathbb{P}\left(T_2^c > \hat{x}\right) \geq 1.$$

Proof of Lemma 6.1: The proof consists of a lower bound and an upper bound which asymptotically coincide. We start with the lower bound. We distinguish between two cases, Assumptions 6.1(a) and (b).

(a) Using Corollary 4.1 with $v = \infty$, $y = (\gamma x)/(\hat{\rho}_1 - \phi_1)$, Lemmas 5.1 and 5.2

$$\liminf_{x\to\infty} \frac{\mathbb{P}(V_1 > x)}{\mathbb{P}\left(V_1^{\phi_1} > x\right)\mathbb{P}\left(T_2^{1-\rho_1} > \frac{x}{\hat{\rho}_1-\phi_1}\right)} \geq F^{1-\rho_1}(\alpha,\gamma,\epsilon).$$

Letting $\alpha, \gamma, \epsilon \downarrow 0$ completes the proof.

(b) Using Corollary 4.1 with $v = 0$, $y = 0$, and Lemma 5.1, we obtain, observing that $P^{\rho_1-\epsilon}(s^*, 0, x, 0) = 1$

$$\liminf_{x\to\infty} \frac{\mathbb{P}(V_1 > x)}{\mathbb{P}\left(V_1^{\phi_1} > x\right)\mathbb{P}\left(T_2^{1-\rho_1} > \frac{x}{\hat{\rho}_1-\phi_1}\right)} \geq F(\alpha).$$

Then let $\alpha \downarrow 0$.

We now turn to the upper bound. Using Corollary 4.2 with $v = \infty$, $y = (\gamma x)/(2(\hat{\rho}_1 - \phi_1))$, Lemmas 5.3–5.5, and Assumptions 6.2 and 6.3, for some $\mu > 0$

$$\limsup_{x\to\infty} \frac{\mathbb{P}(V_1 > x)}{\mathbb{P}\left(V_1^{\phi_1} > x\right)\mathbb{P}\left(T_2^{1-\rho_1} > \frac{x}{\hat{\rho}_1-\phi_1}\right)} \leq G^{1-\rho_1}(\alpha,\gamma,\epsilon).$$

Letting $\alpha, \gamma, \epsilon \downarrow 0$ completes the proof. □

In order to complete the proof of Theorem 3.1, it remains to be shown that flow 2 satisfies Assumptions 6.1–6.3, with $\mathbb{P}(T_2^{1-\rho_1} > x/(\hat{\rho}_1 - \phi_1))$ as in (4)–(6). This is done in the following two sections. In view of space limitations, we focus on the case of instantaneous input processes. Refer to [9] for the corresponding analysis for fluid processes.

Remark 6.1: As the proof shows, Lemma 6.1 and, thus, Theorem 3.1 remain true as long as flow 2 satisfies Assumptions 6.1–6.3 and Lemmas 5.1–5.5 hold for flow 1. Both seem to be the case under somewhat milder assumptions than made in Sections II-A and B.

In particular, for the light-tailed flow, the results in [16] suggest that Lemmas 5.1–5.5 hold for a more general class of arrival processes than just Markov fluid. Upon inspection of the proofs in the previous section, we see that two properties were explicitly exploited. In the first place, it was repeatedly used that the source has a bounded peak rate. Second, it is required that the dependence between $A_1(r, t^*)$ and $A_1(t^*, 0)$ is rather mild. This leads us to believe that the lemmas still hold if the exponential sojourn times of the Markov fluid source are replaced by other light-tailed random variables. Probably, an essential prerequisite is that the light-tailed arrival process allows application of the Gärtner–Ellis large-deviations theorem. In particular, this requires that the log moment generating function of the amount of traffic generated in an interval of length t grows at most linearly

$$\lim_{t\to\infty} \frac{1}{t} \log \mathbb{E} \log \exp\{\theta A_1(0,t)\} < \infty$$

for some positive θ. This rules out input processes such as fractional Brownian motion (with Hurst parameter $H \in (1/2, 1)$), or M/G/∞-type inputs with heavy-tailed job sizes.

For the heavy-tailed flow, the results may be extended to semi-Markov fluid input or mixtures of fluid input and instantaneous input. It may also be possible to extend the results to a larger class of subexponential distributions, although that would require elaborate refinements in the proofs. In a somewhat related context, [8] and [19] provide a sharp demarcation of the distributional conditions for a so-called reduced-load equivalence to hold. We expect that in general there is a complicated tradeoff between the assumptions on the light-tailed flow and the conditions imposed on the heavy-tailed flow.

VII. Preliminary Results Heavy-Tailed Flow

To determine the behavior of $\mathbb{P}(T_2^{1-\rho_1} > x/(\hat{\rho}_1 - \phi_1))$ as $x \to \infty$, we will reduce the space of all relevant sample paths to a single most likely scenario, which occurs with overwhelming probability. In this section, we establish some preliminary results which we will use to neglect the contribution of all nondominant scenarios.

Large-deviations arguments for heavy-tailed distributions suggest that a persistent backlog as associated with the event $T_2^{1-\rho_1} > x/(\hat{\rho}_1 - \phi_1)$, for large x, is most likely due to just a single large burst. To formalize this idea, we first introduce some additional notation. A burst is called large if the size exceeds $\kappa\hat{x}$, with $\kappa > 0$ some small constant, independent of $\hat{x}$. Denote by $\mathcal{N}_{\kappa\hat{x}}[l, r]$ the number of large bursts of flow 2 arriving in the time interval $[l, r]$. Define $N(t) := \{n : U_{2,0}^r + \sum_{i=1}^n U_{2,i} \leq t\}$ as the total number of

bursts of flow 2 arriving in the time interval $[0, t]$. An upper bound for this process is given by

$$N(t) \leq N_U(t) := \left\{ n : \sum_{i=1}^{n} U_{2,i} \leq t \right\} + 1$$

with $U_{2,i}$ i.i.d. random variables representing interarrival times of flow 2.

We now state a crucial lemma which will allow us to limit the attention to large bursts and replace all remaining traffic activity by its average rate. The lemma is a minor modification of Lemma 3 in [31].

Lemma 7.1: Let $S_n = X_1 + \ldots + X_n$ be a random walk with i.i.d. step sizes such that $\mathbb{E}X_1 < 0$ and $\mathbb{E}X_1^p < \infty$ for some $p > 1$. Then, for any $\mu < \infty$, there exists a $\kappa^* > 0$ and a function $\phi(\cdot) \in \mathcal{R}_{-\mu}$ such that for all $\kappa \in (0, \kappa^*]$,

$$\mathbb{P}(S_n > \hat{x} | X_i \leq \kappa\hat{x}, i = 1, \ldots, n) \leq \phi(\hat{x})$$

for all n and $\hat{x}$.

Note that if X_i is the difference of two nonnegative independent random variables X_i^1 and X_i^2, then the lemma remains valid if the X_i's are replaced by the X_i^1's.

We now use the above lemma to show that the workload of flow 2 cannot significantly deviate from the normal drift over intervals of the order $\hat{x}$ when there are no large bursts.

Lemma 7.2: For any $\eta > 0$, $\theta > 0$, there exists a $\kappa^* > 0$ such that for all $\kappa \in (0, \kappa^*]$,

$$\mathbb{P}\left(T_2\left(\left(\theta - (\phi_2 - \rho_2)\eta\right)\hat{x}\right) > \eta\hat{x}, \mathcal{N}_{\kappa\hat{x}}[0, \eta\hat{x}] = 0\right) = o\left(\mathbb{P}\left(B_2^r > \hat{x}(\phi_2 - \rho_2)\right)\right)$$

as $\hat{x} \to \infty$.

Proof: The event $T_2((\theta-(\phi_2-\rho_2)\eta)\hat{x}) > \eta\hat{x}$ means that

$$\inf_{0 \leq u \leq \eta\hat{x}} \{A_2(0, u) - \phi_2 u\} > \left(\theta - (\phi_2 - \rho_2)\eta\right)\hat{x}$$

which in particular implies that

$$A_2(0, \eta\hat{x}) - \phi_2\eta\hat{x} > \left(\theta - (\phi_2 - \rho_2)\eta\right)\hat{x}$$

or equivalently, $A_2(0, \eta\hat{x}) - (\rho_2 + \theta/2\eta)\eta\hat{x} > \theta\hat{x}/2$, so that

$$\sup_{0 \leq u \leq \eta\hat{x}} \left\{ A_2(0, u) - \left(\rho_2 + \frac{\theta}{2\eta}\right) u \right\} > \frac{\theta\hat{x}}{2}.$$

Now, let $S_n := X_1 + \ldots + X_n$ be a random walk with step sizes $X_i := B_{2,i} - (\rho_2 + \theta/2\eta)U_{2,i}$, with $U_{2,i}$ and $B_{2,i}$ i.i.d. random variables representing the interarrival times and burst sizes of flow 2, respectively. Note that X_i represents the net increase in the workload in a queue with service rate $\rho_2 + \theta/2\eta$ between two consecutive bursts, and that $\mathbb{E}X_i < 0$. Because of the sawtooth nature of the process $\{A_2(0, u) - (\rho_2 + \theta/2\eta)u\}$, we have

$$\sup_{0 \leq u \leq t} \left\{ A_2(0, u) - \left(\rho_2 + \frac{\theta}{2\eta}\right) u \right\} \leq B_{2,0} + \sup_{1 \leq n \leq N(t)} S_n.$$

Thus

$$\begin{aligned}
&\mathbb{P}(T_2((\theta - (\phi_2 - \rho_2)\eta)\hat{x}) > \eta\hat{x}, \mathcal{N}_{\kappa\hat{x}}[0, \eta\hat{x}] = 0) \\
&\quad \leq \mathbb{P}\left(B_{2,0} + \sup_{1 \leq n \leq N(\eta\hat{x})} S_n \geq \frac{\theta\hat{x}}{2}, \mathcal{N}_{\kappa\hat{x}}[0, \eta\hat{x}] = 0 \right) \\
&\quad \leq \mathbb{P}\left(B_{2,0} + \sup_{1 \leq n \leq N(\eta\hat{x})} S_n \geq \frac{\theta\hat{x}}{2} | \mathcal{N}_{\kappa\hat{x}}[0, \eta\hat{x}] = 0 \right) \\
&\quad \leq \mathbb{P}\left(B_{2,0} + \sup_{1 \leq n \leq N(\eta\hat{x})} S_n \geq \frac{\theta\hat{x}}{2} | B_{2,i} \leq \kappa\hat{x}, i \geq 0 \right) \\
&\quad \leq \mathbb{P}\left(\sup_{1 \leq n \leq N(\eta\hat{x})} S_n \geq \left(\frac{\theta}{2} - \kappa\right)\hat{x} | B_{2,i} \leq \kappa\hat{x}, i \geq 1 \right) \\
&\quad \leq \mathbb{P}\left(\sup_{1 \leq n \leq (\lambda_2 + \epsilon)\eta\hat{x}} S_n \geq \left(\frac{\theta}{2} - \kappa\right)\hat{x} | B_{2,i} \leq \kappa\hat{x}, i \geq 1 \right) \\
&\qquad + \mathbb{P}(N(\eta\hat{x}) > (\lambda_2 + \epsilon)\eta\hat{x}) \\
&\quad \leq \sum_{i=1}^{(\lambda_2 + \epsilon)\eta\hat{x}} \mathbb{P}\left(S_n \geq \left(\frac{\theta}{2} - \kappa\right)\hat{x} | B_{2,i} \leq \kappa\hat{x}, i = 1, \ldots, n \right) \\
&\qquad + \mathbb{P}(N(\eta\hat{x}) > (\lambda_2 + \epsilon)\eta\hat{x}).
\end{aligned}$$

The second term decays exponentially fast as $\hat{x} \to \infty$. According to Lemma 7.1, there exists a $\kappa^* > 0$ and a function $\phi(\cdot) \in \mathcal{R}_{-\mu}$, $\mu > \nu_2$, such that for all $\kappa \in (0, \kappa^*]$, each of the probabilities in the first term is upper bounded by $\phi(\hat{x})$. The statement then follows. □

We now prove that it is relatively unlikely for flow 2 to generate two large bursts in an interval of order $\hat{x}$.

Lemma 7.3: For any $\alpha < 1$, $\kappa > 0$,

$$\mathbb{P}\left(\mathcal{N}_{\kappa\hat{x}}[0, (1-\alpha)\hat{x}] \geq 2\right) = o\left(\mathbb{P}\left(B_2^r > \hat{x}(\phi_2 - \rho_2)\right)\right)$$

as $\hat{x} \to \infty$.

Proof: By definition

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{N}_{\kappa\hat{x}}[0, (1-\alpha)\hat{x}] \geq 2\right) \\
&\quad \leq \mathbb{P}\left(\#\left\{j \in \{1, \ldots, N_U((1-\alpha)\hat{x})\} : B_{2,j} \geq \kappa\hat{x}\right\} \geq 2\right).
\end{aligned}$$

Conditioning on $N_U((1-\alpha)\hat{x})$, this is upper bounded by

$$\mathbb{E}\left[\left(N_U((1-\alpha)\hat{x})\right)^2\right] \mathbb{P}(B_2 \geq \kappa\hat{x})^2.$$

Finally, observe that $\mathbb{E}[(N_U((1-\alpha)\hat{x}))^2]$ is quadratic in $\hat{x}$ for $\hat{x} \to \infty$. □

The following lemma shows that it is not likely for flow 2 to have a workload of at least order $\hat{x}$ at time 0 and to generate at the same time at least one large burst in an interval of order $\hat{x}$.

Lemma 7.4: For any $0 < \xi < 1 - \alpha$, $\zeta > 0$, $\kappa > 0$,

$$\mathbb{P}\left(\mathcal{N}_{\kappa\hat{x}}[\xi\hat{x}, (1-\alpha)\hat{x}] = 1, V_2^c(0) > \zeta\hat{x}\right) = o\left(\mathbb{P}\left(B_2^r > \hat{x}(\phi_2 - \rho_2)\right)\right) \text{ as } \hat{x} \to \infty.$$

Proof: Because of independence, the probability equals

$$\mathbb{P}\left(\mathcal{N}_{\kappa\hat{x}}[\xi\hat{x}, (1-\alpha)\hat{x}] = 1\right) \mathbb{P}\left(V_2^c(0) > \zeta\hat{x}\right).$$

By conditioning upon $N_U((1-\alpha-\xi)\hat{x})$, we have

$$\mathbb{P}(\mathcal{N}_{\kappa\hat{x}}[\xi\hat{x},(1-\alpha)\hat{x}]=1)\leq\mathbb{E}[N_U((1-\alpha-\xi)\hat{x})]\mathbb{P}(B_2>\kappa\hat{x}).$$

As before, the first term is linear in $\hat{x}$ for $\hat{x}\to\infty$. The statement then follows from the fact that $B_2(\cdot)\in\mathcal{R}_{-\nu_2}$ in combination with Theorem 2.1. □

VIII. Backlog Period Heavy-Tailed Flow

In this section, we consider the case where flow 2 generates instantaneous traffic bursts of regularly varying size. The next theorem shows that flow 2 then satisfies Assumptions 6.1–6.3 and that (4) holds.

Theorem 8.1: For any $c>\rho_2$ and $\alpha,\gamma>0$,

$$\begin{aligned}&\mathbb{P}(T_2^c(\gamma\hat{x})>(1+\alpha)\hat{x})\\&\quad\gtrsim\frac{\rho_2}{c-\rho_2}\mathbb{P}(B_2^r>((\phi_2-\rho_2)(1+\alpha)+\gamma)\hat{x})\end{aligned}\tag{16}$$

$$\begin{aligned}&\mathbb{P}(T_2^c(-\gamma\hat{x})>(1-\alpha)\hat{x})\\&\quad\lesssim\frac{\rho_2}{c-\rho_2}\mathbb{P}\left(B_2^r>\left((\phi_2-\rho_2)(1-\alpha)-\gamma\frac{c+\phi_2-2\rho_2}{\phi_2-\rho_2}\right)\hat{x}\right)\end{aligned}\tag{17}$$

and

$$\mathbb{P}\left(T_2^c>\hat{x}\right)\sim\frac{\rho_2}{c-\rho_2}\mathbb{P}\left(B_2^r>\hat{x}(\phi_2-\rho_2)\right).\tag{18}$$

Before giving the formal proof of the above theorem, we first provide an intuitive argument. Consider a queue with service rate ϕ_2 fed by the arrival process of flow 2. In order for the event $T_2^c>\hat{x}$ to occur, the workload must remain positive throughout the interval [0,$\hat{x}$], given that the initial workload is $V_2^c(0)$. Note that the normal drift in the workload is $\rho_2-\phi_2<0$. Thus, there is a "deficit" $(\phi_2-\rho_2)\hat{x}$, which must be compensated for by the initial workload $V_2^c(0)$ plus possibly flow 2 showing above-average activity during the interval [0, $\hat{x}$].

We claim that the most likely way for the gap to be filled is by a large initial workload only, i.e., $V_2^c(0)>(\phi_2-\rho_2)\hat{x}$. This in turn is most probably due to an extremely large burst of flow 2 at some point before time 0, which is consistent with the usual situation for heavy-tailed distributions that a large deviation is caused by just a single exceptional event. Using Theorem 2.1, we see that the probability of this event is indeed exactly the right-hand side of (18).

Note that it is unlikely for the gap to be filled by flow 2 producing extra traffic during the interval $[0,\hat{x}]$, because this would require a large burst arriving almost immediately after time 0. The probability of this event is negligibly small compared to that of $V_2^c(0)>(\phi_2-\rho_2)\hat{x}$. A combination of both is even less likely, since this would amount to two rare events occurring simultaneously.

The above arguments will be formalized in the proof below. We first prove that the event $V_2^c(0)>(\phi_2-\rho_2)\hat{x}$ indeed implies that $T_2^c>\hat{x}$ for large $\hat{x}$, thus obtaining a lower bound for the probability of the latter event. Next, we show that for large $\hat{x}$ the event $V_2^c(0)>(\phi_2-\rho_2)\hat{x}$ is also necessary for $T_2^c>\hat{x}$ to occur, by proving that the probability of all other possible scenarios is negligibly small.

Proof of Theorem 8.1: We start with the proof of (16). We first prove that for any $\alpha,\gamma,\delta,\theta>0$, the event

$$T_2^c(\gamma\hat{x})>(1+\alpha)\hat{x}\tag{19}$$

is implied by the events

$$V_2^c(z^*)>((\phi_2-\rho_2+\delta)(1+\alpha)+\gamma+\theta)\,\hat{x}+cz^*$$

where z^* is the last time before 0 that a burst arrived, and

$$\sup_{0\leq u\leq(1+\alpha)\hat{x}}\{(\rho_2-\delta)u-A_2(0,u)\}\leq\theta\hat{x}.$$

The second event means that $A_2(0,u)\geq(\rho_2-\delta)u-\theta\hat{x}$, for all $u\in[0,(1+\alpha)\hat{x}]$. Thus, for all $u\in[0,(1+\alpha)\hat{x}]$,

$$\begin{aligned}V_2^c(0)+A_2(0,u)-\phi_2u&=V_2^c(z^*)+A_2(z^*,0)-B_2^c(z^*,0)\\&\quad+A_2(0,u)-\phi_2u\\&\geq V_2^c(z^*)-cz^*+A_2(0,u)-\phi_2u\\&>((\phi_2-\rho_2+\delta)(1+\alpha)+\gamma)\,\hat{x}\\&\quad+(\rho_2-\delta)u-\phi_2u\\&=(\phi_2-\rho_2+\delta)\,((1+\alpha)\hat{x}-u)+\gamma\hat{x}\\&\geq\gamma\hat{x}\end{aligned}$$

where the first equality is obtained using (2), and the first inequality using the fact that $B_2^c(z^*,0)\leq cz^*$. Hence

$$\inf\{u\geq0:V_2^c(0)+A_2(0,u)-\phi_2u\leq\gamma\hat{x}\}>(1+\alpha)\hat{x}$$

which gives (19).

Using independence of $V_2^c(z^*)$ and $A_2(0,u)$

$$\begin{aligned}&\mathbb{P}(T_2^c(\gamma\hat{x})>(1+\alpha)\hat{x})\\&\quad\geq\mathbb{P}(V_2^c(z^*)>((\phi_2-\rho_2+\delta)(1+\alpha)+\gamma+\theta)\hat{x}+cz^*)\\&\qquad\times\mathbb{P}\left(\sup_{0\leq u\leq(1+\alpha)\hat{x}}\{(\rho_2-\delta)u-A_2(0,u)\}\leq\theta\hat{x}\right).\end{aligned}$$

Observe that $V_2^c(0)>0$ implies $V_2^c(0)=V_2^c(z^*)-cz^*$, thus

$$\begin{aligned}&\mathbb{P}\left(V_2^c(z^*)>((\phi_2-\rho_2+\delta)(1+\alpha)+\gamma+\theta)\,\hat{x}+cz^*\right)\\&\quad\geq\mathbb{P}\left(V_2^c(0)>((\phi_2-\rho_2+\delta)(1+\alpha)+\gamma+\theta)\,\hat{x}\right)\\&\quad\sim\frac{\rho_2}{c-\rho_2}\mathbb{P}\left(B_2^r>((\phi_2-\rho_2+\delta)(1+\alpha)+\gamma+\theta)\,\hat{x}\right)\end{aligned}$$

where the last term is due to Theorem 2.1. Also, for all α, δ, $\theta>0$, as $\hat{x}\to\infty$,

$$\begin{aligned}&\mathbb{P}\left(\sup_{0\leq u\leq(1+\alpha)\hat{x}}\{(\rho_2-\delta)u-A_2(0,u)\}\leq\theta\hat{x}\right)\\&\qquad\geq\mathbb{P}\left(\sup_{u\geq0}\{(\rho_2-\delta)u-A_2(0,u)\}\leq\theta\hat{x}\right)\to1\end{aligned}$$

since $\mathbb{E}[A_2(0,u)]=\rho_2u$. Thus, for all $\alpha,\gamma,\delta,\theta>0$,

$$\begin{aligned}&\mathbb{P}\left(T_2^c(\gamma\hat{x})>(1+\alpha)\hat{x}\right)\\&\quad\gtrsim\frac{\rho_2}{c-\rho_2}\mathbb{P}\left(B_2^r>((\phi_2-\rho_2+\delta)(1+\alpha)+\gamma+\theta)\,\hat{x}\right).\end{aligned}$$

Letting $\delta,\theta\downarrow0$, and using $B_2^r(\cdot)\in\mathcal{IR}$, (16) follows.

We now turn to the proof of (17). By partitioning, we obtain for any α, γ, ζ, θ, $\kappa > 0$, $w \geq 0$,

$$\begin{aligned}
&\mathbb{P}\left(T_2^c(-\gamma\hat{x}) > (1-\alpha)\hat{x}\right)\\
&\quad = \mathbb{P}\left(T_2^c(-\gamma\hat{x}) > (1-\alpha)\hat{x},\right.\\
&\qquad \left. V_2^c(w) > ((\phi_2-\rho_2)(1-\alpha)-\gamma-\theta)\,\hat{x}-(c-\rho_2)w\right)\\
&\qquad + \mathbb{P}\left(T_2^c(-\gamma\hat{x}) > (1-\alpha)\hat{x},\right.\\
&\qquad\quad \left. V_2^c(w) \leq ((\phi_2-\rho_2)(1-\alpha)-\gamma-\theta)\,\hat{x}-(c-\rho_2)w,\right.\\
&\qquad\quad \left. \mathcal{N}_{\kappa\hat{x}}[0,w] \leq 1, \mathcal{N}_{\kappa\hat{x}}\left[w,(1-\alpha)\hat{x}\right] = 0\right)\\
&\qquad + \mathbb{P}\left(T_2^c(-\gamma\hat{x}) > (1-\alpha)\hat{x},\right.\\
&\qquad\quad \left. V_2^c(w) \leq ((\phi_2-\rho_2)(1-\alpha)-\gamma-\theta)\,\hat{x}-(c-\rho_2)w,\right.\\
&\qquad\quad \left. \mathcal{N}_{\kappa\hat{x}}[0,w] = 0, \mathcal{N}_{\kappa\hat{x}}\left[w,(1-\alpha)\hat{x}\right] = 1, V_2^c(0) \leq \zeta\hat{x}\right)\\
&\qquad + \mathbb{P}\left(T_2^c(-\gamma\hat{x}) > (1-\alpha)\hat{x},\right.\\
&\qquad\quad \left. V_2^c(w) \leq ((\phi_2-\rho_2)(1-\alpha)-\gamma-\theta)\,\hat{x}-(c-\rho_2)w,\right.\\
&\qquad\quad \left. \mathcal{N}_{\kappa\hat{x}}[0,w] = 0, \mathcal{N}_{\kappa\hat{x}}\left[w,(1-\alpha)\hat{x}\right] = 1, V_2^c(0) > \zeta\hat{x}\right)\\
&\qquad + \mathbb{P}\left(T_2^c(-\gamma\hat{x}) > (1-\alpha)\hat{x}, \mathcal{N}_{\kappa\hat{x}}\left[0,(1-\alpha)\hat{x}\right] \geq 2,\right.\\
&\qquad\quad \left. V_2^c(w) \leq ((\phi_2-\rho_2)(1-\alpha)-\gamma-\theta)\,\hat{x}-(c-\rho_2)w\right)
\end{aligned}$$

which is obviously upper bounded by

$$\begin{aligned}
&\mathbb{P}\left(V_2^c(w) > ((\phi_2-\rho_2)(1-\alpha)-\gamma-\theta)\,\hat{x}-(c-\rho_2)w\right) && (20)\\
&\quad + \mathbb{P}\left(T_2^c(-\gamma\hat{x}) > (1-\alpha)\hat{x}, \mathcal{N}_{\kappa\hat{x}}\left[w,(1-\alpha)\hat{x}\right] = 0,\right.\\
&\qquad \left. V_2^c(w) \leq ((\phi_2-\rho_2)(1-\alpha)-\gamma-\theta)\,\hat{x}-(c-\rho_2)w\right) && (21)\\
&\quad + \mathbb{P}\left(T_2^c(-\gamma\hat{x}) > (1-\alpha)\hat{x}, \mathcal{N}_{\kappa\hat{x}}[0,w] = 0, V_2^c(0) \leq \zeta\hat{x}\right) && (22)\\
&\quad + \mathbb{P}\left(\mathcal{N}_{\kappa\hat{x}}\left[w,(1-\alpha)\hat{x}\right] = 1, V_2^c(0) > \zeta\hat{x}\right) && (23)\\
&\quad + \mathbb{P}\left(\mathcal{N}_{\kappa\hat{x}}\left[0,(1-\alpha)\hat{x}\right] \geq 2\right). && (24)
\end{aligned}$$

Take $w = \xi\hat{x}$, with $\xi := (\gamma+\zeta+\theta)/(\phi_2-\rho_2) < 1-\alpha$. We first concentrate on the event $T_2^c(-\gamma\hat{x}) > (1-\alpha)\hat{x}$, which is equivalent to $\inf_{0\leq u\leq(1-\alpha)\hat{x}}\{V_2^c(0)+A_2(0,u)-\phi_2 u\} > -\gamma\hat{x}$. Observe that the following two inequalities hold:

$$\begin{aligned}
&\inf_{0\leq u\leq(1-\alpha)\hat{x}}\{V_2^c(0)+A_2(0,u)-\phi_2 u\}\\
&\qquad \leq V_2^c(0) + \inf_{0\leq u\leq w}\{A_2(0,u)-\phi_2 u\} && (25)
\end{aligned}$$

and

$$\begin{aligned}
&\inf_{0\leq u\leq(1-\alpha)\hat{x}}\{V_2^c(0)+A_2(0,u)-\phi_2 u\}\\
&\qquad \leq V_2^c(0) + \inf_{w\leq u\leq(1-\alpha)\hat{x}}\{A_2(0,u)-\phi_2 u\}\\
&\qquad = V_2^c(0) + \inf_{w\leq u\leq(1-\alpha)\hat{x}}\{A_2(0,w)-\phi_2 w\\
&\qquad\qquad +A_2(w,u)-\phi_2(u-w)\}\\
&\qquad = V_2^c(0)+A_2(0,w)-\phi_2 w\\
&\qquad\quad + \inf_{w\leq u\leq(1-\alpha)\hat{x}}\{A_2(w,u)-\phi_2(u-w)\}\\
&\qquad \leq V_2^c(w)+(c-\phi_2)w\\
&\qquad\quad + \inf_{w\leq u\leq(1-\alpha)\hat{x}}\{A_2(w,u)-\phi_2(u-w)\}. && (26)
\end{aligned}$$

Consider term (20). Using Theorem 2.1, (20) equals

$$\begin{aligned}
&\mathbb{P}\left(V_2^c > ((\phi_2-\rho_2)(1-\alpha)-\gamma-\theta-(c-\rho_2)\xi)\,\hat{x}\right)\\
&\sim \frac{\rho_2}{c-\rho_2}\mathbb{P}\left(B_2^r > ((\phi_2-\rho_2)(1-\alpha)-\gamma-\theta-(c-\rho_2)\xi)\,\hat{x}\right)\\
&= \frac{\rho_2}{c-\rho_2}\times\\
&\quad \mathbb{P}\left(B_2^r > \left((\phi_2-\rho_2)(1-\alpha)-\gamma-\theta-\frac{(c-\rho_2)(\gamma+\zeta+\theta)}{\phi_2-\rho_2}\right)\hat{x}\right).
\end{aligned}$$

Next, consider term (21). Using (26)

$$\inf_{w\leq u\leq(1-\alpha)\hat{x}}\{A_2(w,u)-\phi_2(u-w)\} > -V_2^c(w)-(c-\phi_2)w-\gamma\hat{x},$$

so that (21) $\leq \mathbb{P}(\mathcal{N}_{\kappa\hat{x}}[w,(1-\alpha)\hat{x}] = 0,$

$$\begin{aligned}
&\inf_{w\leq u\leq(1-\alpha)\hat{x}}\{A_2(w,u)\phi_2(u-w)\} > -V_2^c(w)-(c-\phi_2)w-\gamma\hat{x},\\
&V_2^c(w) \leq ((\phi_2-\rho_2)(1-\alpha)-\gamma-\theta)\,\hat{x}-(c-\rho_2)w)\\
&\qquad \leq \mathbb{P}\left(\inf_{w\leq u\leq(1-\alpha)\hat{x}}\{A_2(w,u)-\phi_2(u-w)\}\right.\\
&\qquad\qquad > \theta\hat{x}-(\phi_2-\rho_2)\left((1-\alpha)\hat{x}-w\right),\\
&\qquad\qquad \left.\mathcal{N}_{\kappa\hat{x}}\left[w,(1-\alpha)\hat{x}\right] = 0\right)\\
&\qquad = \mathbb{P}\left(\inf_{0\leq u\leq(1-\alpha)\hat{x}-w}\{A_2(0,u)-\phi_2 u\}\right.\\
&\qquad\qquad > \theta\hat{x}-(\phi_2-\rho_2)\left((1-\alpha)\hat{x}-w\right),\\
&\qquad\qquad \left.\mathcal{N}_{\kappa\hat{x}}\left[0,(1-\alpha)\hat{x}-w\right] = 0\right)\\
&\qquad = \mathbb{P}\left(\inf_{0\leq u\leq(1-\alpha-\xi)\hat{x}}\{A_2(0,u)-\phi_2 u\}\right.\\
&\qquad\qquad > (\theta-(\phi_2-\rho_2)(1-\alpha-\xi))\,\hat{x},\\
&\qquad\qquad \left.\mathcal{N}_{\kappa\hat{x}}\left[0,(1-\alpha-\xi)\hat{x}\right] = 0\right)\\
&\qquad = \mathbb{P}\left(T_2\left(\theta-(\phi_2-\rho_2)(1-\alpha-\xi)\right)\hat{x}\right)\\
&\qquad\qquad > (1-\alpha-\xi)\hat{x}, \mathcal{N}_{\kappa\hat{x}}\left[0,(1-\alpha-\xi)\hat{x}\right] = 0\right).
\end{aligned}$$

Finally, consider term (22). Using (25), $\inf_{0\leq u\leq w}\{A_2(0,u)-\phi_2 u\} > -V_2^c(0)-\gamma\hat{x}$, so that (22) is less than or equal to

$$\begin{aligned}
&\mathbb{P}\left(\inf_{0\leq u\leq w}\{A_2(0,u)-\phi_2 u\} > -V_2^c(0)-\gamma\hat{x},\right.\\
&\quad \left.\mathcal{N}_{\kappa\hat{x}}[0,w] = 0, V_2^c(0) \leq \zeta\hat{x}\right)\\
&\leq \mathbb{P}\left(\inf_{0\leq u\leq w}\{A_2(0,u)-\phi_2 u\} > -(\gamma+\zeta)\hat{x}, \mathcal{N}_{\kappa\hat{x}}[0,w] = 0\right)\\
&= \mathbb{P}\left(\inf_{0\leq u\leq\xi\hat{x}}\{A_2(0,u)-\phi_2 u\}\right.\\
&\qquad \left.> (\theta-(\phi_2-\rho_2)\xi)\,\hat{x}, \mathcal{N}_{\kappa\hat{x}}[0,\xi\hat{x}] = 0\right)\\
&= \mathbb{P}\left(T_2\left((\theta-(\phi_2-\rho_2)\xi)\,\hat{x}\right) > \xi\hat{x}, \mathcal{N}_{\kappa\hat{x}}[0,\xi\hat{x}] = 0\right).
\end{aligned}$$

Now, taking $\eta = 1-\alpha-\xi$ in Lemma 7.2 for (21), taking $\eta = \xi$ in Lemma 7.2 for (22), using Lemma 7.4 for (23), and using Lemma 7.3 for (24), we obtain

$$\begin{aligned}
&\mathbb{P}(T_2^c(-\gamma\hat{x}) > (1-\alpha)\hat{x}) \lesssim \frac{\rho_2}{c-\rho_2}\times\\
&\quad \mathbb{P}\left(B_2^r > \left((\phi_2-\rho_2)(1-\alpha)-\gamma-\theta-\frac{(c-\rho_2)(\gamma+\zeta+\theta)}{\phi_2-\rho_2}\right)\hat{x}\right).
\end{aligned}$$

Letting $\zeta, \theta \downarrow 0$, and using $B_2^r(\cdot) \in \mathcal{IR}$, (17) follows.

Finally, note that (18) follows from (16) and (17) by letting $\alpha, \gamma \downarrow 0$, and using again $B_2^r(\cdot) \in \mathcal{IR}$. □

IX. Conclusion

We analyzed a GPS queue with two flows, one having light-tailed characteristics, and the other one exhibiting heavy-tailed properties. We showed that the workload distribution of the light-tailed flow is asymptotically equivalent to that when served in isolation at its minimum guaranteed rate, multiplied with a certain pre-factor. The pre-factor may be interpreted as the probability that the heavy-tailed flow is backlogged long enough for the light-tailed flow to reach overflow.

In this paper, we have focused on a scenario with two flows. Observe, however, that the light-tailed flow may be thought of as an aggregate flow, given that the class of Markov-modulated fluid input is closed under superposition of independent processes. In the case of instantaneous input, the heavy-tailed flow also may actually represent an aggregate flow, since the superposition of independent Poisson streams with regularly varying bursts produces again a Poisson stream with regularly varying bursts. Unfortunately, the class of on–off sources is clearly not closed under superposition. In fact, the superposition exhibits a fundamentally more complex structure than a single on–off source, which drastically complicates the analysis of the queueing behavior (see [35]).

Despite the above and earlier observations, it would still be interesting to extend the analysis to general scenarios with several light-tailed flows, say, $N_1 \geq 1$ and $N_2 \geq 1$ heavy-tailed flows. In the case $N_1 = 1, N_2 > 1$, we expect that the workload distribution of the light-tailed flow is equivalent to that when served in isolation at its minimum guaranteed rate, multiplied with a certain pre-factor, exactly as before. In the case $N_1 > 1$, $N_2 = 1$, we conjecture that the workload distribution of the light-tailed flows is equivalent to that in an isolated GPS queue consisting of the light-tailed flows only, multiplied again with a pre-factor. Not surprisingly, the two above-described complicating circumstances conspire in scenarios with $N_1 > 1$, $N_2 > 1$.

Appendix A
Proofs

Proof of Corollary 4.1: Using Lemma 4.1, the fact that $A_1(\cdot)$ and $A_2(\cdot)$ have stationary increments and the independence of $A_1(s,t)$ and $A_2(s,t)$, for all $v, w \geq 0$, and y

$$
\begin{aligned}
&\mathbb{P}\left(V_1(t) > x\right) \\
&\quad \geq \mathbb{P}\left(\exists s^* \in [t-w, t], r^* \in [s^*-v, s^*]: \right. \\
&\qquad A_1(s^*, t) - \phi_1(t-s^*) > x, \\
&\qquad A_1(r^*, s^*) - (\rho_1 - \epsilon)(s^* - r^*) \geq -y, \\
&\qquad \left. \inf_{s^* \leq u \leq t} \{A_2(r^*, u) - (1-\rho_1+\epsilon) \right. \\
&\qquad\qquad \left. \times (s^* - r^*) - \phi_2(u - s^*)\} \geq y\right) \\
&\quad = \mathbb{P}\left(\exists s^* \in [-w, 0], r^* \in [s^*-v, s^*]: \right. \\
&\qquad A_1(s^*, 0) + \phi_1 s^* > x, \\
&\qquad \left. A_1(r^*, s^*) - (\rho_1 - \epsilon)(s^* - r^*) \geq -y, \right. \\
&\qquad \left. \inf_{s^* \leq u \leq s^*+w} \{A_2(r^*, u) - (1-\rho_1+\epsilon) \right. \\
&\qquad\qquad \left. \times (s^* - r^*) - \phi_2(u - s^*)\} \geq y\right) \\
&\quad = \mathbb{P}\left(\exists s^* \in [-w, 0]: A_1(s^*, 0) + \phi_1 s^* > x, \right. \\
&\qquad \left. \inf_{s^*-v \leq r \leq s^*} \{A_1(r, s^*) - (\rho_1 - \epsilon)(s^* - r)\} \geq -y, \right. \\
&\qquad \left. \exists r^* \in [s^*-v, s^*]: \right. \\
&\qquad \left. \inf_{s^* \leq u \leq s^*+w} \{A_2(r^*, u) - (1-\rho_1+\epsilon) \right. \\
&\qquad\qquad \left. \times (s^* - r^*) - \phi_2(u - s^*)\} \geq y\right) \\
&\quad \geq \mathbb{P}\left(\exists s^* \in [-w, 0]: A_1(s^*, 0) + \phi_1 s^* > x\right) \\
&\qquad \times \mathbb{P}\left(\inf_{s^*-v \leq r \leq s^*} \{A_1(r, s^*) - (\rho_1 - \epsilon)(s^* - r)\} \right. \\
&\qquad\qquad \left. \geq -y | A_1(s^*, 0) + \phi_1 s^* > x, \right. \\
&\qquad\qquad \left. \text{with } s^*: A_1(s^*, 0) + \phi_1 s^* > x\right) \\
&\qquad \times \mathbb{P}\left(\exists r^* \in [s^*-v, s^*] \right. \\
&\qquad\qquad \left. (\text{with } s^*: A_1(s^*, 0) + \phi_1 s^* > x): \right. \\
&\qquad\qquad \left. \inf_{s^* \leq u \leq s^*+w} \{A_2(r^*, u) - (1-\rho_1+\epsilon) \right. \\
&\qquad\qquad\qquad \left. \times (s^* - r^*) - \phi_2(u - s^*)\} \geq y\right) \\
&\quad = \mathbb{P}\left(\exists s \in [0, w]: A_1(-s, 0) - \phi_1 s > x\right) \\
&\qquad \times \mathbb{P}\left(\sup_{s^*-v \leq r \leq s^*} \{(\rho_1 - \epsilon)(s^* - r) - A_1(r, s^*)\} \leq y \right| \\
&\qquad\qquad A_1(s^*, 0) + \phi_1 s^* > x, \\
&\qquad\qquad \left. \text{with } s^*: A_1(s^*, 0) + \phi_1 s^* > x\right) \\
&\qquad \times \mathbb{P}\left(\exists r \in [0, v]: \inf_{0 \leq u \leq w} \{A_2(-r, u) \right. \\
&\qquad\qquad \left. -(1-\rho_1+\epsilon)r - \phi_2 u\} \geq y\right) \\
&\quad = \mathbb{P}\left(\sup_{0 \leq s \leq w} \{A_1(-s, 0) - \phi_1 s\} > x\right) P^{\rho_1 - \epsilon}(s^*, v, x, y) \\
&\qquad \times \mathbb{P}\left(\sup_{0 \leq r \leq v} \inf_{0 \leq u \leq w} \{A_2(-r, u) \right. \\
&\qquad\qquad \left. -(1-\rho_1+\epsilon)r - \phi_2 u\} \geq y\right) \\
&\quad = \mathbb{P}\left(V_1^{\phi_1}[w] > x\right) P^{\rho_1 - \epsilon}(s^*, v, x, y) \\
&\qquad \mathbb{P}\left(\inf_{0 \leq u \leq w} \left\{ \sup_{0 \leq r \leq v} \{A_2(-r, 0) - (1-\rho_1+\epsilon)r\} \right. \right. \\
&\qquad\qquad \left. \left. + A_2(0, u) - \phi_2 u\right\} \geq y\right) \\
&\quad = \mathbb{P}\left(V_1^{\phi_1}[w] > x\right) P^{\rho_1 - \epsilon}(s^*, v, x, y) \\
&\qquad \times \mathbb{P}\left(T_2^{1-\rho_1+\epsilon}(v, y) > w\right).
\end{aligned}
$$

Taking $w = ((1+\alpha)x)/(\hat{\rho}_1 - \phi_1)$ completes the proof. □

Proof of Corollary 4.2: Using Lemma 4.2, the independence of $A_1(s,t)$ and $A_2(s,t)$, and the fact that $A_1(\cdot)$ and $A_2(\cdot)$ have stationary increments, for all $w \geq 0$ and y (the numbers indicate the events in the corresponding equations in Lemma 4.2)

$$
\begin{aligned}
&\mathbb{P}\left(V_1(t) > x\right)\\
&\quad \leq \mathbb{P}\left((7) \wedge \{(8) \vee (9) \vee (10)\}\right)\\
&\quad \leq \mathbb{P}\left((7),(8)\right) + \mathbb{P}\left((7),(9)\right) + \mathbb{P}\left((7),(10)\right)\\
&\quad \leq \mathbb{P}\left((7),(8)\right) + \mathbb{P}\left((7)\right) + \mathbb{P}\left((7),(10)\right)\\
&\quad = \mathbb{P}\left(\exists r^* \leq s^* \leq t : A_1(s^*,t) - \phi_1(t-s^*) > x,\right.\\
&\qquad \left. A_1(r^*,s^*) - (\rho_1+\epsilon)(s^*-r^*) > y\right)\\
&\qquad + \mathbb{P}\left(V_1^{\phi_1}(t) > x+y\right)\\
&\qquad + \mathbb{P}\left(\exists r^* \leq s^* \leq t : A_1(s^*,t) - \phi_1(t-s^*) > x,\right.\\
&\qquad\quad \left. \inf_{s^* \leq u \leq t}\{A_2(r^*,u) - (1-\rho_1-\epsilon)(s^*-r^*)\right.\\
&\qquad\qquad \left. -\phi_2(u-s^*)\} > -2y\right)\\
&\quad = \mathbb{P}\left(\exists r^* \leq s^* \leq 0 : A_1(s^*,0) + \phi_1 s^* > x,\right.\\
&\qquad \left. A_1(r^*,s^*) - (\rho_1+\epsilon)(s^*-r^*) > y\right)\\
&\qquad + \mathbb{P}\left(V_1^{\phi_1}(0) > x+y\right)\\
&\qquad + \mathbb{P}\left(\exists r^* \leq s^* \leq 0 : A_1(s^*,0) + \phi_1 s^* > x,\right.\\
&\qquad\quad \left. \inf_{s^* \leq u \leq 0}\{A_2(r^*,u) - (1-\rho_1-\epsilon)(s^*-r^*)\right.\\
&\qquad\qquad \left. -\phi_2(u-s^*)\} > -2y\right)\\
&\quad = \mathbb{P}\left(\exists s^* \leq 0 : A_1(s^*,0) + \phi_1 s^* > x,\right.\\
&\qquad \left. \sup_{r \leq s^*}\{A_1(r,s^*) - (\rho_1+\epsilon)(s^*-r)\} > y\right)\\
&\qquad + \mathbb{P}\left(V_1^{\phi_1}(0) > x+y\right)\\
&\qquad + \mathbb{P}\left(\exists s^* \leq 0 :\right.\\
&\qquad\quad \sup_{r \leq s^*}\inf_{s^* \leq u \leq 0}\{A_2(r,u) - (1-\rho_1-\epsilon)(s^*-r)\\
&\qquad\qquad -\phi_2(u-s^*)\} > -2y,\\
&\qquad\quad \left. A_1(s^*,0) + \phi_1 s^* > x\right)\\
&\quad \leq \mathbb{P}\left(\exists s^* \leq 0 : A_1(s^*,0) + \phi_1 s^* > x\right)\\
&\qquad \times \mathbb{P}\left(\sup_{r \leq s^*}\{A_1(r,s^*) - (\rho_1+\epsilon)(s^*-r)\} > y \middle|\right.\\
&\qquad\quad \left. A_1(s^*,0) + \phi_1 s^* > x\right) + \mathbb{P}\left(V_1^{\phi_1}(0) > x+y\right)\\
&\qquad + \mathbb{P}\left(\exists s^* \in [-w,0] : A_1(s^*,0) + \phi_1 s^* > x\right)\\
&\qquad + \mathbb{P}\left(\exists s^* \leq -w :\right.\\
&\qquad\quad \sup_{r \leq s^*}\inf_{s^* \leq u \leq s^*+w}\{A_2(r,u) - (1-\rho_1-\epsilon)(s^*-r)\\
&\qquad\qquad -\phi_2(u-s^*)\} > -2y,\\
&\qquad\quad \left. A_1(s^*,0) + \phi_1 s^* > x\right)\\
&\quad = \mathbb{P}\left(\exists s \geq 0 : A_1(-s,0) - \phi_1 s > x\right) Q^{\rho_1+\epsilon}(s^*,x,y)\\
&\qquad + \mathbb{P}\left(V_1^{\phi_1}(0) > x+y\right)\\
&\qquad + \mathbb{P}\left(\exists s \in [0,w] : A_1(-s,0) - \phi_1 s > x\right)\\
&\qquad + \mathbb{P}\left(\exists s^* \leq -w : A_1(s^*,0) + \phi_1 s^* > x\right)\\
&\qquad \times \mathbb{P}\left(\sup_{r \leq s^*}\inf_{s^* \leq u \leq s^*+w}\right.\\
&\qquad\quad \left. \{A_2(r,u) - (1-\rho_1-\epsilon)(s^*-r)\right.\\
&\qquad\qquad \left. -\phi_2(u-s^*)\} > -2y\right)\\
&\quad \leq \mathbb{P}\left(\sup_{s \geq 0}\{A_1(-s,0) - \phi_1 s\} > x\right) Q^{\rho_1+\epsilon}(s^*,x,y)\\
&\qquad + \mathbb{P}\left(V_1^{\phi_1}(0) > x+y\right)\\
&\qquad + \mathbb{P}\left(\sup_{0 \leq s \leq w}\{A_1(-s,0) - \phi_1 s\} > x\right)\\
&\qquad + \mathbb{P}\left(\exists s \geq 0 : A_1(-s,0) - \phi_1 s > x\right)\\
&\qquad \times \mathbb{P}\left(\sup_{r \geq 0}\inf_{0 \leq u \leq w}\{A_2(-r,u)\right.\\
&\qquad\qquad \left. -(1-\rho_1-\epsilon)r - \phi_2 u\} > -2y\right)\\
&\quad = \mathbb{P}\left(V_1^{\phi_1} > x\right) Q^{\rho_1+\epsilon}(s^*,x,y)\\
&\qquad + \mathbb{P}\left(V_1^{\phi_1}(0) > x+y\right) + \mathbb{P}\left(V_1^{\phi_1}[w] > x\right)\\
&\qquad + \mathbb{P}\left(\sup_{s \geq 0}\{A_1(-s,0) - \phi_1 s\} > x\right)\\
&\qquad \times \mathbb{P}\left(\inf_{0 \leq u \leq w}\left\{\sup_{r \geq 0}\{A_2(-r,0)\right.\right.\\
&\qquad\qquad \left.\left. -(1-\rho_1-\epsilon)r + A_2(0,u)\right.\right.\\
&\qquad\qquad \left.\left. -\phi_2 u\}\right\} > -2y\right)\\
&\quad = \mathbb{P}\left(V_1^{\phi_1} > x\right) Q^{\rho_1+\epsilon}(s^*,x,y)\\
&\qquad + \mathbb{P}\left(V_1^{\phi_1}(0) > x+y\right) + \mathbb{P}\left(V_1^{\phi_1}[w] > x\right)\\
&\qquad + \mathbb{P}\left(V_1^{\phi_1} > x\right)\mathbb{P}\left(T_2^{1-\rho_1-\epsilon}(-2y) > w\right).
\end{aligned}
$$

Taking $w = ((1-\alpha)x)/(\hat{\rho}_1 - \phi_1)$ completes the proof. □

REFERENCES

[1] V. Anantharam, "How large delays build up in a GI/G/1 queue," *Queueing Syst.*, vol. 5, pp. 345–368, 1988.

[2] A. Arvidsson and P. Karlsson, "On traffic models for TCP/IP," in *Teletraffic Engineering in a Competitive World, Proc. ITC-16*, P. Key and D. Smith, Eds. Amsterdam, The Netherlands: North-Holland, 1999, pp. 457–466.

[3] J. Beran, R. Sherman, M. S. Taqqu, and W. Willinger, "Long-range dependence in variable-bit-rate video traffic," *IEEE Trans. Commun.*, vol. 43, pp. 1566–1579, Feb./Mar./Apr. 1995.

[4] N. H. Bingham, C. M. Goldie, and J. L. Teugels, *Regular Variation*. Cambridge, U.K.: Cambridge Univ. Press, 1987.

[5] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An architecture for differentiated services," IETF, RFC 2475, 1998.

[6] S. C. Borst, O. J. Boxma, and P. R. Jelenković, "Reduced-load equivalence and induced burstiness in GPS queues with long-tailed traffic flows," *Queueing Syst.*, vol. 43, pp. 273–306, 2003.

[7] S. C. Borst, O. J. Boxma, and M. J. G. van Uitert, "The asymptotic workload behavior of two coupled queues," *Queueing Syst.*, vol. 43, pp. 81–102, 2003.

[8] S. C. Borst, K. Dębicki, and A. P. Zwart, "Subexponential asymptotics of hybrid fluid and ruin models," Eindhoven Univ. Technol., Eindhoven, The Netherlands, Tech. Rep., 2003.

[9] S. C. Borst, M. Mandjes, and M. J. G. van Uitert. (2001) Generalized processor sharing queues with heterogeneous traffic classes. Ctr. for Mathematics and Computer Science (CWI), Amsterdam, The Netherlands. [Online]. Available: http://www.cwi.nl/static/publications/reports/abs/PNA-R0106.html

[10] S. C. Borst and A. P. Zwart, "A reduced-peak equivalence for queues with a mixture of light-tailed and heavy-tailed input flows," *Adv. Appl. Prob.*, vol. 35, pp. 793–805, 2003.

[11] O. J. Boxma, Q. Deng, and A. P. Zwart, "Waiting-time asymptotics for the M/G/2 queue with heterogeneous servers," *Queueing Syst.*, vol. 40, pp. 5–31, 2002.
[12] J. Cao and K. Ramanan, "A Poisson limit for buffer overflow probabilities," in *Proc. IEEE INFOCOM*, New York, 2002, pp. 994–1003.
[13] M. Crovella and A. Bestavros, "Self-similarity in World Wide Web traffic: Evidence and possible causes," in *Proc. ACM Sigmetrics*, 1996, pp. 160–169.
[14] A. I. Elwalid and D. Mitra, "Analysis and design of rate-based congestion control of high speed networks, I: Stochastic fluid models, access regulation," *Queueing Syst.*, vol. 9, pp. 29–64, 1991.
[15] ——, "Effective bandwidth of general Markovian traffic sources and admission control of high speed networks," *IEEE/ACM Trans. Networking*, vol. 1, pp. 329–343, June 1993.
[16] P. W. Glynn and W. Whitt, "Logarithmic asymptotics for steady-state tail probabilities in a single-server queue," *J. Appl. Prob.*, vol. 31A, pp. 131–156, 1994.
[17] M. Grossglauser and J.-C. Bolot, "On the relevance of long-range dependence in network traffic," *IEEE/ACM Trans. Networking*, vol. 7, pp. 629–640, Oct. 1999.
[18] P. R. Jelenković and A. A. Lazar, "Asymptotic results for multiplexing subexponential on-off processes," *Adv. Appl. Prob.*, vol. 31, pp. 394–421, 1999.
[19] P. R. Jelenković, P. Momčilović, and A. P. Zwart, "Reduced-load equivalence and subexponentiality," Queueing Syst., to be published.
[20] G. Kesidis, J. Walrand, and C.-S. Chang, "Effective bandwidths for multi-class Markov fluids and other ATM sources," *IEEE/ACM Trans. Networking*, vol. 1, pp. 424–428, Aug. 1993.
[21] C. Kotopoulos, N. Likhanov, and R. R. Mazumdar, "Asymptotic analysis of the GPS system fed by heterogeneous long-tailed sources," in *Proc. IEEE INFOCOM*, Anchorage, AK, 2001, pp. 299–308.
[22] L. Kosten, "Liquid models for a type of information buffer problem," *Delft Progr. Rep.*, vol. 11, pp. 71–86, 1986.
[23] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of Ethernet traffic (extended version)," *IEEE/ACM Trans. Networking*, vol. 2, pp. 1–15, Feb. 1994.
[24] M. Mandjes and J.-H. Kim, "Large deviations for small buffers: An insensitivity result," *Queueing Syst.*, vol. 37, pp. 349–362, 2001.
[25] M. Mandjes and A. Ridder, "Finding the conjugate of Markov fluid processes," *Prob. Eng. Inform. Sci.*, vol. 9, pp. 297–315, 1995.
[26] L. Massoulié, "Large deviations estimates for polling and weighted fair queueing service systems," *Adv. Perform. Anal.*, vol. 2, pp. 103–128, 1999.
[27] A. G. Pakes, "On the tails of waiting-time distributions," *J. Appl. Prob.*, vol. 12, pp. 555–564, 1975.
[28] A. K. Parekh and R. G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: the single node case," *IEEE/ACM Trans. Networking*, vol. 1, pp. 344–357, June 1993.
[29] K. Park and W. Willinger, Eds., *Self-Similar Network Traffic and Performance Evaluation.* New York: Wiley, 2000.
[30] A. Paxson and S. Floyd, "Wide area traffic: the failure of Poisson modeling," *IEEE/ACM Trans. Networking*, vol. 3, pp. 226–244, June 1995.
[31] S. Resnick and G. Samorodnitsky, "Activity periods of an infinite server queue and performance of certain heavy tailed fluid queues," *Queueing Syst.*, vol. 33, pp. 43–71, 1999.
[32] B. Ryu and A. Elwalid, "The importance of long-range dependence of VBR video traffic in ATM traffic engineering: myths and realities," *Comput. Commun. Rev.*, vol. 13, pp. 1017–1027, 1996.
[33] K. Sigman, Ed., *Queueing Syst.*, 1999, vol. 33, Special Issue on Queues with Heavy-Tailed Distributions.
[34] Z.-L. Zhang, "Large deviations and the generalized processor sharing scheduling for a multiple-queue system," *Queueing Syst.*, vol. 28, pp. 349–376, 1998.
[35] A. P. Zwart, S. C. Borst, and M. Mandjes, "Exact queueing asymptotics for multiple heavy-tailed on-off flows," in *Proc. IEEE INFOCOM*, Anchorage, AK, 2001, pp. 279–288.

Sem Borst received the M.Sc. degree in applied mathematics from the University of Twente, Enschede, The Netherlands, in 1990 and the Ph.D. degree from the University of Tilburg, Tilburg, The Netherlands, in 1994.

In 1994, he was a Visiting Scholar at the Statistical Laboratory of the University of Cambridge, Cambridge, U.K. In 1995, he joined the Mathematics of Networks and Systems Department, Bell Laboratories, Lucent Technologies, Murray Hill, NJ, as a Member of Technical Staff. Since 1998, he has also been a member of the Center for Mathematics and Computer Science (CWI), Amsterdam, The Netherlands. In addition, he has a part-time appointment as a Professor of Stochastic Operations Research at Eindhoven University of Technology, Eindhoven, The Netherlands. His main research interests are in the performance evaluation of communication networks and computer systems.

Michel Mandjes received the M.Sc. degrees in mathematics and econometrics and the Ph.D. degree, all from the Free University, Amsterdam, The Netherlands.

After having worked at KPN Research and Bell Laboratories, Lucent Technologies, he currently has a joint position as Department Head with the Center for Mathematics and Computer Science (CWI), Amsterdam, The Netherlands, and as a Full Professor at the University of Twente, Enschede, The Netherlands. His research interests include large deviations analysis of multiplexing systems, queueing theory, traffic management and control in IP networks, and pricing in multiservice networks.

Miranda van Uitert received the M.A. degree in econometrics from Tilburg University, Tilburg, The Netherlands, in 1999. Since 1999, she has been working toward the Ph.D. degree in the Probability, Networks, and Algorithms Department of the Center for Mathematics and Computer Science (CWI), Amsterdam, The Netherlands.

In 2000 and 2001, she had a part-time appointment at KPN Research, The Netherlands. Her main research interests are in queueing theory and its application in the performance analysis of communication networks.