Fluid queues with heavy-tailed M/G/\infty input

Citation for published version (APA):

Document status and date:
Published: 01/01/2001

Document Version:
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:
• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher’s website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl
providing details and we will investigate your claim.
Fluid queues with heavy-tailed $M/G/\infty$ input

S.G. Borst
A.P. Zwart

department of mathematics and computing science
Fluid Queues with Heavy-Tailed $M/G/\infty$ Input

Sem Borst*,†‡, Bert Zwart*

*Department of Mathematics & Computing 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 fluid queue fed by several heterogeneous $M/G/\infty$ input processes with regularly varying session lengths. Under fairly mild assumptions, we derive the exact asymptotic behavior of the stationary workload distribution. As a by-product, we obtain asymptotically tight bounds for the transient workload distribution.

The results are strongly inspired by the large-deviations idea that overflow is typically due to some minimal combination of extremely long concurrent sessions causing positive drift. The typical configuration of long sessions is identified through a simple integer program, paving the way for the exact computation of the asymptotic workload behavior. The calculations provide crucial insight in the typical overflow scenario.

2000 Mathematics Subject Classification: 60K25 (primary), 60F10, 90B18, 90B22 (secondary).

Keywords and Phrases: fluid queues, heavy-tailed input, infinite-server queue, large deviations, transient analysis, workload distribution.
1 Introduction

Fluid models have found widespread use as a versatile approach for analyzing burst-scale traffic behavior in high-speed communication networks. The canonical model comprises a superposition of On-Off sources, independently alternating between activity phases (bursts) and silence periods. When active, each source generates traffic at some constant rate. An alternative yet closely related model is that of $M/G/\infty$ input, where sessions arrive as a Poisson process, and remain in the system for a randomly distributed period of time. While in the system, each session generates traffic at some constant rate. Note that the number of active sessions behaves as the number of customers in an $M/G/\infty$ system, hence the term $M/G/\infty$ input. An $M/G/\infty$ input process may also be viewed as the limit of the superposition of On-Off sources when the number of sources grows large, and the fraction of On-time gets correspondingly small as shown in Jelenković & Lazar [11].

While incorporating session-level dynamics, the $M/G/\infty$ model avoids the intricate temporal dependence structure of ordinary On-Off sources. At the same time, the $M/G/\infty$ model retains the usual versatility of fluid models in covering a wide spectrum of possible traffic characteristics through the distribution of the activity periods. As is typically the case for fluid queues, the distribution of the activity periods strongly influences the workload behavior. In particular, there is a sharp dichotomy in the qualitative workload behavior, depending on whether the activity periods have light-tailed or heavy-tailed characteristics. It turns out that for light-tailed (exponentially bounded) activity periods, the tail distribution of the workload generally exhibits exponential decay as well, see Parulekar & Makowski [23]. These results fall in the realm of logarithmic asymptotics for general single-server queues with light-tailed arrival processes as in Duffield & O'Connell [7] and Glynn & Whitt [9].

The past few years have witnessed a surging interest in fluid models with heavy-tailed activity periods. Extensive measurements in high-speed communication networks indicate that bursty traffic behavior may extend over a wide range of time scales, manifesting itself in long-range dependence and self-similarity, see Leland et al. [16] and Paxson & Floyd [24]. The occurrence of these phenomena is commonly attributed to extreme variability and heavy-tailed characteristics in the underlying activity patterns (connection times, file sizes, scene lengths), see Beran et al. [2], Crovella & Bestavros [5] and Willinger et al. [28]. Fluid models with heavy-tailed activity periods provide a natural paradigm for capturing these characteristics, see in particular Likhanov et al. [19] for the case of $M/G/\infty$ input. We refer to Boxma & Dumas [4] for a survey paper; see also the recent publications [22] and [27].

Fluid queues with heavy-tailed $M/G/\infty$ input have been extensively studied in several earlier papers. Likhanov [17] and Liu et al. [20] obtain asymptotic lower and upper bounds for the workload distribution. Under a certain peak rate condition, the bounds are shown to be tight (up to a constant factor) for Pareto-distributed session lengths, thus yielding the exact decay rate. The peak rate condition essentially implies that just a single long session is enough to cause overflow. Under roughly similar assumptions, Boxma [3], Jelenković & Lazar [10], and Resnick & Samorodnitsky [26] also determine the corresponding pre-factor, resulting in the
exact workload asymptotics. Duffield [6] obtains logarithmic ‘many-sources’ asymptotics (as opposed to ‘large-buffer’ asymptotics) for a regime where the arrival rate, service rate, and buffer size are scaled up in proportion, see also Mandjes [21].

Recently, several authors have considered heterogeneous heavy-tailed $M/G/\infty$ input, where sessions belong to one of several classes with distinct characteristics (arrival rates, session lengths, peak rates). Likhanov & Mazumdar [18] obtain asymptotic lower and upper bounds for the workload distribution, which are shown to be tight up to a constant factor. Under a similar peak rate condition as described above, the bounds coincide, yielding the exact asymptotics. An elegant treatment of this special case is also given in Jelenković [13]. Remarkably enough, the bounds in [18] are asymptotically exact for finite buffers as well.

As mentioned above, the $M/G/\infty$ model is closely related to the classical model with a fixed set of On-Off sources. Despite some subtle differences, the similarity manifests itself in the qualitative way that overflow occurs for heavy-tailed input, and is also reflected in the tail asymptotics of the workload. For example, the results in [8] and [13] for a fixed set of On-Off sources are reminiscent of the results in [18] for $M/G/\infty$ input. Also, the $M/G/\infty$ asymptotics in [3], [10], and [26] for the special case where a single long session can cause overflow are accompanied (in [3] and [10]) by conceptual counterparts for a scenario where a single regularly varying On-Off source is multiplexed with several light-tailed sources.

It is interesting to observe that the exact workload asymptotics for the $M/G/\infty$ model with infinite buffers have only been obtained under the condition that a single long session is sufficient to cause positive drift. Although technically convenient, this condition is rather restrictive from a practical perspective. The degree of multiplexing is typically so high, that the peak rate of an individual session is relatively small compared to the link rate. Thus, under moderate loading, several long sessions must coincide in order for the drift to turn positive. In the present paper, we derive the exact asymptotic workload behavior under such general circumstances where a combination of several long sessions is involved in causing overflow. Besides the practical relevance, these scenarios are also theoretically challenging, since the combinatorial structure of the overlap of the various sessions significantly adds to the complexity. The analysis unifies and generalizes the results in [13], [18], and [26], and complements the exact tail asymptotics for a fixed set of On-Off sources which were recently obtained in [29].

The remainder of the paper is organized as follows. In Section 2, we present a detailed model description, and state an important preliminary result. Next, we provide some intuitive arguments, and summarize the main results of the paper in Section 3. The arguments are grounded on the large-deviations idea that overflow is typically due to some minimal combination of extremely long concurrent sessions causing positive drift. The typical configuration of long sessions is identified through a simple integer linear program, which corresponds to the set optimization problem defined in [18].

The subsequent sections are devoted to the detailed proofs. In particular, in Section 4, we extend the probabilistic arguments developed in [26], enabling the exact calculation of the asymptotic workload behavior. In addition, the computations provide fundamental insight in the typical overflow scenario.
The analysis in fact focuses on the transient behavior, from which the steady-state asymptotics easily follow after showing in Section 5 that overflow occurs in linear time. As a by-product, we obtain asymptotically tight bounds for the transient workload distribution. The exact transient asymptotics in full generality remain a challenging open problem. In Section 6, we combine the transient and steady-state asymptotics to obtain the limiting distribution of the most probable time to overflow.

2 Model description and preliminaries

In this section, we present a detailed model description, discuss some useful concepts, and state an important auxiliary result.

We first introduce some notational conventions that we use throughout the paper. For any two real functions $f(\cdot)$ and $g(\cdot)$, we write $f(x) \sim g(x)$ to indicate $\lim_{x \to \infty} f(x)/g(x) = 1$, or equivalently, $f(x) = g(x)(1 + o(1))$ as $x \to \infty$. Also, we use $f(x) \lesssim g(x)$ to denote $\limsup_{x \to \infty} f(x)/g(x) \leq 1$. Similarly, $f(x) \gtrsim g(x)$ denotes $\liminf_{x \to \infty} f(x)/g(x) \geq 1$. For any two random variables $X$ and $Y$, we write $X \overset{d}{=} Y$ to indicate that $X$ and $Y$ are equal in distribution. We denote by $\mathbb{N} = \{0, 1, \ldots\}$ the set of non-negative integers. The class of functions which are regularly varying of index $-\alpha$ is denoted by $\mathcal{R}_{-\alpha}$.

2.1 Basic input and workload processes

We consider a fluid queue of unit capacity fed by $K$ heterogeneous $M/G/\infty$ input processes. Class-$k$ sessions arrive as a Poisson process of rate $\lambda_k$, and remain in the system for a random period $B_k$ having distribution $B_k(\cdot)$ with mean $\beta_k$, $k = 1, \ldots, K$. While in the system, each class-$k$ session generates traffic at constant rate $r_k$. We assume that $B_k(\cdot) \in \mathcal{R}_{-\nu_k}$, $\nu_k > 1$, so that $\beta_k < \infty$. (This assumption can be relaxed somewhat, see Remark 3.1 below.) Let $\rho_k := \lambda_k \beta_k r_k$ be the traffic intensity associated with class-$k$ sessions. Define $\bar{\rho}_k := \lambda_k \beta_k = \rho_k/r_k$. Let $\rho := \sum_{k=1}^{K} \rho_k$ be the total traffic intensity. We assume $\rho < 1$ for stability. Denote by $B_k^r(\cdot)$ the distribution of the residual lifetime of $B_k$, and by $B_k^r$ a stochastic variable with that distribution.

Define $A_k(s, t)$ as the amount of class-$k$ traffic generated in the time interval $[s, t]$. Denote by $A(s, t) := \sum_{k=1}^{K} A_k(s, t)$ the total amount of traffic generated in the time interval $[s, t]$. The workload in the system at time $t \geq 0$ is $V(t) := \sup_{0 \leq s \leq t} \{A(s, t) - (t - s)\}$, assuming the system is empty at time $t = 0$. Let $V$ be the weak limit of $V(t)$ for $t \to \infty$.

2.2 Auxiliary processes: separating short and long sessions

One of the first steps of the analysis will be to split the arriving sessions into two groups, short and long ones. In this subsection we introduce some notation for the corresponding processes.
We denote by \( A_{k \leq z}(s, t) \) the amount of traffic generated \((s, t]\) by class-\(k\) sessions of length at most \(z\). The corresponding traffic intensity is denoted by 
\[
\rho_{k \leq z} := \lambda_k r_k \mathbb{E}[B_k \mid B_k \leq z] = \rho_k B_k^r(z).
\]
Define \( A_{s \leq z}(s, t) := \sum_{k=1}^{K} A_{k \leq z}(s, t) \), and 
\[
\rho_{s \leq z} := \sum_{k=1}^{K} \rho_{k \leq z}.
\]
Similarly, we denote by \( A_{k > z}(s, t) \) the amount of traffic generated in \((s, t]\) by class-\(k\) sessions of length exceeding \(z\). The corresponding traffic intensity \( \rho_{k > z} \) is given by 
\[
\rho_{k > z} = \rho_k (1 - B_k^r(z)).
\]
Define \( A_{s > z}(s, t) := \sum_{k=1}^{K} A_{k > z}(s, t) \), and 
\[
\rho_{s > z} := \sum_{k=1}^{K} \rho_{k > z}. \]
Denote \( \hat{\rho}_{k > z} = \rho_{k > z} / \tau_k \).

Denote by \( V_{s \leq z}(t) \) and \( V_{s > z}(t) \) the workload at time \(t\) in a system of capacity \(c\) fed by the processes \( A_{s \leq z}(s, t) \) and \( A_{s > z}(s, t) \), respectively. Note that \( A(t) = A_{s > 0}(t) \) and \( V(t) = V_{s > 0}(t) \).

### 2.3 Representation for the workload

In this subsection we give a convenient representation for the transient and stationary workload. These representations require a detailed description of the input process. Denote by \( N_{k > z}(t) \) the number of class-\(k\) sessions exceeding length \(z\) active at time \(t\). Assume that the processes \( \{N_{k > z}(t)\} \) are stationary (see for instance Kelly [15] for a construction). Note that \( N_{s > z}(t) := (N_{1 > z}(t), \ldots, N_{K > z}(t)) \) has a multi-dimensional Poisson distribution with parameters \( (\hat{\rho}_{1 > z}, \ldots, \hat{\rho}_{K > z}) \), i.e.,

\[
\mathbb{P}\{N_{s > z}(t) = (n_1, \ldots, n_K)\} = \prod_{k=1}^{K} \frac{e^{-\hat{\rho}_{k > z}} \hat{\rho}_{k > z}^{n_k}}{n_k!} = \prod_{k=1}^{K} e^{-\hat{\rho}_{k > z}} \frac{\hat{\rho}_{k > z}^{n_k}}{n_k!} (\mathbb{P}\{B_k^r > z\})^{n_k}. \quad (2.1)
\]

The process \( A_{k > z}(\cdot, \cdot) \) may be related to the process \( N_{k > z}(u) \) as

\[
A_{k > z}(0, t) := r_k \int_{0}^{t} N_{k > z}(u)du.
\]

Using the expression \( V_{s > z}(t) = \sup_{0 \leq s \leq t} \{A_{s > z}(s, t) - c(t-s)\} \) and noting that the process \( A_{s > z}(\cdot, \cdot) \) has stationary and reversible increments, the transient workload may be represented as

\[
V_{s > z}^c(t) = \sup_{0 \leq s \leq t} \{A_{s > z}(s, t) - c(t-s)\} \overset{d}{=} \sup_{0 \leq s \leq t} \{A_{s > z}(0, s) - cs\}.
\]

In the sequel, we proceed similarly as in [26] and [29], and use the latter expression as the definition of \( V_{s > z}^c(t) \). Accordingly, for \( c > \rho_{s > z} \), the stationary workload as \( t \to \infty \) may be expressed as

\[
V_{s > z}^c := \sup_{t \geq 0} \{A_{s > z}(0, t) - ct\}.
\]

Note that these representations should be handled with care, e.g. when considering the joint distribution of \( V_{1}^c(t_1) \) and \( V_{2}^c(t_2) \). This problem however is not analyzed here.
2.4 Bound for truncated heavy-tailed distributions

We conclude the section with an important auxiliary result due to Resnick & Samorodnitsky [25]. This result will be instrumental in taming all 'short' sessions.

**Lemma 2.1** Let $S_n = X_1 + \ldots + X_n$ be a random walk with i.i.d. step sizes such that $E[X_1] < 0$ and $E[(X_1^+)^p] < \infty$ for some $p > 1$. Then, for any $\alpha < \infty$, there exists an $\epsilon^* > 0$ and a function $\phi(\cdot) \in R_{-\alpha}$ such that for $\epsilon \in (0, \epsilon^*]$,

$$P\{S_n > x | X_j \leq \epsilon x, j = 1, \ldots, n\} \leq \phi(x),$$

for all $n$ and all $x$.

Exact asymptotics in the above setting, for both $S_n$ and $\text{sup}_n S_n$ and a regularly varying right tail of $X_1$, have been computed by Jelenković [12]. For our purposes however, it suffices to use the above lemma. Note that if $X_j$ can be represented as the difference of two non-negative independent random variables $X^{(1)}_j$ and $X^{(2)}_j$, then the lemma remains valid if the $X^{(2)}_j$'s are replaced by $X^{(1)}_j$.

3 Main results

In this section we present the main results of the paper, which characterize the exact asymptotic behavior of $P\{V > x\}$ as $x \to \infty$.

3.1 Intuitive arguments

Before formally stating the results, we first provide some intuitive arguments. Large-deviations results for heavy-tailed distributions suggest that a large workload level is typically due to some 'minimal combination' of extremely long overlapping sessions causing positive drift. In a homogeneous context, the typical combination simply consists of the minimal number of long sessions needed for the drift to turn positive. However, in a heterogeneous setting, not only the number of long sessions counts, but also the class characteristics. Note that the number of long sessions required for a positive drift varies with the peak rates $r_k$ of the various classes. In addition, the relative frequency of long sessions differs across the various classes as governed by the tail exponents $\nu_k$.

Informally speaking, the typical combination may be interpreted as the one most likely to occur among those producing positive drift. Specifically, let the typical configuration of long sessions be $n = (n_1, \ldots, n_K)$. For the workload to reach a large level $x$, the associated drift must be strictly positive, i.e.,

$$\sum_{k=1}^{K} n_k r_k + \rho - 1 > 0. \quad (3.1)$$
In addition, the sessions must last for a period of the order \( x \), which happens with probability of the order \( x^{K-1} \).

\[
\mathcal{L} = \sum_{k=1}^{K} n_k (\nu_k - 1). \tag{3.2}
\]

The supposition that \( n = (n_1, \ldots, n_K) \) is the most likely combination, means that it should maximize (3.2) for large values of \( x \), i.e., minimize the exponent \( \sum_{k=1}^{K} n_k (\nu_k - 1) \), subject to the drift condition (3.1). Thus, the most likely configuration of long sessions may be identified as follows.

\[
\begin{align*}
\min_{n_k} \mu &= \sum_{k=1}^{K} n_k (\nu_k - 1) \\
\text{subject to} \quad &\sum_{k=1}^{K} n_k r_k \geq 1 - \rho \\
&n_k \in \mathbb{N}, \quad k = 1, \ldots, K.
\end{align*}
\]

The above integer linear program corresponds to the set optimization problem defined in [18]. In general, the optimal solution cannot be obtained in closed form due to the integrality constraints. However, if the integrality constraints are relaxed, then the optimization problem may be easily solved. The optimal solution is then given by \( n^* = (1 - \rho) e_k / r_k \), with \( k^* := \arg \max_{k=1,\ldots,K} r_k/(\nu_k - 1) \), and \( e_k \) denoting the \( k \)-th unit vector. This suggests that sessions of class \( k^* \) are likely to be involved in the typical configuration of long sessions that causes overflow. This is especially the case when the peak rates \( r_k \) are relatively small compared the slack capacity \( 1 - \rho \), so that the typical combination consists of a relatively large number of sessions. However, in general the optimal combination may include sessions of other classes as well due to the integrality constraints, and in extreme cases may not contain a single session of class \( k^* \) at all.

Let \( S^* \subseteq \mathbb{N}^K \) be the set of optimal solutions of the above linear program (there may be several in general). Denote by \( \mu^* \) the corresponding optimal value. Also, define \( r_{\min} := \min_{n \in S^*} \sum_{k=1}^{K} n_k r_k \).

Throughout the paper, we assume that \( r_{\min} > 1 - \rho \). This assumption ensures that the drift in all plausible overflow scenarios is strictly positive. (In general, some overflow scenarios may involve only zero drift.)

### 3.2 Steady-state workload asymptotics

We now state the central result of the paper, which characterizes the exact asymptotic behavior of the stationary workload distribution. For given \( n \in \mathbb{N}^K \), denote \( d_n := \sum_{k=1}^{K} n_k r_k + \rho - 1 \).

**Theorem 3.1** Assume that \( r_{\min} > 1 - \rho \). Then,

\[
P \{ V > x \} \sim \sum_{n \in S^*} \sum_{j \leq n} \prod_{k=1}^{K} \frac{\hat{p}_k^{n_k}}{j_k!} P_{j,n}(x), \tag{3.3}
\]
where \( j = (j_1, \ldots, j_K) \), and \( P_{j,n}(x) \) satisfies
\[
P_{j,n}(x) \sim \kappa_{j,n} \prod_{k=1}^{K} \mathbb{P}\{B_k > \frac{x}{d_n}\}^{n_k},
\]
for some constant \( \kappa_{j,n} \).

In particular, \( \mathbb{P}\{V > x\} \) is regularly varying of index \( -\mu^* \).

Explicit expressions for \( P_{j,n}(x) \) and \( \kappa_{j,n} \) are given in Subsection 4.5.

**Remark 3.1** Recall that we assumed that \( B_k(\cdot) \in \mathcal{R}_{-\nu_k}, \nu_k > 1 \), for all \( k = 1, \ldots, K \). In fact, Theorem 3.1 continues to hold if \( B_k(\cdot) \) is light-tailed for some \( k \), provided \( B_k(\cdot) \) is regularly varying for at least one class \( k \). (By light-tailed, we mean that \( 1 - B_k(x) = o(x^{-\alpha}) \) as \( x \to \infty \) for any \( \alpha \).) Theorem 3.1 and all the results stated below which follow from it, formally go through if we simply define \( \nu_k := \infty \) in case \( B_k(\cdot) \) is light-tailed.

### 3.3 Single-session overflow scenario

The expressions for the coefficients \( \kappa_{j,n} \) may in principle be computable, but are in general not very explicit. However, as described in the Introduction, rather tractable results are available for scenarios where just a single long session can cause overflow. We now specialize the general result stated in Theorem 3.1 to these scenarios in order to obtain more explicit expressions, and recover these results. Define \( T^* = \{k : e_k \in S^*\} \), with \( e_k \) denoting the \( k \)-th unit vector.

**Theorem 3.2** Assume that \( S^* \subseteq \{e_1, \ldots, e_K\} \). If \( r_{\min} = \min_{k \in T^*} r_k > 1 - \rho \), then
\[
\mathbb{P}\{V > x\} \sim \sum_{k \in T^*} \frac{\rho_k}{1 - \rho} \mathbb{P}\{B_k^{e_k} > \frac{x}{r_k + \rho - 1}\}. \tag{3.4}
\]

This result is obtained in [13] under the condition that \( r_k > 1 - \rho \) and \( B_k(\cdot) \) is of intermediate regular variation for all \( k = 1, \ldots, K \). The discrete-time analogue for Pareto-distributed session lengths may be found in [18].

### 3.4 Single-class input

We now consider the important special case of a single input class, i.e., homogeneous input. For conciseness, we suppress the class index 1. We have \( S^* = \{n^*\} \) and \( \mu^* = n^*(\nu - 1) \), with \( n^* := \lfloor (1 - \rho)/r \rfloor \).

**Theorem 3.3** Assume that \( r_{\min} = n^*r > 1 - \rho \). Then,
\[
\mathbb{P}\{V > x\} \sim \sum_{j=0}^{n^*} \frac{\rho^{n^*}}{j!} P_{j,n^*}(x), \tag{3.5}
\]
where $P_{j,n^*}(x)$ satisfies

$$P_{j,n^*}(x) \sim \kappa_{j,n^*} \mathbb{P}\{B^r > \frac{x}{d_{n^*}}\}^{n^*},$$

for some constant $\kappa_{j,n^*}$.

In particular, $\mathbb{P}\{V > x\}$ is regularly varying of index $-n^*(\nu - 1)$.

An explicit expression for $\kappa_{j,n^*}$ is given in Subsection 4.5.

### 3.5 Single-class input with single-session overflow scenario

Finally, we consider the intersection of single-class input with a single-session overflow scenario. Taking $T^* = \{1\}$ in Theorem 3.2, or $n^* = 1$ in Theorem 3.3, we find

$$\mathbb{P}\{V > x\} \sim \frac{\rho}{1-\rho} \mathbb{P}\{B^r > \frac{x}{r + \rho - 1}\}. \quad (3.6)$$

This result is also obtained in [3], [10], and [26].

**Remark 3.2** It is worth observing that the qualitative resemblance of (3.6) with (3.4) is markedly stronger than with (3.5). Thus, the extension to a multiple-session overflow scenario has greater ramifications than the issue of heterogeneous input. This confirms that the fundamental problem lies in the plurality of the set $S^*$ rather than the heterogeneity of the input or non-uniqueness of the set $S^*$.

**Remark 3.3** It is also interesting to compare (3.6) with the corresponding result for a single On-Off source. Specifically, consider a fluid queue of capacity $c$ fed by a single On-Off source with the same On-periods $B$, Off-periods with mean $1/\lambda'$, peak rate $r'$, fraction Off-time $p = (1 + \lambda' \beta)^{-1}$, and traffic intensity $\rho' = (1 - p)r'$, with $\rho' < c < r'$. Then the asymptotic behavior of the workload is given by [10],

$$\mathbb{P}\{V' > x\} \sim p' \frac{\rho'}{c - \rho'} \mathbb{P}\{B^r > \frac{x}{r' - c}\}. \quad (3.7)$$

Now suppose that we choose $r = r' - \rho' = pr'$, $\lambda' = (1/(\lambda' + \beta))^{-1}$, so that $\rho = \lambda \beta r = (1 - p)r = (1 - p)pr' = \rho p'$, and $c = 1 + \rho' - \rho$. Then (3.7) agrees with (3.6). In other words, if $r + \rho > 1$, then the workload in a queue of unit capacity fed by $M/G/\infty$ input with $\lambda' = (1/(\lambda' + \beta))^{-1}$ and $r = r' - \rho$ is asymptotically equivalent to that in a queue of capacity $c = 1 + \rho' - \rho$ fed by a single On-Off source with the same On-periods $B$, peak rate $r'$, and Off-periods with mean $1/\lambda'$.

This may be understood as follows. In both situations, a large workload level is most likely due to a single extreme event causing a persistent positive drift, either a long session in the $M/G/\infty$ case, or a long On-period in the On-Off case. By assumption, the sessions in the $M/G/\infty$ case have the same distribution as the On-periods in the On-Off case. The chosen parameter values imply that the frequency of sessions and On-periods is also equal. The mean number of On-periods per unit of time is $(1/(\lambda' + \beta))^{-1} = \lambda$, the rate at which sessions arrive.
As a result, the occurrence of long sessions and long On-periods matches. The workload dynamics during long sessions and long On-periods coincide as well. With \(M/G/\infty\) input, the workload has positive drift \(r + \rho - 1\) when a long session is active, and negative drift \(\rho - 1\) otherwise. With On-Off input, the workload increases at rate \(r' - c = r + \rho - 1\) during a long On-period, and decreases approximately at rate \(\rho' - c = \rho - 1\) otherwise. Unfortunately, this equivalence does not seem to extend to more general scenarios.

4 Proof of Theorem 3.1

In this section we analyze the asymptotic behavior of \(\mathbb{P}\{V(ax) > x\}\) for fixed \(a\) and \(x \to \infty\). As the next theorem shows, this directly yields the steady-state asymptotics after letting \(a \to \infty\).

**Theorem 4.1** If \(r_{\min} > 1 - \rho\), then

\[
\lim_{a \to \infty} \lim_{x \to \infty} \frac{\mathbb{P}\{V(ax) > x\}}{\mathbb{P}\{V > x\}} = 1.
\]

The proof of the above theorem is deferred to Section 5.

In order to analyze \(\mathbb{P}\{V(ax) > x\}\), it will be convenient to use the representation

\[
V(ax) = \sup_{0 \leq s \leq ax} \{A(0, s) - s\},
\]

see Subsection 2.3. For the tail behavior of \(\mathbb{P}\{V(ax) > x\}\), similar heuristic arguments apply as those sketched in Subsection 3.1. The only difference is that in general a positive drift alone is not enough for the process \(\{A(0, s) - s\}\) to reach level \(x\) before time \(ax\). Instead, the drift should be at least \(\frac{1}{a}\). Therefore, the integer linear program as formulated in Subsection 3.1 needs to be modified as follows.

\[
\begin{align*}
\min & \quad \mu = \sum_{k=1}^{K} n_k (\nu_k - 1) \\
\text{sub} & \quad \sum_{k=1}^{K} n_k r_k \geq 1 - \rho + \frac{1}{a} \\
& \quad n_k \in \mathbb{N}, \quad k = 1, \ldots, K.
\end{align*}
\]

Let \(S^*_a \subseteq \mathbb{N}^K\) be the set of optimal solutions of the above linear program. Denote by \(\mu^*_a\) the corresponding optimal value. Also, define \(r^*_a := \min_{n \in S^*_a} \sum_{k=1}^{K} n_k r_k\).

The analysis of the tail behavior of \(\mathbb{P}\{V(ax) > x\}\) involves several steps.

- We first separate 'short' and 'long' sessions. A session is called 'long' if it exceeds length \(\varepsilon x\), with \(\varepsilon\) some small positive constant, independent of \(x\). Otherwise, it is called
'short'. We show that the ‘short’ sessions can be asymptotically ignored if the capacity is reduced by \( \rho \), in the sense that for \( \epsilon \) sufficiently small,

\[
P\{V(ax) > x\} \sim P\{V_{>ax}^{1-\rho}(ax) > x\}.
\]

- Next, we determine the typical combination of long sessions involved in causing overflow. Specifically, we prove that, for overflow of level \( x \) to occur within time \( ax \), the configuration of long sessions in the interval \([0, ax]\) must be \( n = (n_1, \ldots, n_K) \), for some \( n \in S^*_a \).

- Subsequently, we identify a stopping time \( \tau_f(\epsilon x) \) (conditional upon the event that the configuration of long sessions is \( n \in S^*_a \)) such that for \( a \) sufficiently large and \( c \) sufficiently close to \( 1 - \rho \),

\[
P\{\sup_{0 \leq s \leq ax} \{A_{>\epsilon x}(0, s) - cs\} > x\} \sim P\{A_{>\epsilon x}(0, \tau_f^n(\epsilon x)) - c\tau_f^n(\epsilon x) > x\}.
\]

- Last, we compute the asymptotic behavior of \( P\{A_{>\epsilon x}(0, \tau_f^n(\epsilon x)) - c\tau_f^n(\epsilon x) > x\} \) as \( x \to \infty \), which involves a rather tedious but straightforward calculation.

Subsections 4.1-4.4 elaborate upon the above four steps, which prepare the way for the proof of Theorem 3.1 in Subsection 4.5. As a by-product of the analysis, we obtain asymptotically tight lower and upper bounds for the transient workload distribution in Subsection 4.6. The various steps involve similar probabilistic arguments as developed in [26] for the special case where a single long session is enough to cause overflow. The first two steps are also used in [18] to derive asymptotic lower and upper bounds for \( P\{V > x\} \) which coincide up to a constant factor. Remarkably enough, these bounds are asymptotically exact for finite buffers. The exact asymptotics for infinite buffers however entail a detailed calculation as in the last two steps listed above.

4.1 Discarding short sessions

As a first step, we separate short and long sessions. We show that – as far as asymptotic behavior is concerned – the short sessions can be deleted if the capacity is reduced by \( \rho \). Formally, we derive asymptotic lower and upper bounds for \( P\{V(ax) > x\} \) of the form

\[
P\{V_{>ax}^{1-\rho+\delta}(ax) > (1 \pm \theta)x\} \text{ for arbitrarily small } \delta, \theta.
\]

We first establish a simple sample-path lower bound. For any \( c > 0 \), define \( Z_{c, \leq t}^{c}(t) := \sup_{0 \leq s \leq t} \{cs - A_{\leq t}(0, s)\} \).

**Proposition 4.1** For any \( c \in (0, \rho_{\leq}) \),

\[
P\{V(t) > x\} \geq P\{V_{>\leq}^{1-c}(t) > x + y\}P\{Z_{c, \leq}^{c}(t) \leq y\}.
\]
Proof
Sample-path wise,

\[
V(t) = \sup_{0 \leq s \leq t} \{A(0, s) - s\} \\
= \sup_{0 \leq s \leq t} \{A_{>z}(0, s) - (1 - c)s + A_{\leq z}(0, s) - cs\} \\
\geq \sup_{0 \leq s \leq t} \{A_{>z}(0, s) - (1 - c)s\} - \sup_{0 \leq s \leq t} \{cs - A_{\leq z}(0, s)\} \\
= V^{1-c}_{>z}(t) - Z^c_{\leq z}(t).
\]

We now use the above sample-path bound to obtain an asymptotic lower bound for \(\mathbb{P}\{V(ax) > x\}\) as \(x \to \infty\).

**Proposition 4.2** For any \(\delta > 0, \epsilon > 0, \theta > 0\),

\[
\mathbb{P}\{V(ax) > x\} \gtrsim \mathbb{P}\{V^{1-\rho+\delta}_{\leq x}(ax) > (1 + \theta)x\}.
\]

**Proof**

Since \(\rho_{\leq z} \uparrow \rho\) for \(z \to \infty\), there exists an \(x_0\) such that \(\rho_{\leq x_0} > \rho - \delta\) for all \(x \geq x_0\).

From Proposition 4.1, taking \(c = \rho - \delta, y = \theta x, z = \epsilon x\), for all \(x \geq x_0\),

\[
\frac{\mathbb{P}\{V(ax) > x\}}{\mathbb{P}\{V^{1-\rho+\delta}_{\leq x}(ax) > (1 + \theta)x\}} \geq \frac{\mathbb{P}\{Z^{\rho-\delta}_{\leq x}(ax) \leq \theta x\}}{\mathbb{P}\{Z^{\epsilon-\delta}_{\leq \epsilon x_0}(ax) \leq \theta x\}}.
\]

The statement then easily follows.

We now proceed with a simple sample-path upper bound.

**Proposition 4.3** For any \(c \in (\rho_{\leq z}, 1 - \rho_{>z})\),

\[
\mathbb{P}\{V(t) > x\} \leq \mathbb{P}\{V^{1-c}_{>z}(t) > x - y\} + \mathbb{P}\{V^c_{\leq z}(t) > y\}.
\]

**Proof**

Sample-path wise,

\[
V(t) = \sup_{0 \leq s \leq t} \{A(0, s) - s\} \\
= \sup_{0 \leq s \leq t} \{A_{>z}(0, s) - (1 - c)s + A_{\leq z}(0, s) - cs\} \\
\leq \sup_{0 \leq s \leq t} \{A_{>z}(0, s) - (1 - c)s\} + \sup_{0 \leq s \leq t} \{A_{\leq z}(0, s) - cs\} \\
= V^{1-c}_{>z}(t) + V^c_{\leq z}(t).
\]
The next proposition provides an upper bound which indicates that the workload from the short sessions can be asymptotically neglected.

**Proposition 4.4** For any $c > \rho$, $\theta > 0$, $\mu > 0$, there exists an $\epsilon^* > 0$ such that for all $\epsilon < \epsilon^*$,

$$\mathbb{P}\{V_{\leq \epsilon x}^c(ax) > \theta x\} = o(x^{-\mu})$$

as $x \to \infty$.

**Proof**

Define $\delta := (c - \rho)/K$. Then

$$V_{\leq \epsilon x}^c(ax) = \sup_{0 \leq s \leq ax} \{A_{\leq \epsilon x}(0, s) - cs\}$$

$$= \sup_{0 \leq s \leq ax} \left\{ \sum_{k=1}^{K} A_{k,\leq \epsilon x}(0, s) - \sum_{k=1}^{K} (\rho_k + \delta)s \right\}$$

$$\leq \sum_{k=1}^{K} \sup_{0 \leq s \leq ax} \{A_{k,\leq \epsilon x}(0, s) - (\rho_k + \delta)s\}$$

$$= \sum_{k=1}^{K} V_{k,\leq \epsilon x}^{\rho_k + \delta}(ax).$$

This implies

$$\mathbb{P}\{V_{\leq \epsilon x}^c(ax) > x\} \leq \sum_{k=1}^{K} \mathbb{P}\{V_{k,\leq \epsilon x}^{\rho_k + \delta}(ax) > x/K\}. $$

Thus, it suffices to show that

$$\mathbb{P}\{V_{k,\leq \epsilon x}^{\rho_k + \delta}(ax) > x/K\} = o(x^{-\mu})$$

as $x \to \infty$ for all $k = 1, \ldots, K$.

Now observe that

$$V_{k,\leq \epsilon x}^{\rho_k + \delta}(ax) \leq A_{k,\leq \epsilon x}^{(0)} + \sup_{0 \leq s \leq ax} \{A_{k,\leq \epsilon x}^{(>0)}(0, s) - (\rho_k + \delta)s\},$$

where the two terms correspond to the traffic generated by the sessions already active at and starting after time 0, respectively. Hence,

$$\mathbb{P}\{V_{k,\leq \epsilon x}^{\rho_k + \delta}(ax) > x/K\} \leq \mathbb{P}\{A_{k,\leq \epsilon x}^{(0)} > x/(2K)\}$$

$$+ \mathbb{P}\{ \sup_{0 \leq s \leq ax} \{A_{k,\leq \epsilon x}^{(>0)}(0, s) - (\rho_k + \delta)s\} > x/(2K)\}$$

$$= I + II.$$

In the remainder of the proof, we bound the terms $I$ and $II$. 

13
We first consider Term I. Note that $A_{k, \leq \varepsilon x}^{(0)}$ is stochastically smaller than $r_k \sum_{i=1}^{N_k(0)} B_{k,i}(\varepsilon x)$, where $B_{k,i}(\varepsilon x) \overset{d}{=} B_{k,i} | B_{k,i} \leq \varepsilon x$. Thus,

$$I \leq \mathbb{P}\{r_k \sum_{i=1}^{N_k(0)} B_{k,i}(\varepsilon x) > x/(2K)\} + \mathbb{P}\{N_k(0) > \sqrt{x}\}.$$  

Since $N_k(0)$ is Poisson distributed, the second term decays exponentially fast in $x$. Using Lemma 2.1, the first term can be bounded as follows, for $\varepsilon$ sufficiently small:

$$\mathbb{P}\{r_k \sum_{i=1}^{\sqrt{x}} B_{k,i}(\varepsilon x) > x/(2K)\} \leq \mathbb{P}\{\sum_{i=1}^{\sqrt{x}} [B_{k,i}(\varepsilon x) - 2\mathbb{E}[B_k]] > x/(2Kr_K) - 2\mathbb{E}[B_k] \sqrt{x}\} \leq \phi(x/(2Kr_K) - 2\mathbb{E}[B_k] \sqrt{x}),$$

with $\phi(\cdot) \in \mathbb{R}_{-\alpha}$, $\alpha > \mu + \frac{1}{2}$.

We now turn to Term II. Note that $\sup_{0 \leq s \leq ax} \{A_{k, \leq \varepsilon x}^{(0)}(0, s) - (\rho_k + \delta) s\}$ is stochastically smaller than $W_{\rho_k + \delta, k, \leq \varepsilon x}(ax)$, where the latter quantity represents the workload if the entire amount of traffic generated over the duration of a session were released instantaneously upon the arrival of the session. Thus,

$$II \leq \mathbb{P}\{W_{\rho_k + \delta, k, \leq \varepsilon x}(ax) > x/(2K)\}.$$  

Now observe that $W_{\rho_k + \delta, k, \leq \varepsilon x}(ax)$ is the workload at time $ax$ in an $M/G/1$ queue of capacity $\rho_k + \delta$ with arrival rate $\lambda_k B_k(\varepsilon x)$ and service time distribution $B_k(y/r_k)/B_k(\varepsilon x)$, $0 \leq y \leq \varepsilon rx_k$. Let $B_{k,n}(\varepsilon x)$, $n \geq 1$ be an i.i.d. sequence of random variables with this distribution, and let $U_{k,n}$, $n \geq 1$, be an i.i.d. sequence of interarrival times. Denote by $N_k(ax) := \sup\{n : U_{k,1} + \ldots + U_{k,n} \leq ax\}$ the number of arrivals in this $M/G/1$ queue up to time $ax$. Define $S_{k,n}(\varepsilon x) := \sum_{i=1}^{n} X_{k,i}$, with $X_{k,i} := B_{k,i}(\varepsilon x) - (\rho_k + \delta) U_{k,i}$. Then, for any $\Lambda$,

$$\mathbb{P}\{W_{\rho_k + \delta, k, \leq \varepsilon x}(ax) > x/(2K)\} = \mathbb{P}\left\{\sup_{n \leq N_k(ax)} S_{k,n}(\varepsilon x) > x/(2K)\right\} \leq \mathbb{P}\left\{\sup_{n \leq \Lambda ax} S_{k,n}(\varepsilon x) > x/(2K)\right\} + \mathbb{P}\{N_k(ax) > \Lambda ax\}.$$  

The second term decays exponentially fast in $x$ for $\Lambda > \lambda_k$. Using Lemma 2.1, noting that $\mathbb{E}[X_1] < 0$, the first term can be bounded by, for $\varepsilon^* > 0$ sufficiently small,

$$\sum_{n=1}^{\Lambda ax} \mathbb{P}\{S_{k,n}(\varepsilon x) > x/(2K)\} \leq \Lambda ax \phi(x/(2K)),$$

with $\phi(\cdot) \in \mathbb{R}_{-\alpha}$, $\alpha > \mu + 1$. This completes the proof. 

$\square$
We now combine the above two bounds to obtain an asymptotic upper bound for \( \mathbb{P}\{V(ax) > x\} \) as \( x \to \infty \).

**Proposition 4.5** For any \( \delta > 0, \theta > 0, \mu > 0 \), there exists an \( \epsilon^* > 0 \) such that for all \( \epsilon < \epsilon^* \),

\[
\mathbb{P}\{V(ax) > x\} \leq \mathbb{P}\{V_{\geq \epsilon x}^{1-\rho-\delta}(ax) > (1 - \theta)x\} + o(x^{-\mu})
\]

as \( x \to \infty \).

**Proof**
The proof follows directly from Propositions 4.3 and 4.4 taking \( c = \rho + \delta \).

Combined, Propositions 4.2 and 4.5 allow us to restrict the attention to long sessions only, and focus on probabilities of the form \( \mathbb{P}\{V_{\geq \epsilon x}^{1-\rho+\delta}(ax) > (1 \pm \theta)x\} \).

### 4.2 Configuration of long sessions

In this subsection, we determine the typical combination of long sessions involved in causing overflow. Specifically, we show that, for overflow of level \( x \) to occur within time \( ax \), the configuration of long sessions in the interval \([0, ax]\) must be in the set \( S^*_a \). As we argued before, these configurations of long sessions may be interpreted as the most likely ones to occur among those producing sufficiently high drift. All other combinations are unlikely to cause overflow, either because the resulting drift is simply too low, or because the corresponding probability is too small (or both).

In order to formalize these statements, we need to keep track of the number of long sessions in the time interval \([0, ax]\). With minor abuse of notation, define \( N_{k,>\epsilon x}(T) \) as the number of class-\( k \) sessions exceeding length \( \epsilon x \) in the time interval \( T \). Denote \( N_{>\epsilon x}(T) := (N_{1,>\epsilon x}(T), \ldots, N_{K,>\epsilon x}(T)) \). Formally, we will show that for \( \delta, \theta \) sufficiently small,

\[
\mathbb{P}\{V_{>\epsilon x}^{1-\rho+\delta}(ax) > (1 \pm \theta)x\} \sim \sum_{n \in S^*_a} \mathbb{P}\{V_{>\epsilon x}^{1-\rho+\delta}(ax) > (1 \pm \theta)x; N_{>\epsilon x}([0, ax]) = n\}.
\]

We first exclude the possibility that overflow is caused by some configuration which fails to generate at least a drift \( 1/a \).

Let \( S^*_a(c) \) be the set of optimal solutions of the integer linear program formulated at the beginning of this section with the constraint value \( 1 - \rho + 1/a \) replaced by \( c + 1/a \). Denote by \( \mu^*_a(c) \) the corresponding optimal value. Define \( S^-_a(c) := \{n \in \mathbb{N}^K : c + 1/a \geq \sum_{k=1}^K n_k r_k < \sum_{k=1}^K n_k r_k < c + 1/a\} \), \( S^+_a(c) := \{n \in \mathbb{N}^K : \sum_{k=1}^K n_k r_k > c + 1/a\} \), and \( r^\max_a(c) := \max_{n \in S^-_a(c)} \sum_{k=1}^K n_k r_k \).

**Proposition 4.6** For \( \theta \) sufficiently small, and all \( \epsilon > 0, x > 0 \),

\[
\mathbb{P}\{V_{>\epsilon x}^c(ax) > (1 \pm \theta)x; N_{>\epsilon x}([0, ax]) \in S^-_a(c)\} = 0.
\]
Proof
The idea of the proof is as follows. If \( N_{> \varepsilon x}([0, ax]) \in S_a(c) \), then during the time interval \([0, ax]\) the drift of the workload is always less than \(1/a\). Hence, the workload cannot reach level \((1 \pm \theta)x\) before time \(ax\) for \(\theta\) sufficiently small.

Formally, denote \( v_a(c) := c + 1/a - r_a^{\max}(c) > 0 \). If \( N_{> \varepsilon x}([0, ax]) \in S_a(c) \), then the left derivative \( \frac{d}{dx} A_{> \varepsilon x}(0, s) \leq r_a^{\max}(c) \) for all \( s \in [0, ax] \), so that \( A_{> \varepsilon x}(0, s) \leq r_a^{\max}(c)s \) for all \( s \in [0, ax] \). Therefore,

\[
V_{> \varepsilon x}(ax) = \sup_{0 \leq s \leq ax} \{ A_{> \varepsilon x}(0, s) - cs \} \leq \sup_{0 \leq s \leq ax} \{ (r_a^{\max}(c) - c)s \} = \\
\sup_{0 \leq s \leq ax} \{ (1/a - u_a(c))s \} = (1/a - u_a(c))ax.
\]

The latter quantity is less than \((1 \pm \theta)x\) for \(\theta < au_a(c)\). \(\square\)

We now eliminate all configurations of long sessions that do generate at least a drift \(1/a\), but that are relatively unlikely compared to other combinations that do so.

**Proposition 4.7** There exists \( \mu > \mu_a^*(c) \) such that, for all \( \varepsilon > 0 \), \( n \in S_a(c) \setminus S_a^e(c) \),

\[
P\{ N_{> \varepsilon x}([0, ax]) \geq n \} = o(x^{-\mu})
\]

as \( x \to \infty \).

**Proof**
Note that \( N_{k,> \varepsilon x}([0, ax]) \) has a Poisson distribution with parameter \( \bar{\rho}_k \mathbb{P}\{ B_k > \varepsilon x \} + \lambda_k ax \mathbb{P}\{ B_k > \varepsilon x \} \). A straightforward computation then shows that \( \mathbb{P}\{ N_{k,> \varepsilon x}([0, ax]) \geq n_k \} \) is upper bounded by a function which is regularly varying of index \(-n_k(\nu_k - 1)\). Since

\[
\mathbb{P}\{ N_{> \varepsilon x}([0, ax]) \geq n \} = \prod_{k=1}^{K} \mathbb{P}\{ N_{k,> \varepsilon x}([0, ax]) \geq n_k \},
\]

the left-hand side is upper bounded by a function which is regularly varying of index \(-\sum_{k=1}^{K} n_k(\nu_k - 1)\). The fact that \( n \in S_a^+(c) \setminus S_a^e(c) \) implies \( \sum_{k=1}^{K} n_k(\nu_k - 1) > \mu_a^*(c) \), because otherwise \( n \in S_a^e(c) \). \(\square\)

Combined, the above two propositions allow us to limit the attention to scenarios with \( N_{> \varepsilon x}([0, ax]) \in S_a^e(c) \), as formalized in the following lemma.
Lemma 4.1 Assume that \( r_{a}^{\min} > 1 - \rho \). Then there exists a \( \mu > \mu_{a}^{\rho} \) such that for \( \delta, \theta \) sufficiently small, and all \( \epsilon > 0 \),

\[
\mathbb{P}\{ V_{>\varepsilon \pi}^{1-\rho+\delta} (ax) > (1 \pm \theta)x \} = \sum_{n \in \mathbb{S}^{*}_{\varepsilon}} \mathbb{P}\{ V_{>\varepsilon \pi}^{1-\rho+\delta} (ax) > (1 \pm \theta)x; N_{>\varepsilon \pi}\{\{0, ax\}\} = n \} + o(x^{-\mu}).
\]

Proof

The proof follows directly from Propositions 4.6, 4.7, noting that \( S_{a}(1 - \rho \pm \delta) = S_{a}^{*} \) for \( \delta \) sufficiently small as \( r_{a}^{\min} > 1 - \rho \).

\qed

Combined with the earlier results, we have now obtained asymptotic lower and upper bounds for \( \mathbb{P}\{ V > x \} \) in terms of the probabilities \( \mathbb{P}\{ V_{>\varepsilon \pi}^{1-\rho+\delta} (ax) > (1 \pm \theta)x; N_{>\varepsilon \pi}\{\{0, ax\}\} = n \} \).

What thus remains is to determine the asymptotic behavior of these probabilities as \( x \to \infty \), which is the subject of the next subsection.

4.3 Computing \( \mathbb{P}\{ V_{>\varepsilon \pi}^{1-\rho+\delta} (ax) > (1 \pm \theta)x; N_{>\varepsilon \pi}\{\{0, ax\}\} = n \} \)

In this subsection we identify a stopping time \( \tau_{f}^{n}(\varepsilon x) \) (conditional upon the event \( N_{>\varepsilon \pi}\{\{0, ax\}\} = n \)) such that for \( a \) sufficiently large and \( c \) sufficiently close to \( 1 - \rho \),

\[
\mathbb{P}\{ \sup_{0 \leq s \leq \varepsilon x} \{ A_{>\varepsilon \pi}(0, s) - cs \} > x \} \sim \mathbb{P}\{ A_{>\varepsilon \pi}(0, \tau_{f}^{n}(\varepsilon x)) - c\tau_{f}^{n}(\varepsilon x) > x \}.
\]

We first introduce some additional notation. Assume that \( N_{>\varepsilon \pi}(0) \leq n \). In this case, we define \( A_{>\varepsilon \pi}^{n}(0, t) \) as the amount of traffic generated up to time \( t \) by the first \( n_{k} \) class-\( k \) sessions only, \( k = 1, \ldots, K \). Define \( V_{>\varepsilon \pi}^{c, n}(t) := \sup_{0 \leq s \leq t} \{ A_{>\varepsilon \pi}^{n}(0, s) - cs \} \).

Let \( \tau_{s,k}^{n}(\varepsilon x) \) and \( \tau_{f,k}^{n}(\varepsilon x) \) be the respective starting and finishing times of the \( n \)-th class-\( k \) session exceeding length \( \varepsilon x \). For any \( n \in \mathbb{N}^{K} \), let

\[
\tau_{s}^{n}(\varepsilon x) := \max_{k=1,\ldots,K} \tau_{s,k}^{n}(\varepsilon x),
\]

and

\[
\tau_{f}^{n}(\varepsilon x) := \min_{k=1,\ldots,K} \tau_{f,k}^{n}(\varepsilon x).
\]

Thus, for a configuration \( n \in \mathbb{N}^{K} \) of long sessions, \( \tau_{s}^{n}(\varepsilon x) \) is the time at which the last long session begins, and \( \tau_{f}^{n}(\varepsilon x) \) is the time at which the first long session ends. To account for the case \( \tau_{f}^{n}(\varepsilon x) > ax \), we define \( \tau_{f}^{n}(\varepsilon x) := \min\{ ax, \tau_{f}^{n}(\varepsilon x) \} \). This turns out to be the relevant stopping time, as is demonstrated by the following lemma.

Lemma 4.2 There exists a \( \mu > \mu_{a}^{\rho}(c) \) such that for \( \theta \) sufficiently small and all \( n \in S_{a}^{*}(c) \),

\[
\mathbb{P}\{ V_{>\varepsilon \pi}^{c}(ax) > (1 \pm \theta)x; N_{>\varepsilon \pi}\{\{0, ax\}\} = n \} \gtrsim \mathbb{P}\{ N_{>\varepsilon \pi}(0) \leq n; A_{>\varepsilon \pi}^{n}(0, \tau_{f}^{n}(\varepsilon x)) - c\tau_{f}^{n}(\varepsilon x) > (1 - \theta)x \} + o(x^{-\mu}).
\]
In case $r_{\alpha}^{\max}(c) < c$, there also exists a $\mu > \mu_{\alpha}^*(-c)$ such that for $\theta$ sufficiently small and all $n \in S_{\alpha}(c)$,

$$\mathbb{P}\{V_{>\varepsilon x}^{\leq \varepsilon x}(ax) > (1 \pm \theta)x; N_{>\varepsilon x}([0, ax]) = n\} \approx \mathbb{P}\{N_{>\varepsilon x}(0) \leq n; A_{>\varepsilon x}^n(0, \tau_{\varepsilon x}(\varepsilon x)) - c\tau_{\varepsilon x}^n(\varepsilon x) > (1 \pm \theta)x + o(x^{-\mu})\}.$$

**Proof**

We first prove the second statement. Since $V_{>\varepsilon x}^{c,n}(ax) \leq V_{>\varepsilon x}^{c,n}(ax)$, with strict equality under the event $N_{>\varepsilon x}([0, ax]) = n$, and the latter event also implies that $N_{>\varepsilon x}(0) \leq n$, we have

$$\mathbb{P}\{V_{>\varepsilon x}^{c,n}(ax) > (1 \pm \theta)x; N_{>\varepsilon x}([0, ax]) = n\} = \mathbb{P}\{V_{>\varepsilon x}^{c,n}(ax) > (1 \pm \theta)x; N_{>\varepsilon x}([0, ax]) = n; N_{>\varepsilon x}(0) \leq n\}.$$

First observe that

$$\mathbb{P}\{V_{>\varepsilon x}^{c,n}(ax) > (1 \pm \theta)x; N_{>\varepsilon x}([0, ax]) = n; N_{>\varepsilon x}(0) \leq n\} \leq \mathbb{P}\{V_{>\varepsilon x}^{c,n}(ax) > (1 \pm \theta)x; N_{>\varepsilon x}(0) \leq n\} = \mathbb{P}\{\sup_{0 \leq s \leq ax} \{A_{>\varepsilon x}^n(0, s) - cs\} > (1 \pm \theta)x; N_{>\varepsilon x}(0) \leq n\}.$$

Note that before time $\tau_{\varepsilon x}(\varepsilon x)$ and after time $\tau_{\varepsilon x}(\varepsilon x)$ the drift of the process $A_{>\varepsilon x}^n(0, s)$ is at most $r_{\alpha}^{\max}(c) < c$. Thus, the drift of the process $\{A_{>\varepsilon x}^n(0, s) - cs\}$ is only positive between times $\tau_{\varepsilon x}(\varepsilon x)$ and $\tau_{\varepsilon x}(\varepsilon x)$. Hence, $\sup_{0 \leq s \leq ax} \{A_{>\varepsilon x}^n(0, s) - cs\} > (1 \pm \theta)x$ implies that $A_{>\varepsilon x}^n(0, \tau_{\varepsilon x}(\varepsilon x)) - c\tau_{\varepsilon x}^n(\varepsilon x) > (1 \pm \theta)x$. Thus, the last probability in the above display is smaller than $\mathbb{P}\{N_{>\varepsilon x}(0) \leq n; A_{>\varepsilon x}^n(0, \tau_{\varepsilon x}(\varepsilon x)) - c\tau_{\varepsilon x}^n(\varepsilon x) > (1 \pm \theta)x\}$.

We now turn to the first statement. Observe that $V_{>\varepsilon x}^{c,n}(ax) = 0$, unless $N_{>\varepsilon x}([0, ax]) \geq n$, so for $\theta$ sufficiently small, using Proposition 4.7,

$$\mathbb{P}\{V_{>\varepsilon x}^{c,n}(ax) > (1 \pm \theta)x; N_{>\varepsilon x}([0, ax]) = n; N_{>\varepsilon x}(0) \leq n\} \geq \mathbb{P}\{V_{>\varepsilon x}^{c,n}(ax) > (1 \pm \theta)x; N_{>\varepsilon x}(0) \leq n\} - \mathbb{P}\{N_{>\varepsilon x}(0, ax) > n\}$$

$$\geq \mathbb{P}\{V_{>\varepsilon x}^{c,n}(ax) > (1 \pm \theta)x; N_{>\varepsilon x}(0) \leq n\} + o(x^{-\mu})$$

$$= \mathbb{P}\{\sup_{0 \leq s \leq ax} \{A_{>\varepsilon x}^n(0, s) - cs\} > (1 \pm \theta)x; N_{>\varepsilon x}(0) \leq n\} + o(x^{-\mu})$$

$$\geq \mathbb{P}\{N_{>\varepsilon x}(0) \leq n; A_{>\varepsilon x}^n(0, \tau_{\varepsilon x}(\varepsilon x)) - c\tau_{\varepsilon x}^n(\varepsilon x) > (1 \pm \theta)x\} + o(x^{-\mu}).$$

Combined with the earlier results, we have now obtained asymptotic lower and upper bounds for $\mathbb{P}\{V > x\}$ in terms of the probabilities $\mathbb{P}\{N_{>\varepsilon x}(0) \leq n; A_{>\varepsilon x}^n(0, \tau_{\varepsilon x}(\varepsilon x)) - c\tau_{\varepsilon x}^n(\varepsilon x) > (1 \pm \theta)x\}$ with $c = 1 - \rho \pm \delta$. What thus remains is to determine the asymptotic behavior of these probabilities as $x \to \infty$, which is the subject of the next subsection.
4.4 Computing \( \mathbb{P}\{N_{>\varepsilon x}(0) \leq n; A_{>\varepsilon x}(0, \tau_f^\varepsilon(\varepsilon x)) - c\tilde{\tau}_f^\varepsilon(\varepsilon x) > (1 \pm \theta)x\} \)

As a final step, we compute the asymptotic behavior of \( \mathbb{P}\{N_{>\varepsilon x}(0) \leq n; A_{>\varepsilon x}(0, \tau_f^\varepsilon(\varepsilon x)) - c\tilde{\tau}_f^\varepsilon(\varepsilon x) > (1 \pm \theta)x\} \) for fixed \( n \in S^*_a(c) \) and \( x \to \infty \). Throughout this section, we assume that \( a \) is large enough for the condition \( r^\varepsilon a(x) < c \) to hold.

We start by conditioning upon the configuration of long sessions active at time 0. For \( j = (j_1, \ldots, j_K) \), define the event \( D_j(\varepsilon x) \) by \( D_j(\varepsilon x) := \{N_{>\varepsilon x}(0) = j\} \). In words, \( D_j(\varepsilon x) \) is the event that the number of long class-\( k \) sessions active at time 0 is \( j_k \), \( k = 1, \ldots, K \).

Denote \( \mathbb{P}_j \{\cdot\} = \mathbb{P}\{\cdot\mid D_j(\varepsilon x)\} \). Then

\[
\mathbb{P}\{N_{>\varepsilon x}(0) \leq n; A_{>\varepsilon x}(0, \tau_f^\varepsilon(\varepsilon x)) - c\tilde{\tau}_f^\varepsilon(\varepsilon x) > (1 \pm \theta)x\} = \sum_{j \leq n} \mathbb{P}\{D_j(\varepsilon x)\} \mathbb{P}_j \{A_{>\varepsilon x}(0, \tau_f^\varepsilon(\varepsilon x)) - c\tilde{\tau}_f^\varepsilon(\varepsilon x) > (1 \pm \theta)x\}.
\]

Denote \( J: \mathbb{P}\{B_k > \varepsilon x\}\) by \( \sum_{k=1}^K \sum_{i=1}^{n_k-j_k} \mathbb{P}\{B_k > \varepsilon x\} \).

It remains to compute the asymptotic behavior of \( \mathbb{P}_j \{A_{>\varepsilon x}(0, \tau_f^\varepsilon(\varepsilon x)) - c\tilde{\tau}_f^\varepsilon(\varepsilon x) > (1 \pm \theta)x\} \) as \( x \to \infty \). In order to so, we need to condition upon the arrival times of the remaining sessions as well. Denote the interarrival times of the class-\( k \) sessions by \( E_{ki}(\varepsilon x), k = 1, \ldots, K, i = 1, 2, \ldots \). Note that \( E_{ki}(\varepsilon x) \) is an exponentially distributed random variable with parameter \( \lambda_k \mathbb{P}\{B_k > \varepsilon x\} \).

To obtain an expression for \( A_{>\varepsilon x}(0, \tau_f^\varepsilon(\varepsilon x)) - c\tilde{\tau}_f^\varepsilon(\varepsilon x) \) under the event \( D_j(\varepsilon x) \), note that, if all long sessions had been active already at time 0, the expression would equal \( cn\tau_f^\varepsilon(\varepsilon x) \), with \( c_n := \sum_{k=1}^K n_k r_k - c \). However, some sessions may have started later. To account for this, it is not hard to see that we need to subtract \( H(\varepsilon x) \), which is defined by

\[
H(\varepsilon x) := \sum_{k=1}^K r_k \sum_{i=1}^{n_k-j_k} \sum_{l=1}^i E_{ki}(\varepsilon x).
\]

This is summarized in the following lemma.

**Lemma 4.3** Under the event \( D_j(\varepsilon x) \), \( A_{>\varepsilon x}(0, \tau_f^\varepsilon(\varepsilon x)) - c\tilde{\tau}_f^\varepsilon(\varepsilon x) \) can be represented as

\[
A_{>\varepsilon x}(0, \tau_f^\varepsilon(\varepsilon x)) - c\tilde{\tau}_f^\varepsilon(\varepsilon x) = c_n\tau_f^\varepsilon(\varepsilon x) - H(\varepsilon x),
\]

with \( \tau_f^\varepsilon(\varepsilon x) = \min\{ax, \tau_f^\varepsilon(\varepsilon x)\} \) the stopping time defined earlier and

\[
\tau_f^\varepsilon(\varepsilon x) = \min_{k=1, \ldots, K} \min_{i=1, \ldots, n_k-j_k} \min_{l=1, \ldots, j_k} \sum_{i=1}^{n_k-j_k} E_{ki}(\varepsilon x) + \ldots + E_{ki}(\varepsilon x) + \tilde{B}_{ki}(\varepsilon x)] \}
\]

Here \( \tilde{B}_{ki}(\varepsilon x) \) \( \tilde{B}_{ki}(\varepsilon x) \) is an exponentially distributed random variable with parameter \( \lambda_k \mathbb{P}\{B_k > \varepsilon x\} \), and \( \tilde{B}_{ki}(\varepsilon x) \) is an exponentially distributed random variable with parameter \( \lambda_k \mathbb{P}\{B_k > \varepsilon x\} \).
We proceed to compute the asymptotic behavior of \( \mathbb{P}_j \{ A_{\geq \epsilon x}^n(0, \tau^n_j(\epsilon x)) - c \tau^n_j(\epsilon x) > (1 \pm \theta)x \} \), using the above representation.

Define the sets \( E_0 := \{(k, i) : k = 1, \ldots, K, i = 1, \ldots, j_k \} \) and \( E_1 := \{(k, i) : k = 1, \ldots, K, i = 1, \ldots, n_k - j_k \} \). Write \( y = y_{(k, i)} \in E_0 \) (we interpret \( y \) as a vector), and let \( h(y) \) be a realization of \( H(\epsilon x) \), i.e., if \( E_{ki}(\epsilon x) = y_{ki} \) for \( (k, i) \in E_1 \), then

\[
h(y) = \sum_{k=1}^{K} \sum_{i=1}^{n_k-j_k} \sum_{l=1}^{j_k} y_{kl}.
\]

Let \( t(y) \) be distributed as \( \tau^n_j(\epsilon x) \) conditional upon \( E_{ki}(\epsilon x) = y_{ki} \) for \( (k, i) \in E_1 \). Note that \( t(y) \) is still a random variable. Hence, using Lemma 4.3,

\[
P_j \{ A_{\geq \epsilon x}^n(0, \tau^n_j(\epsilon x)) - c \tau^n_j(\epsilon x) > (1 \pm \theta)x \} \approx \frac{1}{K} \prod_{k=1}^{K} \mathbb{P}\{ B_k^r > \epsilon x \} \int_{y \geq 0, h(y) \leq (\epsilon a - 1)x} \left[ \prod_{(k, i) \in E_1} \left( \frac{\lambda_k e^{-y_{ki} \lambda_k} \mathbb{P}\{ B_{ki} > \epsilon x \}}{\prod_{(k, i) \in E_0} \mathbb{P}\{ c_{n} (y_{ki} + \ldots + y_{ki} + B_k) > (1 \pm \theta)x + h(y) \}} \right) \right] dy.
\]

This implies (using bounded convergence)

\[
P_j \{ D_j(\epsilon x) \} \approx \frac{1}{K} \prod_{k=1}^{K} \mathbb{P}\{ B_k^r > \epsilon x \} \int_{y \geq 0, h(y) \leq (\epsilon a - 1)x} \left[ \prod_{(k, i) \in E_1} \mathbb{P}\{ c_{n} (y_{ki} + \ldots + y_{ki} + B_k) > (1 \pm \theta)x + h(y) \} \right] dy.
\]

For given \( n \) and \( j \), define the \(|E_1|\)-dimensional row vector \( g = g^c, y, n \) by \( g = (g_1, \ldots, g_K) \). Here \( g_k \) is a row vector of dimension \( n_k - j_k \) with all elements equal to \( r_k/c_n \). In the sequel, we write \( g := (g_{(k, i)})_{(k, i) \in E_1} \). Let \( G \) be a square matrix with all rows equal to \( g \). Define \( \hat{G} := G - I \). Note that \( |\hat{G}| = e g - 1 \) and that the inverse \( H \) of \( \hat{G} \) is given by \( H = \frac{1}{eg - 1} G - I \). Here \( e := (1, \ldots, 1) \) is the unit vector with all elements equal to 1. Note that \( gH = \frac{1}{eg - 1} g \).

Set \( z := (z_{ki})_{(k, i) \in E_1} \), where \( z_{ki} = y_{ki} + \ldots + y_{ki} \). Define \( w := \hat{G} z \). Note that \( h(y) = c_n g z \).
Straightforward computations yield

\[
\prod_{k=1}^{K} \frac{1}{\tilde{p}_k^{\sum n_k - j_k}} \int_{y \geq 0, h(y) \leq (c_n a - 1) x} \left[ \prod_{(k,i) \in E_0} \mathbb{P}(c_n B_k > (1 \pm \theta) x + h(y)) \right] dy
\]

\[
= \prod_{k=1}^{K} \frac{1}{\tilde{p}_k^{\sum n_k - j_k}} \int_{z \geq 0, gz \leq (a - 1/c_n) x} \left[ \prod_{(k,i) \in E_0} \mathbb{P}(B_{ki} > (1 \pm \theta) \frac{x}{c_n} + gz) \right] dz
\]

\[
= \prod_{k=1}^{K} \frac{1}{\tilde{p}_k^{\sum n_k - j_k}} \frac{1}{eg - 1} \int_{w \geq 0, gw \leq (eg - 1)(a - 1/c_n) x} \left[ \prod_{(k,i) \in E_0} \mathbb{P}(B_{ki} > (1 \pm \theta) \frac{x}{c_n} + \frac{1}{eg - 1} gw) \right] dw
\]

\[
= \frac{1}{eg - 1} \int_{w \geq 0, gw \leq (eg - 1)(a - 1/c_n) x} \left[ \prod_{(k,i) \in E_0} \mathbb{P}(B_{ki} > (1 \pm \theta) \frac{x}{c_n} + \frac{1}{eg - 1} gw) \right] dw
\]

In the last expression, \( B_{E_1} := (B_{ki})_{(k,i) \in E_1} \).

Using the fact that \( B_{ki} \) is regularly varying of index \( 1 - \nu_k \), it is easy to show that

\[
\mathbb{P}(B_{ki} > (1 \pm \theta) \frac{x}{c_n}) = (1 + c_n y)^{1 - \nu_k}.
\]

Take the \( Z_{ki} \) independent. Then, with obvious notation,

\[
(eg - 1) P_{j,n,a}^{c}(x)
\]

\[
= \mathbb{P}(B_{ki} > \frac{x}{c_n}, k = 1, \ldots, K, i = 1, \ldots, n_k; B_{ki} - \frac{x}{c_n} \geq \frac{1}{eg - 1} g \left( B_{E_1} - (1 \pm \theta) \frac{x}{c_n} e \right), (k,i) \in E_0;
\]
\[
\frac{1}{e^g-1} g \left( B_{E_1} - \frac{x}{c_n} \right) \leq x \left( a - \frac{1}{c_n} \right)
\]

\[
\sim \mathbb{P}\{Z_{ki} \geq \frac{1}{e^g-1} g Z_{E_1}, (k, i) \in E_0; (a - \frac{1}{c_n}) \geq \frac{1}{e^g-1} g Z_{E_1}\} \prod_{k=1}^{K} \mathbb{P}\{B_k^c > \frac{x}{c_n}\}^{j_k}.
\]

The above calculations are summarized in the following lemma.

Lemma 4.4 For \( n \in S^*_a(c) \) there exists an \( \epsilon^* > 0 \) such that for all \( \epsilon < \epsilon^* \),

\[
\mathbb{P}\{N_{>ex}(0) \leq n; A_{>ex}(0, \tau^p_f(\epsilon x)) - c\tau^p_f(\epsilon x) > (1 \pm \theta)x\} \sim \sum_{j \leq n} \prod_{k=1}^{K} \frac{\alpha_k}{\beta_k} P_{j,n,a}^c((1 \pm \theta)x),
\]

where

\[
P_{j,n,a}^c(x) = \frac{1}{e^g-1} \mathbb{P}\{B_{ki}^c > \frac{x}{c_n}, k = 1, \ldots, K, i = 1, \ldots, n_k; \}
\]

\[
\left( B_{ki}^c - \frac{x}{c_n} \right) \geq \frac{1}{e^g-1} g \left( B_{E_1}^c - \frac{x}{c_n} \right), (k, i) \in E_0;
\]

\[
\frac{1}{e^g-1} g \left( B_{E_1}^c - \frac{x}{c_n} \right) \leq x \left( a - \frac{1}{c_n} \right),
\]

with \( g = g^{c,j,n} \) as defined earlier.

In particular, we have

\[
P_{j,n,a}^c(x) \sim \kappa_{j,n,a}^c \prod_{k=1}^{K} \mathbb{P}\{B_k^c > \frac{x}{c_n}\}^{n_k},
\]

with \( \kappa_{j,n,a}^c = 1 \), and for \( j \leq n, j \neq n \),

\[
\kappa_{j,n,a}^c = \frac{1}{e^g-1} \mathbb{P}\{Z_{ki} \geq \frac{1}{e^g-1} g Z_{E_1}; (k, i) \in E_0; (a - \frac{1}{c_n}) \geq \frac{1}{e^g-1} g Z_{E_1}\}.
\]

The coefficient \( \kappa_{j,n,a}^c \) is a continuous function of \( c \) in a neighborhood of \( c = 1 - \rho \).

The continuity property of the coefficient \( \kappa_{j,n,a}^c \) follows immediately from its definition.

4.5 Proof of Theorem 3.1

We have now gathered all the ingredients for the proof of Theorem 3.1, which is restated below in extended form. Recall that \( d_n = \sum_{k=1}^{K} n_k r_k + \rho - 1 \).

Theorem 3.1

Assume that \( r^{\min} > 1 - \rho \). Then,

\[
\mathbb{P}\{V > x\} \sim \sum_{n \in S^*} \sum_{j \leq n} \prod_{k=1}^{K} \frac{\beta_k^{n_k}}{\alpha_k^{j_k}} P_{j,n}(x),
\]

22
where \( j = (j_1, \ldots, j_K) \) and \( P_{j,n}(x) := \lim_{a \to \infty} P_{j,n,a}^{1-\rho}(x) \) satisfies

\[
P_{j,n}(x) \sim \kappa_{j,n} \prod_{k=1}^{K} \mathbb{P}\{B_k^* > \frac{x}{d_n}\}^{n_k},
\]

for some constant \( \kappa_{j,n} := \lim_{a \to \infty} \kappa_{j,n,a} \), with \( \kappa_{j,n,a} := \kappa_{j,n,a}^{1-\rho} \), which is given by

\[
\kappa_{j,n} = \frac{1}{eg-1} \mathbb{P}\{|Z_{ki} - gZ_{E_1} - 1\} > (k,i) \in E_0\},
\]

with \( g = g^{1-\rho,j,n} \) as defined earlier.

In particular, \( \mathbb{P}\{V > x\} \) is regularly varying of index \(-\mu^*\).

**Proof**

For compactness, denote

\[
P_c^e(x) := \sum_{n \in S^*} \sum_{j \leq n} \prod_{k=1}^{K} P_{j,n,a}^{n_k} (x),
\]

and

\[
P(x) := \sum_{n \in S^*} \sum_{j \leq n} \prod_{k=1}^{K} \frac{n_k}{j^k} P_{j,n}(x).
\]

We need to show that

\[
\lim_{x \to \infty} \frac{\mathbb{P}\{V > x\}}{P(x)} = 1.
\]

We may write, for any \( a > 0 \),

\[
\mathbb{P}\{V > x\} = \frac{\mathbb{P}\{V > x\}}{P(x)} \mathbb{P}\{V(az) > x\} = \mathbb{P}\{V(ax) > x\} \mathbb{P}\{V(ax) > x\} / P(x).
\]

Because of Theorem 4.1, it thus suffices to show that

\[
\lim_{a \to \infty} \lim_{x \to \infty} \frac{\mathbb{P}\{V(ax) > x\}}{P(x)} = 1. \tag{4.8}
\]

First observe that if \( \min_{a} > 1 - \rho \), then there exists an \( a_0 \) such that \( S^*_a = S^* \) for all \( a \geq a_0 \).

Also, combining Lemmas 4.1, 4.2, 4.4, we have that for \( \delta, \theta \) sufficiently small,

\[
\mathbb{P}\{V_{\geq \theta}(ax) > (1 \pm \theta)x\} \sim P_{a}^{1-\rho \pm \delta}((1 \pm \theta)x). \tag{4.9}
\]

The proof of (4.8) consists of a lower and an upper bound.

**Lower bound**

Using Proposition 4.2 and Equation (4.9), we obtain that for \( \delta > 0, \theta > 0 \) sufficiently small

\[
\mathbb{P}\{V(ax) > x\} \gtrsim P_{a}^{1-\rho + \delta}((1 + \theta)x).
\]
Thus, for all \( a \geq a_0 \),
\[
\frac{\mathbb{P}\{V(ax) > x\}}{P(x)} \geq \min_{n \in S^*} \sum_{j = n}^{n+K} \mathbb{E}_{j^k} P_{j,n,a}^{1-\rho+\delta} ((1+\theta)x) \left( \sum_{k=1}^{K} \mathbb{P}_{j,k} P_{j,n}(x) \right) \geq \min_{n \in S^*} \sum_{j = n}^{n+K} \mathbb{E}_{j^k} P_{j,n,a}^{1-\rho+\delta} ((1+\theta)x) \left( \sum_{k=1}^{K} \mathbb{P}_{j,k} P_{j,n}(x) \right).
\]

Letting \( \theta \downarrow 0 \), using the fact that \( P_{j,n,a}(x) \) is regularly varying, we find
\[
\liminf_{x \to \infty} \frac{\mathbb{P}\{V(ax) > x\}}{P(x)} \geq \min_{n \in S^*} \sum_{j = n}^{n+K} \mathbb{E}_{j^k} P_{j,n,a}^{1-\rho+\delta} ((1+\theta)x) \left( \sum_{k=1}^{K} \mathbb{P}_{j,k} P_{j,n}(x) \right) \geq \min_{n \in S^*} \sum_{j = n}^{n+K} \mathbb{E}_{j^k} P_{j,n,a}^{1-\rho+\delta} ((1+\theta)x) \left( \sum_{k=1}^{K} \mathbb{P}_{j,k} P_{j,n}(x) \right).
\]

Letting \( \delta \downarrow 0 \), recalling that \( \kappa_{j,n,a}^c \) is continuous in \( c \) in a neighborhood of \( 1 - \rho \), and then \( a \to \infty \), the desired lower bound follows.

**Upper bound**

Using Proposition 4.4 and Equation (4.9), we obtain that for \( \delta > 0 \), \( \theta > 0 \) sufficiently small
\[
\mathbb{P}\{V(ax) > x\} \lesssim P_a^{1-\rho-\delta} (1-\theta)x.
\]

Thus, for all \( a \geq a_0 \),
\[
\frac{\mathbb{P}\{V(ax) > x\}}{P(x)} \leq \max_{n \in S^*} \sum_{j = n}^{n+K} \mathbb{E}_{j^k} P_{j,n,a}^{1-\rho+\delta} ((1-\theta)x) \left( \sum_{k=1}^{K} \mathbb{P}_{j,k} P_{j,n}(x) \right) \leq \max_{n \in S^*} \sum_{j = n}^{n+K} \mathbb{E}_{j^k} P_{j,n,a}^{1-\rho+\delta} ((1-\theta)x) \left( \sum_{k=1}^{K} \mathbb{P}_{j,k} P_{j,n}(x) \right).
\]

Letting \( \theta \downarrow 0 \), using the fact that \( P_{j,n,a}(x) \) is regularly varying, we conclude
\[
\limsup_{x \to \infty} \frac{\mathbb{P}\{V(ax) > x\}}{P(x)} \leq \max_{n \in S^*} \sum_{j = n}^{n+K} \mathbb{E}_{j^k} P_{j,n,a}^{1-\rho+\delta} ((1-\theta)x) \left( \sum_{k=1}^{K} \mathbb{P}_{j,k} P_{j,n}(x) \right) \leq \max_{n \in S^*} \sum_{j = n}^{n+K} \mathbb{E}_{j^k} P_{j,n,a}^{1-\rho+\delta} ((1-\theta)x) \left( \sum_{k=1}^{K} \mathbb{P}_{j,k} P_{j,n}(x) \right).
\]

Letting \( \delta \downarrow 0 \), recalling that \( \kappa_{j,n,a}^c \) is continuous in \( c \) in a neighborhood of \( 1 - \rho \), and then \( a \to \infty \), the desired upper bound follows.

\[\square\]

### 4.6 Transient workload asymptotics

Recall that the steady-state workload asymptotics were obtained from an analysis of the asymptotic behavior of \( \mathbb{P}\{V(ax) > x\} \) for \( x \to \infty \) after letting \( a \to \infty \). This raises the question whether it is possible to obtain the exact asymptotics of \( \mathbb{P}\{V(ax) > x\} \) for \( x \to \infty \) for any value of \( a \). To answer this question, we first consider the case where \( a \) is large enough for the condition \( \rho_a^\text{max} (1-\rho) < 1-\rho \) to hold, which implies that the overflow scenarios in the transient and steady-state case coincide.
Theorem 4.2 If \( r_a^{\text{max}}(1 - \rho) < 1 - \rho \), then

\[
P(V(ax) > x) \sim P_a^{1-\rho}(x).
\]

Proof
The proof is largely similar to that of Theorem 3.1 in the previous subsection, except that the use of Theorem 4.1 is not needed now.

Unfortunately, it seems difficult to remove the condition \( r_a^{\text{max}}(1 - \rho) < 1 - \rho \) in the above theorem. This condition is induced by the use of Lemma 4.2, where it is needed to ensure that the process \( \{A_{i>ex}(0,s) - cs\} \) reaches its supremum over the interval \([0,ax]\) at time \( \bar{t}_f^l(ex) \). This is no longer guaranteed to be the case when \( r_a^{\text{max}}(1 - \rho) > 1 - \rho \). In that case, the event \( A_{i>ex}(0,\bar{t}_f^l(ex)) - c\bar{t}_f^l(ex) > x \) is by far not necessary for the event \( V_{c,ex}^n(ax) > x \) to occur because the drift of the process \( \{A_{i>ex}(0,s) - cs\} \) may remain positive after time \( \bar{t}_f^l(ex) \). This necessitates a detailed analysis of the process \( \{A_{i>ex}(0,s) - cs\} \) after time \( \bar{t}_f^l(ex) \), which seems rather difficult, even in the single-class case \( K = 1 \).

Nevertheless, it is possible to apply the earlier results to obtain asymptotically tight lower and upper bounds for \( P(V(ax) > x) \) as \( x \to \infty \), which hold for any value of \( a \), under the considerably milder condition \( r_a^{\text{min}} > 1 - \rho \).

Theorem 4.3 Assume that \( r_a^{\text{min}} > 1 - \rho \). Then,

\[
P_a^{1-\rho}(x) \preceq P(V(ax) > x) \preceq P_a^{1-\rho}(x(1 - a(r_a^{\text{max}} + \rho - 1))).
\]

Proof
The lower bound follows directly from Lemmas 4.2 and 4.4 and the results of Subsection 4.1. For the upper bound, note that the drift of the process \( \{A_{i>ex}(0,s) - cs\} \) is at most \( r_a^{\text{max}} - c \) after time \( \bar{t}_f^l(ex) \). Hence, this process can increase by at most \( a(r_a^{\text{max}} - c)x \) until time \( ax \). This implies that one must have \( A_{i>ex}(0,\bar{t}_f^l(ex)) - c\bar{t}_f^l(ex) > x(1 - a(r_a^{\text{max}} - c)) \) in order for the event \( V_{c,ex}^n(ax) > x \) to occur. The proof of the upper bound is then completed by using Lemma 4.4.

\[\square\]

Note that the upper bound in the above theorem is non-trivial because \( r_a^{\text{max}} + \rho - 1 < \frac{1}{\alpha} \). Moreover, the bounds asymptotically coincide up to a constant factor, since the function \( P_a^{1-\rho}(\cdot) \) is regularly varying of index \(-\mu_{\alpha}^*\).

5 Proof of Theorem 4.1

In this section we provide the proof of Theorem 4.1. We first collect some preparatory results. For conciseness, we drop \( a = \infty \) from the previously introduced notation to denote steady-state quantities. For example \( S^*(c) := S^*_{\infty}(c) \) is the set of optimal solutions of the linear
program formulated at the beginning of Section 4, \( r_{\min}(c) := r_{\infty, x}(c) = \min_{n \in S^*(c)} \sum_{k=1}^{K} n_k r_k \),
\( S^{-}(c) := S_{\infty}^{-}(c) = \{ n \in \mathbb{N}^{K} : \sum_{k=1}^{K} n_k r_k < c \}, \)
\( S^{+}(c) := S_{\infty}^{+}(c) = \{ n \in \mathbb{N}^{K} : \sum_{k=1}^{K} n_k r_k \geq c \}, \) and
\( r_{\max}(c) := r_{\infty}^{\max}(c) = \max_{n \in S^{-}(c)} \sum_{k=1}^{K} n_k r_k < c. \)

**Proposition 5.1** Assume that \( r_{\min}(c) > c. \) Then for all \( \epsilon < \frac{1}{r_{\min}(c)-c}, \)
\[ \Pr\{ V_{c, \epsilon} > x \} \geq \sum_{n \in S^*(c)} \prod_{k=1}^{K} e^{-\rho_k^* n_k} \left( \frac{x}{r_{\min}(c)-c} \right)^{n_k}. \]

**Proof**
Consider the event that at some arbitrary time \( t \) there are exactly \( n_k \) active class-\( k \) sessions, \( k = 1, \ldots, K, n \in S^*(c), \) which all started before time \( t - \frac{x}{r_{\min}(c)-c}. \)
Since \( \epsilon < \frac{1}{r_{\min}(c)-c}, \) this event implies that \( V_{c, \epsilon}(t) \) is larger than
\[ \sum_{k=1}^{K} n_k r_k - c \cdot \frac{x}{r_{\min}(c)-c} \geq x, \]
while it occurs with probability
\[ \prod_{k=1}^{K} e^{-\rho_k^* n_k} \left( \frac{x}{r_{\min}(c)-c} \right)^{n_k}. \]

\( \square \)

**Proposition 5.2** Consider a queue of capacity \( c \) fed by a process which generates traffic at rate \( r_n \) for a fraction of the time \( p_n, n = 1, \ldots, N \) (possibly \( N = \infty \)). Assume \( r_1 \leq r_2 \leq \ldots r_{K-1} < c \leq r_K \leq \ldots \leq r_N, \) and \( \sum_{n=1}^{N} p_n r_n < c \) for stability. Let \( V^c \) be the stationary workload. Then for any \( x > 0 \)
\[ \Pr\{ V^c > 0 \} \leq \frac{1}{c - r_{K-1}} \sum_{n=K}^{N} p_n (r_n - r_{K-1}). \]

**Proof**
First observe that \( \Pr\{ V^c > 0 \} \leq \pi, \) where the latter quantity represents the stationary probability that the workload is non-zero if the rate \( r_n \) were increased to \( r_{K-1} \) for all \( n = 1, \ldots, K-1. \)
From a simple balance argument, noting the workload cannot be zero when traffic is generated at a rate \( p_n > c, \)
\[ \sum_{n=K}^{N} p_n (r_n - c) = (\pi - \sum_{n=K}^{N} p_n) (c - r_{K-1}), \]
yielding
\[
\pi = \frac{1}{c - r_{K-1}} \sum_{n=K}^{N} p_{n}(r_{n} - r_{K-1}),
\]
which completes the proof.

**Proposition 5.3**

\[
\mathbb{P}(V_{>\varepsilon} > 0) \sim \frac{\max_{k=1,\ldots,K} r_k}{c - r_{\max}(c)} \sum_{n \in S^*(c)} \prod_{k=1}^{K} \frac{\beta_{nk}^{n_k}}{n_k!} (\mathbb{P}(B_k^r > \varepsilon))^{n_k}.
\]

**Proof**

Using Equation (2.1) and Proposition 5.2,

\[
\mathbb{P}(V_{>\varepsilon} > 0) \leq \frac{1}{c - r_{\max}(c)} \sum_{n \in S^*(c)} \prod_{k=1}^{K} \frac{\beta_{nk}^{n_k}}{n_k!} (\mathbb{P}(B_k^r > \varepsilon))^{n_k} \leq \frac{1}{c - r_{\max}(c)} \sum_{n \in S^*(c)} \prod_{k=1}^{K} \frac{\beta_{nk}^{n_k}}{n_k!} (\mathbb{P}(B_k^r > \varepsilon))^{n_k}.
\]

Note that

\[
\sum_{n \in S^*(c)} \left( \prod_{k=1}^{K} \frac{\beta_{nk}^{n_k}}{n_k!} (\mathbb{P}(B_k^r > \varepsilon))^{n_k} \right) \leq \max_{n \in S^*(c)} \sum_{k=1}^{K} n_k r_k - r_{\max}(c) \leq \max_{k=1,\ldots,K} r_k.
\]

From the definition of \(S^*(c)\) it follows that there exists an \(x_0\) such that for all \(x \geq x_0\),

\[
\prod_{k=1}^{K} \mathbb{P}(B_k^r > \varepsilon)^{m_k} \leq H(x) \prod_{k=1}^{K} \frac{\beta_{nk}^{n_k}}{n_k!} (\mathbb{P}(B_k^r > \varepsilon))^{n_k},
\]

for all \(m \in S^+(c) \setminus S^*(c), n \in S^*(c)\), with \(H(x) = o(1)\) as \(x \to \infty\), so that

\[
\sum_{m \in S^+(c) \setminus S^*(c)} \left( \prod_{k=1}^{K} \frac{\beta_{nk}^{n_k}}{n_k!} (\mathbb{P}(B_k^r > \varepsilon))^{m_k} \right) \leq \max_{k=1,\ldots,K} r_k.
\]
\[
\begin{align*}
&\leq H(x) \sum_{m \in S^+(c) \setminus S^*(c)} \sum_{k=1}^{K} m_k r_k - r^{\text{max}}(c) \prod_{k=1}^{K} \frac{\bar{\rho}^{m_k}}{m_k!} \\
&\leq H(x) \sum_{m \geq 0} \left( \sum_{k=1}^{K} m_k r_k \right) \prod_{k=1}^{K} \frac{\bar{\rho}^{m_k}}{m_k!} \\
&= H(x) \left( \sum_{k=1}^{K} \bar{\rho}_k r_k \right) e^{\sum_{k=1}^{K} \rho_k} \\
&= \rho H(x) e^{\sum_{k=1}^{K} \rho_k}.
\end{align*}
\]

Hence,

\[
\limsup_{x \to \infty} \frac{\mathbb{P}\{\mathbf{V}^{\leq \epsilon_x} > 0\}}{\mathbb{E}\left\{ \mathbb{E}\left[ A(0,t) \right] \right\}} \leq \max_{k=1,\ldots,K} \frac{r_k}{c - r^{\text{max}}(c)}.
\]

We have now gathered all the ingredients for the proof of Theorem 4.1 which is repeated below.

**Theorem 4.1**

If \( r^{\text{min}} > 1 - \rho \), then

\[
\lim_{a \to \infty} \lim_{x \to \infty} \frac{\mathbb{P}\{\mathbf{V}(ax) > x\}}{\mathbb{P}\{\mathbf{V} > x\}} = 1.
\]

**Proof**

By definition,

\[
\mathbb{P}\{\mathbf{V} > x\} = \mathbb{P}\{\sup_{t \geq 0} A(0,t) - t > x\} \\
\leq \mathbb{P}\{\sup_{t \leq ax} A(0,t) - t > x\} + \mathbb{P}\{\sup_{t \geq ax} A(0,t) - t > x\} \\
= \mathbb{P}\{\mathbf{V}(ax) > x\} + \mathbb{P}\{\sup_{t \geq ax} A(0,t) - t > x\}.
\]

Thus, it suffices to show that

\[
\lim_{a \to \infty} \limsup_{x \to \infty} \frac{\mathbb{P}\{\sup_{t \geq ax} A(0,t) - t > x\}}{\mathbb{P}\{\mathbf{V} > x\}} = 0.
\]

For \( t \geq ax \), write

\[
A(0,t) - t = A(0,ax) - ax + A(ax, t) - (t - ax),
\]

and observe that \( A(ax, t) \overset{d}{=} A(0, t - ax) \) since the process \( A(0,t) \) has stationary increments.
Thus, for $\delta > 0$ sufficiently small,
\[
\mathbb{P}\{\sup_{t \geq ax} \{A(0, t) - t\} > 0\} = \mathbb{P}\{\sup_{t \geq ax} \{A(0, ax) - ax + A(ax, t) - (t - ax)\} > 0\} \leq \mathbb{P}\{A(0, ax) - ax > -\delta ax\} + \mathbb{P}\{\sup_{t \geq ax} \{A(0, t - ax) - (t - ax)\} > \delta ax\}
\]
\[
= \mathbb{P}\{A(0, ax) - ax > (1 - 2\delta)ax > \delta ax\} + \mathbb{P}\{\sup_{t \geq ax} \{A(0, t - ax) - (t - ax)\} > \delta ax\}
\]
\[
\leq \mathbb{P}\{\sup_{t \geq 0} \{A(0, t - (1 - 2\delta)t)\} > \delta ax\} + \mathbb{P}\{\sup_{t \geq 0} \{A(0, t - t)\} > \delta ax\}
\]
\[
\leq 2\mathbb{P}\{V^{1-2\delta} > \delta ax\} + \mathbb{P}\{V > \delta ax\}.
\]
Hence, using Propositions 4.2, 4.5, for $\theta > 0$ sufficiently small,
\[
\limsup_{x \to \infty} \frac{\mathbb{P}\{\sup_{t \geq ax} \{A(0, t) - t\} > x\}}{\mathbb{P}\{V > x\}} \leq 2\limsup_{x \to \infty} \frac{\mathbb{P}\{V^{1-2\delta} > \delta ax\}}{\mathbb{P}\{V > x\}} \leq 2\limsup_{x \to \infty} \frac{\mathbb{P}\{V^{1-\rho-3\delta} > (1 - \theta)\delta ax\}}{\mathbb{P}\{V^{1-\rho+\delta} > x\}} \leq 2\limsup_{x \to \infty} \frac{\mathbb{P}\{V^{1-\rho-3\delta} > (1 - \theta)\delta x/(1 + \theta)\}}{\mathbb{P}\{V^{1-\rho+\delta} > x/a\}} \leq 2\limsup_{x \to \infty} \frac{\mathbb{P}\{V^{1-\rho-3\delta} > 0\}}{\mathbb{P}\{V^{1-\rho+\delta} > x/a\}}.
\]

The assumption that $r_{\min} > 1 - \rho$ ensures that there exists a $\delta^* > 0$ such that $r_{\min} > 1 - \rho + \delta^*$, $r_{\max} < 1 - \rho - 3\delta^*$, and $S^*(1 - \rho - 3\delta^*) = S^*(1 - \rho + \delta^*) = S^*$.

Using Propositions 5.1, 5.3, we then find that there exists an $\epsilon^* > 0$ such that for all $\epsilon < \epsilon^*$,
\[
\limsup_{x \to \infty} \frac{\mathbb{P}\{\sup_{t \geq ax} \{A(0, t) - t\} > x\}}{\mathbb{P}\{V > x\}} \leq 2\max_{k=1, \ldots, K} \frac{\rho_k}{1 - \rho - r_{\max} - 3\delta^*} \left( \sum_{n \in S^*} \frac{\rho_n}{n!} (\mathbb{P}\{B_k^n > x\})^{n_k} \right) \leq 2 \max_{k=1, \ldots, K} \frac{\rho_k}{1 - \rho - r_{\max} - 3\delta^*} \left( \sum_{n \in S^*} \frac{\rho_n}{n!} (\mathbb{P}\{B_k^n > x\})^{n_k} \right) \max_{n \in S^*} \limsup_{x \to \infty} \left( \frac{\mathbb{P}\{B_k^n > x\}}{\mathbb{P}\{B_k^n > x/a\}} \right)^{n_k}.
\]

Now first let $x \to \infty$ and then $a \to \infty$ (use the fact that $\mathbb{P}\{B_k^n > x\}$ is of regular variation).
6 Most probable time to overflow

As a direct application of the workload asymptotics which we derived in the previous sections, we now establish a conditional limit theorem for the most probable time to overflow, given that the process \( \{A(0, t) - ct\} \) reaches a large level \( x \). Define \( \tau(x) = \inf\{t \geq 0 : A(0, t) - ct > x\} \). Note that \( V \geq x \) iff \( \tau(x) < \infty \). We will give an expression for the asymptotic distribution of \( \tau(x) \) conditional upon \( \tau(x) < \infty \) for \( x \to \infty \). Define the probability measure \( \mathbb{P}_x \{ \cdot \} := \mathbb{P}\{ \cdot | \tau(x) < \infty \} \). In this section we compute the limiting \( \mathbb{P}_x \)-distribution of \( \frac{\tau(x)}{x} \) for \( x \to \infty \).

A similar problem has been investigated by Asmussen & Klüppelberg [1] for random walks and Lévy processes with negative drift and heavy-tailed jumps. As has been shown in [1], this class of processes allows for a general subexponential jump size distribution. Here though, like in the rest of the paper, we consider the case of regular variation. In fact, since slowly varying functions may be difficult to compare in the multi-class case, we assume that the session lengths are Pareto distributed, i.e.,

\[
\mathbb{P}(B_k^r > x) \sim \gamma_k x^{1 - \nu_k}, \quad k = 1, \ldots, K.
\]

This assumption may be weakened, as will be discussed below.

In order to state the result, we need to introduce some additional notation. For given \( a \), define the set \( S_a \) as \( S_a := \{ n \in S^* : \sum_{k=1}^K r_k n_k \geq 1 - \rho + \frac{1}{a} \} \). We will also make extensive use of the coefficients \( \kappa_{j,n} \) and \( \kappa_{j,n,a} \) defined earlier. The definition of \( \kappa_{j,n,a} \) as given in Subsection 4.4 only makes sense for \( \sum_{k=1}^K r_k n_k > 1 - \rho + \frac{1}{a} \). If \( \sum_{k=1}^K r_k n_k = 1 - \rho + \frac{1}{a} \), we define \( \kappa_{j,n,a} = 1_{\{j=n\}} \).

**Theorem 6.1** The quantity \( \frac{\tau(x)}{x} \) converges in \( \mathbb{P}_x \)-distribution for \( x \to \infty \) to a random variable \( Y \), which has distribution function

\[
G(a) := \mathbb{P}\{Y \leq a\} = \frac{\sum_{n \in S_a} \sum_{j \leq n} d_n^{\ast} \kappa_{j,n,a} \prod_{k=1}^K (\frac{\beta_k \gamma_k}{J_k})^{n_k}}{\sum_{n \in S^*} \sum_{j \leq n} d_n^{\ast} \kappa_{j,n} \prod_{k=1}^K (\frac{\beta_k \gamma_k}{J_k})^{n_k}}, \tag{6.1}
\]

with \( d_n = \sum_{k=1}^K n_k r_k + \rho - 1 \) as before.

**Proof**

First observe that the extended definition of \( \kappa_{j,n,a} \) ensures that \( \kappa_{j,n,a} \) is right-continuous in \( a \) if \( a \) is such that \( \sum_{k=1}^K r_k n_k = 1 - \rho + \frac{1}{a} \). This then implies that the function \( G(\cdot) \) is right-continuous. From the analysis in the previous sections, it follows that \( G(\cdot) \) is non-decreasing and that \( G(a) \to 1 \) as \( a \to \infty \). Hence, \( G(\cdot) \) is a proper distribution function, so that \( Y \) is a well-defined random variable.

We need to show that \( \mathbb{P}_x \{ \tau(x) < ax\} \to G(a) \) as \( x \to \infty \) for each continuity point of \( G(\cdot) \). Using the definition of \( S_a \) and the (extended) definition of \( \kappa_{j,n,a} \), it is easy to see that \( G(\cdot) \) is.
continuous in $a$ iff $\sum_{k=1}^{K} r_kn_k > 1 - \rho + \frac{1}{a}$ for all $n \in S_a$ (look at the structure of $S_a$). Hence, we may assume that $a$ is such that $\sum_{k=1}^{K} r_kn_k > 1 - \rho + \frac{1}{a}$ for all $n \in S_a$.

Now write

$$P\{\tau(x) \leq ax | \tau(x) < \infty\} = \frac{P\{\tau(x) \leq ax\}}{P\{\tau(x) < \infty\}} = \frac{P\{V(ax) \geq x\}}{P\{V \geq x\}} \sim \frac{P\{V(ax) > x\}}{P\{V > x\}}.$$  

Note that $P\{V > x\}$ is regularly varying of index $-\mu^*$. If $\sum_{k=1}^{K} r_kn_k < 1 - \rho + \frac{1}{a}$ for all $n \in S^*$ (i.e. $S_a = \emptyset$), then it is obvious that $P\{V(ax) > x\}$ is regularly varying of index $-\mu^*_a < -\mu^*$. This implies that $P\{V(ax) > x\}/P\{V > x\} \to 0$ if $a$ is small enough for $S_a$ to be empty.

Now suppose that $a$ is large enough such that $S_a$ is non-empty. It is then easy to see that

$$S_a = S^*_a.$$  

If we combine this identity with Theorems 3.1 and 4.2, we find, noting that $\sum_{k=1}^{K} (\nu_k - 1) = \mu^*$ for all $n \in S^*$,

$$P_x\{\tau(x) \leq ax\} = \frac{P\{V(ax) \geq x\}}{P\{V > x\}} \sum_{n \in S_a} \sum_{j \leq n} \sum_{k \in K} \frac{\kappa_{j,n,a} \prod_{k=1}^{K} \frac{\rho_k^{n_k}}{j_k!} (P\{B_k > a\})^{n_k}}{\sum_{n \in S_a} \sum_{j \leq n} \sum_{k \in K} \frac{\kappa_{j,n,a} \prod_{k=1}^{K} \frac{\rho_k^{n_k}}{j_k!} (\gamma_k d_n) - n_k (\nu_k - 1)}} \sum_{n \in S^*} \sum_{j \leq n} \sum_{k \in K} \frac{\kappa_{j,n,a} \prod_{k=1}^{K} (\rho_k \gamma_k)^{n_k}}{j_k!} \frac{\kappa_{j,n,a} \prod_{k=1}^{K} (\rho_k \gamma_k)^{n_k}}{j_k!}.$$

If the set $S^*$ is a singleton, then it is easy to see that regular variation suffices in the last two lines of the above proof. In particular, this is true in the single-class case $K = 1$.

We conclude the section with the most basic single-class scenario where overflow is caused by a single long session, which occurs when $r > 1 - \rho$. In this case, the distribution of $Y$ takes the explicit form

$$P\{Y \leq a\} = \frac{1 - \rho}{r} + \left(1 - \frac{1 - \rho}{r}\right) P\{\frac{r}{1 - \rho} Z \leq a - \frac{1}{r - (1 - \rho)}\},$$

31
where $\mathbb{P}\{Z > a\} = (1 + (r - (1 - \rho))a)^{1-\nu}$. This expression reduces to the results for the case of compound Poisson input in [1] when we let $r \to \infty$. The results in [1] further include conditional limit theorems for the behavior of the process $\{A(0, t) - ct\}$ up to time $\tau(x)$. Similar results may be derived for the case of $M/G/\infty$ input considered here.

References


