Exact Queueing Asymptotics for Multiple Heavy-Tailed On-Off Flows

Bert Zwart*, Sem Borst*,†,‡, Michel Mandjes‡

*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 multiple On-Off flows with heavy-tailed (regularly varying) On-periods. Under fairly mild assumptions, we prove that the workload distribution is asymptotically equivalent to that in a reduced system. The reduced system consists of a 'dominant' subset of the flows, with the original service rate subtracted by the mean rate of the other flows. We describe how a dominant set may be determined from a simple knapsack formulation. We exploit a powerful intuitive argument to obtain the exact asymptotics for the reduced system. Combined with the reduced-load equivalence, the results for the reduced system provide an asymptotic characterization of the buffer behavior.**

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

Keywords and Phrases: **fluid models, heavy-tailed distributions, knapsack problem, large deviations, queueing theory, reduced-load equivalence.**

I. Introduction

Over the past few decades, fluid models have gained strong ground as a versatile approach for analyzing burst-scale traffic behavior. The basic model is that of several On-Off sources, each alternating between activity phases (commonly referred to as bursts) and silence periods. When active, each source generates traffic at some constant rate.

Classical papers of Anick, Mitra, & Sondhi [2] and Kosten [19] considered a queue fed by the superposition of several homogeneous On-Off sources with exponentially distributed activity and silence periods. Subsequent work extended the model in various directions, such as heterogeneous source characteristics, several source states to account for various activity levels, or activity periods with a general Markovian structure, see for instance Kosten [20] and Stern & Elwalid [32]. Under traditional statistical assumptions, it turns out that the tail of the backlog distribution typically exhibits exponential decay.

In recent years, empirical findings have triggered a strong interest in fluid models with non-Markovian activity periods. Extensive measurements 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.* [21] and Paxson & Floyd [27]. The occurrence of these phenomena is commonly attributed to extreme variability and long-tailed characteristics in the underlying activity patterns (connection times, file sizes, scene lengths), see Beran *et al.* [4], Crovella & Bestavros [11] and Willinger *et al.* [33]. Fluid queues with long-tailed activity periods provide a natural paradigm for capturing these characteristics. We refer to Boxma & Dumas [9] for a survey paper.

Although the presence of long-tailed traffic characteristics is widely acknowledged, the practical implications for network performance and traffic engineering remain to be fully resolved. Analytical studies show potentially dramatic performance repercussions for infinite buffers. For moderate buffer sizes though, the impact of long-tailed traffic characteristics is not as pronounced, see Grossglauser & Bolot [14], Heyman & Lakshman [15], Mandjes & Kim [24], and Ryu & Elwalid [31]. For larger buffer sizes, flow control mechanisms play a critical role in preventing badly-behaved traffic from overwhelming the buffer content, see Arvidsson & Karlsson [3]. However, the amount of backlogged traffic at the user, and thus the end-to-end quality-of-service, may still be significantly affected by long-tailed activity patterns.

The effect of long-tailed traffic characteristics on buffer behavior also crucially depends on the relative amount of heavy-tailed traffic, in particular whether or not activity of heavy-tailed flows alone can cause the buffer to fill. Asymptotic bounds in Dumas & Simonian [12] indeed show a sharp dichotomy in the qualitative behavior of the workload, depending on whether the mean rate of the light-tailed flows plus the peak rate of the heavy-tailed flows exceeds the link rate or not. In case the link rate is larger, the workload distribution has light-tailed characteristics, whereas the link rate being smaller results in heavy-tailed characteristics. The exact asymptotics for the former case were recently obtained in [6]. For the latter case, the bounds of [12] indicate that one can usually identify a 'dominant' set, which is a minimal set of flows that can cause a positive drift in the buffer. As far as bounds is concerned, all other flows can essentially be accounted for by subtracting their aggregate mean rate from the link rate. Somewhat related notions are developed in Likhanov & Mazumdar [22] in the setting of $M/G/\infty$ input with heterogeneous sessions.

Exact results however, have remained elusive for all but a few special cases. Results of Agrawal *et al.* [1] show that the

dominance principle described above in fact extends to the exact asymptotics in the case of a *single* dominant flow. This may be expressed in terms of a 'reduced-load equivalence', implying that the workload is asymptotically equivalent to that in a reduced system. The reduced system consists only of the dominant flow, with the link rate subtracted by the aggregate mean rate of all other flows. This extends results of Boxma [8], Jelenković & Lazar [16], and Rolski *et al.* [30] for multiplexing a single (intermediately) regularly varying flow with several exponential flows. Related results are derived in Jelenković & Lazar [16] and Resnick & Samorodnitsky [29] in the context of $M/G/\infty$ input. Like the reduced-load equivalence, however, all these results rely on the assumption that a single active flow is sufficient for a positive drift in the buffer.

In the present paper we determine the exact asymptotics for the case where several On-Off flows must be active for the buffer to fill (under the assumption that the distribution of the On-periods is regularly varying [5]). From a practical perspective, this case appears particularly relevant, as the peak rate of a single flow is usually substantially smaller than the link rate. However, the rather subtle interaction of several flows that is involved in filling the buffer drastically complicates the analysis, reflecting the sharp demarcation in known results described above. We start with extending the reduced-load equivalence to the case of a reduced system consisting of several flows, using sample-path arguments. We then build on a qualitative understanding of the large-deviations behavior to obtain the exact asymptotics for the reduced system. This part of the analysis is related to recent work of Resnick & Samorodnitsky [29] on fluid queues with M/G/∞ input.

The remainder of the paper is organized as follows. In Section II, we present a detailed model description. In Section III, we give a broad overview of the main results of the paper, and describe how the dominant set may be determined from a simple knapsack formulation. We also discuss the relationship between the asymptotic regime considered here ('large buffers') and a many-sources regime. Section IV gives some preliminary results. The reduced-load equivalence result is established in Section V. Section VI develops the detailed probabilistic arguments involved in deriving the tail asymptotics for the reduced system.

II. Model description

We first present a detailed model description. We consider a queue of unit capacity fed by several flows indexed by the set $\mathcal{I}$. For any subset $E \subseteq \mathcal{I}$, denote by $A_E(s,t) := \sum_{i\in E} A_i(s,t)$ the aggregate amount of traffic generated by the flows $i \in E$ during the time interval $(s,t]$. Denote by $\rho_E := \sum_{i\in E} \rho_i$ the aggregate traffic intensity of the flows $i \in E$ (as will be specified in detail below). We assume $\rho := \rho_{\mathcal{I}} < 1$ for stability.

For any $c \geq 0$, $E \subseteq \mathcal{I}$, define $V_E^c(t) := \sup_{0\leq s\leq t}\{A_E(s,t) - c(t-s)\}$ as the workload at time t in a queue of capacity c fed by the flows $i \in E$ (assuming $V_E^c(0) = 0$). For $c > \rho_E$, let $\mathbf{V}_E^c$ be a random variable with the limiting distribution of $V_E^c(t)$ for $t \to \infty$. In particular, $V(t) := V_{\mathcal{I}}^1(t)$ is the total workload, and $\mathbf{V} := \mathbf{V}_{\mathcal{I}}^1$ is a random variable with the limiting distribution of $V(t)$ for $t \to \infty$.

We assume the flows may be partitioned into two sets: $\mathcal{I}_1$ is the set of 'light-tailed' flows; $\mathcal{I}_2$ is the set of 'heavy-tailed' flows. For the flows $i \in \mathcal{I}_1$ we make the following assumption.

Assumption II.1: For any $c > \rho_{\mathcal{I}_1}$, $\mu > 0$,

$$\lim_{x\to\infty} x^\mu \mathbb{P}\{\mathbf{V}_{\mathcal{I}_1}^c > x\} = 0.$$

The above assumption is satisfied for many input processes of practical interest, e.g. by On-Off flows with light-tailed or Weibullian On-periods.

We assume the flows in $\mathcal{I}_2$ generate traffic according to independent On-Off processes, each alternating between On- and Off-periods. The Off-periods of flow i are generally distributed with mean $1/\lambda_i$. The On-periods $\mathbf{A}_i$ have a heavy-tailed distribution $A_i(\cdot)$ with mean $\alpha_i < \infty$. While On, flow i produces traffic at constant rate r_i, so the mean burst size is $\alpha_i r_i$. The fraction of time that flow i is On is

$$p_i = \frac{\alpha_i}{1/\lambda_i + \alpha_i} = \frac{\lambda_i\alpha_i}{1+\lambda_i\alpha_i}.$$

Thus the traffic intensity of flow i is

$$\rho_i := p_i r_i = \frac{\lambda_i\alpha_i r_i}{1+\lambda_i\alpha_i}.$$

Before stating an important preliminary result, we first introduce some useful notation.

For any two real functions $f(\cdot)$ and $g(\cdot)$, we use the notational convention $f(x) \sim g(x)$ to denote $\lim_{x\to\infty} f(x)/g(x) = 1$. 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 positive stochastic variable $\mathbf{X}$ with distribution function $F(\cdot)$, $\mathbb{E}\{\mathbf{X}\} < \infty$, denote by $F^r(\cdot)$ the distribution function of the residual life-time of $\mathbf{X}$, i.e.,

$$F^r(x) := \frac{1}{\mathbb{E}\{\mathbf{X}\}} \int_0^x (1 - F(y))\mathrm{d}y,$$

and by $\mathbf{X}^r$ a stochastic variable with that distribution.

The classes of *long-tailed*, *subexponential*, *regularly varying*, and *intermediately regularly varying* distributions are denoted with the symbols $\mathcal{L}$, $\mathcal{S}$, $\mathcal{R}$, and $\mathcal{IR}$, respectively (note that $\mathcal{R} \subset \mathcal{IR} \subset \mathcal{S} \subset \mathcal{L}$). Background on heavy-tailed distributions may be found in Embrechts *et al.* [13].

For each flow $i \in \mathcal{I}_2$, we assume that the On-period distribution is regularly varying of index $-\nu_i$, i.e., $A_i(\cdot) \in \mathcal{R}_{-\nu_i}$ for some $\nu_i > 1$. The next result which is due to Jelenković & Lazar [16] then yields the tail behavior of the workload distribution.

Theorem II.1: If $A_i^r(\cdot) \in \mathcal{S}$, $\rho_i < c < r_i$, then

$$\mathbb{P}\{\mathbf{V}_i^c > x\} \sim (1-p_i)\frac{\rho_i}{c-\rho_i}\mathbb{P}\{\mathbf{A}_i^r > \frac{x}{r_i - c}\}.$$

III. Overview of the results

We now give a broad overview of the main results of the paper. As mentioned in the introduction, asymptotic bounds in Dumas & Simonian [12] show a sharp dichotomy in the qualitative behavior of $\mathbb{P}\{\mathbf{V} > x\}$, depending on the value of

$\rho_{\mathcal{I}_1} + r_{\mathcal{I}_2}$ (i.e. the mean rate of the light-tailed flows plus the peak rate of the heavy-tailed flows) relative to the service rate. In case $\rho_{\mathcal{I}_1} + r_{\mathcal{I}_2} < 1$, the workload has light-tailed characteristics, whereas $\rho_{\mathcal{I}_1} + r_{\mathcal{I}_2} > 1$ implies heavy-tailed characteristics. In the present paper we determine the exact asymptotics of $\mathbb{P}\{\mathbf{V} > x\}$ in the latter case.

A. Intuitive arguments

Before formulating our main theorems, we first provide a heuristic derivation of the tail behavior of $\mathbb{P}\{\mathbf{V} > x\}$.

Large-deviations theory suggests that, given that a 'rare event' occurs, with overwhelming probability 'it happens in the most likely way'. In the asymptotic regime considered here ('large buffers'), the most likely way usually consists of a linear build-up of the workload, due to temporary instability of the system. In case of heavy-tailed distributions, the temporary instability typically arises from a 'minimal set' of potential causes. The minimal set corresponds to the minimal *number* of causes when these are homogeneous in nature. In general however, when the potential causes have heterogeneous characteristics, not only the number of them matters, but also their relative likelihood, and their relative contribution to the occurrence of the rare event under consideration.

Translated to our situation, temporary instability is most likely caused by a 'minimal set' of flows generating an extreme amount of traffic, while all other flows show roughly average behavior. These considerations give rise to the following characterization of the tail behavior of $\mathbb{P}\{\mathbf{V} > x\}$:

$$\mathbb{P}\{\mathbf{V} > x\} \sim \mathbb{P}\{\mathbf{V}_{S^*}^{c_{S^*}} > x\},$$

with S^* representing the 'minimal set', and $c_{S^*} := 1 - \rho_{\mathcal{I}\setminus S^*}$ the service rate subtracted by the aggregate traffic intensity of all other flows.

We now introduce some helpful notions in order to formalize the above intuitive arguments. For any subset $S \subseteq \mathcal{I}_2$, define $c_S := 1 - \rho_{\mathcal{I}\setminus S}$ as the service rate subtracted by the aggregate traffic intensity of all other flows $j \in \mathcal{I} \setminus S$. Observe that the stability condition implies $\rho_S < c_S$ for any $S \subseteq \mathcal{I}_2$.

For any subset $S \subseteq \mathcal{I}_2$, denote by $r_S := \sum_{j\in S} r_j$ the aggregate peak rate of the flows $j \in S$. Define $d_S := r_S - c_S = r_S + \rho_{\mathcal{I}\setminus S} - 1$ as the net input rate (i.e. the drift) when all flows in S are On and all other flows show average behavior.

A set $S \subseteq \mathcal{I}_2$ is called (strictly) *critical* if $d_S \geq (>) 0$, i.e., if

$$r_S + \rho_{\mathcal{I}\setminus S} \geq (>)\ 1.$$

Thus, when all flows in a (strictly) critical set are On, the workload has a (strictly) positive drift. A critical set S is termed *minimally-critical* if no proper subset of S is critical, i.e., $d_S < \min_{j\in S}\{r_j - \rho_j\}$.

For any subset $S \subseteq \mathcal{I}_2$, denote $\mu_S := \sum_{j\in S}(\nu_j - 1)$. A strictly critical set $S \subseteq \mathcal{I}_2$ is said to be (weakly) *dominant* if $\mu_S < (\leq)\ \mu_U$ for any other critical set $U \subseteq \mathcal{I}_2$. Observe that for a set $S \subseteq \mathcal{I}_2$ to be dominant, it must be minimally-critical (because otherwise the defining property would be violated for any critical subset $U \subset S$).

The quantity μ_S may be interpreted as a measure for the 'cost' associated with a temporary drift d_S: the probability of all flows in S being On for a time of the order x in steady state is roughly equal to $x^{-\mu_S}$. Thus, a set S is (weakly) dominant if the flows in S being On causes the drift to be positive in the cheapest possible way.

In case of light-tailed distributions, the cost minimization is usually not so simple; one then also needs to consider how long a certain positive drift must be be maintained in order for a given workload level x to be reached. This issue does not arise in case of regularly varying On periods, since $\mathbb{P}\{\mathbf{A}_i^r > ax\}$ is of the same order of magnitude (up to a constant) as $\mathbb{P}\{\mathbf{A}_i^r > x\}$ for any constant $a > 1$. This implies that the value of the temporary drift is not relevant as long as it is positive.

B. Tail behavior of the workload distribution

We now state our main theorem.

Theorem III.1: (Reduced-load equivalence)

Suppose the set of flows $S^* \subseteq \mathcal{I}_2$ is dominant. If $A_j(\cdot) \in \mathcal{R}$ for all $j \in \mathcal{I}_2$, then

$$\mathbb{P}\{\mathbf{V} > x\} \sim \mathbb{P}\{\mathbf{V}_{S^*}^{c_{S^*}} > x\}, \tag{3.1}$$

with

$$\mathbb{P}\{\mathbf{V}_{S^*}^{c_{S^*}} > x\} \sim \prod_{j\in S^*} p_j \sum_{\mathcal{J}_0 \subseteq S^*} P_{\mathcal{J}_0}(x), \tag{3.2}$$

where $P_{\mathcal{J}_0}(x)$ is given by (with $\mathcal{J}_1 = S^* \setminus \mathcal{J}_0$, and $d_{S^*} = r_{S^*} - c_{S^*}$ as defined earlier)

$$P_{\mathcal{J}_0}(x) = \frac{1}{\prod_{i\in\mathcal{J}_1}\mathbb{E}\{\mathbf{A}_i\}} \int_{y_i\in(0,\infty), i\in\mathcal{J}_1} \tag{3.3}$$
$$\prod_{i\in\mathcal{J}_1} \mathbb{P}\{d_{S^*}\mathbf{A}_i > \sum_{j\in\mathcal{J}_1} y_j(r_j - \rho_j) - d_{S^*}y_i + x\}$$
$$\prod_{i\in\mathcal{J}_0} \mathbb{P}\{d_{S^*}\mathbf{A}_i^r > \sum_{j\in\mathcal{J}_1} y_j(r_j - \rho_j) + x\} \prod_{i\in\mathcal{J}_1} \mathrm{d}y_i.$$

The proof of the above theorem may be found in Section V (Equation (3.1)) and Section VI (Equations (3.2) and (3.4) and the regular variation property).

Note that in case the reduced system consists of just a single flow, i.e., $S^* = \{i^*\}$, the tail asymptotics follow directly from Theorem II.1. This is in fact the reduced-load equivalence established in Agrawal, Makowski & Nain [1] (under somewhat weaker distributional assumptions). Note that in this case the right-hand side of (3.2) takes the form $p_{i^*}[P_\emptyset(x) + P_{i^*}(x)]$, which is consistent with Theorem II.1.

In case the reduced system consists of several flows, the tail asymptotics cannot be obtained from known results. In fact, the analysis of the reduced system then poses a major challenge because of the rather subtle mechanics involved in reaching a large workload level. By definition though, the reduced system has the special feature that all flows must be On for the drift in the workload to be positive, i.e., $r_{S^*} - \min_{j\in S^*}\{r_j - \rho_j\} < c_{S^*} < r_{S^*}$. In Section VI we determine the exact asymptotics for systems satisfying this property, yielding the integral expression given in Theorem III.1.

C. Knapsack formulation for determining a dominant set

We now describe how a dominant set may be determined from a simple knapsack formulation (for a related optimization problem, see [22]). Recall that the On-period distributions of the flows $i \in \mathcal{I}_2$ are regularly varying of index $-\nu_i$.

For a strictly critical set $S \subseteq \mathcal{I}_2$ to be dominant, it must necessarily solve the optimization problem

$$\begin{aligned} \min_{S\subseteq\mathcal{I}_2} \quad & \sum_{j\in S}(\nu_j - 1) \\ \text{sub} \quad & \sum_{j\in S} r_j + \sum_{j\in\mathcal{I}_2\setminus S} \rho_j > 1 - \rho_{\mathcal{I}_1}. \end{aligned}$$

Note that the constraint is equivalent to $d_S > 0$. If we define $\theta_i := r_i - \rho_i$ for all $i \in \mathcal{I}_2$, then the above problem may be expressed in the standard knapsack form as

$$\begin{aligned} \max_{U\subseteq\mathcal{I}_2} \quad & \sum_{j\in U}(\nu_j - 1) \\ \text{sub} \quad & \sum_{j\in U} \theta_j \leq \rho_{\mathcal{I}_1} + r_{\mathcal{I}_2} - 1 - \epsilon, \end{aligned}$$

with $U = \mathcal{I}_2 \setminus S$ and ϵ some small positive number. The above problem may not always have a unique solution. In case it does, the corresponding set S is dominant, except for the case when some set T exists which is critical but not strictly critical (i.e. $r_T + \rho_{I\setminus T} = 1$), with $\mu_T \leq \mu_S$ (see the definition of a dominant set). Although intriguing, this 'critical case' is not further considered in the present paper. In this case, the temporary drift may be *zero* for a long period of time during the path to overflow.

In case the knapsack problem has several solutions, the corresponding sets are weakly dominant (except for the critical case again). The next theorem extends the reduced-load equivalence to the case of weakly dominant sets.

Theorem III.2: (Generalized reduced-load equivalence)

Let $\Upsilon \subseteq 2^{\mathcal{I}_2}$ be the collection of all weakly dominant sets. If $A_j(\cdot) \in \mathcal{R}$ for all $j \in S$, $S \in \Upsilon$, then

$$\mathbb{P}\{\mathbf{V} > x\} \sim \sum_{S\in\Upsilon} \mathbb{P}\{\mathbf{V}_S^{c_S} > x\}, \tag{3.4}$$

with $\mathbb{P}\{\mathbf{V}_S^{c_S} > x\}$ as in (3.2), (3.3).

D. Homogeneous On-Off flows

We briefly consider the case of homogeneous On-Off flows as an important special case with weakly dominant sets. Assume that the flows $i \in \mathcal{I}_2$ have identical characteristics. With some minor abuse of notation, let $A(\cdot) := A_i(\cdot)$, $\nu := \nu_i$, $\rho := \rho_i$, $r := r_i$, $p_i \equiv p$. Define $N^* := \arg\min\{N : Nr + (|\mathcal{I}_2| - N)\rho > 1 - \rho_{\mathcal{I}_1}\}$. (Observe that the assumption $\rho_{\mathcal{I}_1} + r_{\mathcal{I}_2} > 1$ ensures $N^* \leq |\mathcal{I}_2|$.) To exclude the critical case, assume that $(N^*-1)r + (|\mathcal{I}_2| - N^* + 1)\rho < 1 - \rho_{\mathcal{I}_1}$, so that the drift remains negative (and cannot be zero) when only $N^* - 1$ flows are On.

Corollary III.1: If $A(\cdot) \in \mathcal{R}$, then

$$\mathbb{P}\{\mathbf{V} > x\} \sim \binom{|\mathcal{I}_2|}{N^*} \mathbb{P}\{\bar{\mathbf{V}} > x\},$$

with

$$\mathbb{P}\{\bar{\mathbf{V}} > x\} \sim p^{N^*} \sum_{n=0}^{N^*} \binom{N^*}{n} P_{\{1,\ldots,n\}}(x).$$

where $P_{\{1,\ldots,n\}}(x)$ is given by (3.4). In particular, $\mathbb{P}\{\mathbf{V} > x\}$ and $P_{\{1,\ldots,n\}}(x)$ are regularly varying of index $-N^*(\nu - 1)$.

E. K heterogeneous classes

We finally consider the important special case where each On-Off flow in $\mathcal{I}_2$ belongs to one of K heterogeneous classes. We will show how an approximate solution to the knapsack problem may be obtained using a simple index rule. The approximation is in fact asymptotically exact in the many-sources regime.

Specifically, consider the superposition of n On-Off flows, each belonging to one of K heterogeneous classes. Let a_k be the fraction of flows of class $k \in \{1, \ldots, K\}$, with peak rate r_k, mean rate ρ_k, and an On-period distribution which is regularly varying of index $-\nu_k$. Let the service rate be n (instead of 1), and let $\mathbf{V}^{(n)}$ be the stationary workload. The knapsack problem then takes the form

$$\begin{aligned} \min_{n_k\in\{0,\ldots,na_k\}} \quad & \sum_{k=1}^{K} n_k(\nu_k - 1) \\ \text{sub} \quad & \sum_{k=1}^{K} n_k r_k + \sum_{k=1}^{K}(na_k - n_k)\rho_k > n. \end{aligned}$$

Unfortunately, the above problem cannot be easily solved due to the integrality constraints. Intuitively however, one may expect that as n grows large, the integrality constraints should have a negligible effect, so that a continuous relaxation with $n_k \in [0, na_k]$ should give a good approximate solution.

This relaxation may be solved using a simple index rule. Index the K classes in non-decreasing order of the ratios

$$\gamma_k := (\nu_k - 1)/(r_k - \rho_k).$$

For any $k \in \{1, \ldots, K\}$, define $\sigma_k := \sum_{m=1}^{k-1} a_m r_m + \sum_{m=k}^{K} a_m \rho_m$. Determine the (unique) index ℓ such that $1 \in (\sigma_{\ell-1}, \sigma_\ell]$. Then take $n_k^* = na_k$ for all classes $k < \ell$, $n_k^* = 0$ for all classes $k > \ell$, and $n_\ell^* = n(1 - \sigma_{\ell-1})/(r_\ell - \rho_\ell)$.

This yields the (crude) approximation

$$\mathbb{P}\{\mathbf{V}^{(n)} > x\} \approx x^{-n\mu}, \tag{3.5}$$

with $\mu := \sum_{k=1}^{\ell-1} a_k(\nu_k - 1) + (1 - \sigma_{\ell-1})\gamma_\ell$. In [35] we prove that the above approximation is logarithmically exact in the many-sources regime. In particular, one may show that the limits for $x \to \infty$ and $n \to \infty$ commute if one considers logarithmic asymptotics.

Theorem III.3: (Robustness of logarithmic asymptotics)

$$\lim_{n\to\infty}\lim_{x\to\infty} \frac{1}{n}\frac{\log\mathbb{P}\{\mathbf{V}^{(n)} > nx\}}{\log x} = \lim_{x\to\infty}\lim_{n\to\infty} \frac{1}{n}\frac{\log\mathbb{P}\{\mathbf{V}^{(n)} > nx\}}{\log x}.$$

The proof of the above theorem can be found in [35]. Although logarithmically exact, the approximation (3.5) may not be appropriate from a practical perspective. In particular, it is shown in [35] that an analogue of Theorem III.3 cannot hold if one considers exact asymptotics.

IV. Preliminary results

In this section we collect some preliminary results which will be used in later sections. We first give a convenient representation for the stationary distribution of the workload $\mathbf{V}_E^c$. Starting point is the definition $V_E^c(t) := \sup_{0\leq s\leq t}\{A_E(s,t) - c(t-s)\}$ (assuming $V_E^c(0) = 0$). Since the process $A_E(0,t)$ has stationary and reversible increments (see [35] for a detailed description of$A_E(0,t)$), we have

$$\sup_{0\leq s\leq t}\{A_E(s,t) - c(t-s)\} \stackrel{d}{=} \sup_{0\leq s\leq t}\{A_E(0,s) - cs\}.$$

In the sequel we will use the latter expression as the *definition* of $V_E^c(t)$. Accordingly, for $c > \rho_E$, the stationary workload as $t \to \infty$ may be represented as

$$\mathbf{V}_E^c := \sup_{t\geq 0}\{A_E(0,t) - ct\}. \tag{4.1}$$

We now derive some simple bounds for the workload distribution $\mathbb{P}\{\mathbf{V}_S^c > x\}$ for subsets $S \subseteq \mathcal{I}_2$. For any subset $S \subseteq \mathcal{I}_2$, $c < r_S$, define

$$P_S^c(x) := \prod_{j\in S} p_j \mathbb{P}\{\mathbf{A}_j^r > \frac{x}{r_S - c}\}.$$

The first result may also be found in Choudhury & Whitt [10].

Lemma IV.1: For $S \subseteq \mathcal{I}_2$, $c < r_S$,

$$\mathbb{P}\{\mathbf{V}_S^c > x\} \geq P_S^c(x).$$

For any subset $S \subseteq \mathcal{I}_2$, $c < r_S$, define

$$K_S^c := \prod_{j\in S} \frac{r_j - \rho_j}{r_j - \rho_j + c - r_S}.$$

Lemma IV.2: Let $S \subseteq \mathcal{I}_2$. If $c \in (r_S - \min_{j\in S}\{r_j - \rho_j\}, r_S)$, and $A_j^r(\cdot) \in \mathcal{S}$ for all $j \in S$, then

$$\mathbb{P}\{\mathbf{V}_S^c > x\} \lesssim K_S^c P_S^c(x).$$

Proof: For any $i \in S$, denote $d_i := c - r_S + r_i$. Note that $d_i > \rho_i$ since $c > r_S - (r_i - \rho_i)$. Then, sample-path wise, $V_S^c(t) \leq V_i^{d_i}(t)$ for all $i \in S$. Theorem II.1 then yields,

$$\mathbb{P}\{\mathbf{V}_S^c > x\} \leq \mathbb{P}\{\mathbf{V}_j^{d_j} > x \text{ for all } j \in S\} \sim K_S^c P_S^c(x).$$

■

We now derive some general bounds for the tail of the total workload distribution $\mathbb{P}\{\mathbf{V} > x\}$. For any $c \geq 0$, $E \subseteq \mathcal{I}$, define $Z_E^c(t) := \sup_{0\leq s\leq t}\{cs - A_E(0,s)\}$. For $c < \rho_E$, let $\mathbf{Z}_E^c$ be a random variable with the limiting distribution of $Z_E^c(t)$ for $t \to \infty$.

We first present a lower bound. The idea behind its derivation as follows: $\mathbf{V}_E^{c_E}$ being large for some minimally-critical set $E \in \Lambda$ basically implies that $\mathbf{V}$ must be large too, unless the other flows $j \notin E$ persist in below-average behavior. Excluding such below-average behavior (reflected in large values of $\mathbf{Z}_{\mathcal{I}\setminus E}^c$) from the event $\{\mathbf{V} > x\}$ yields the following lower bound for $\mathbb{P}\{\mathbf{V} > x\}$.

Lemma IV.3: For any $E \subseteq \mathcal{I}_2$, $\delta > 0$, and $y \geq 0$,

$$\mathbb{P}\{\mathbf{V} > x\} \geq \mathbb{P}\{\mathbf{V}_E^{c_E+\delta} > x + y\}\mathbb{P}\{\mathbf{Z}_{\mathcal{I}\setminus E}^{\rho_{\mathcal{I}\setminus E}-\delta} \leq y\}.$$

Proof: Sample-path wise, using properties of the sup-operator,

$$V(t) \geq V_E^{c_E+\delta}(t) - Z_{\mathcal{I}\setminus E}^{\rho_{\mathcal{I}\setminus E}-\delta}(t)$$

for any $E \subseteq \mathcal{I}_2$. Next, let $t \to \infty$ to obtain the corresponding lower bound in the stationary regime. ■

Denote by $\mathcal{N} := |\mathcal{I}|$ the total number of flows, and let $\Omega \subseteq 2^{\mathcal{I}_2}$ be the collection of all minimally-critical sets.

We now provide a corresponding upper bound, which is somewhat more involved. The idea is as follows: $\mathbf{V}$ being large essentially means that $\mathbf{V}_E^{c_E}$ must be large for some minimally-critical set $E \in \Lambda$ too, unless the other flows $j \notin E$ exhibit above-average behavior. Extending the event $\{\mathbf{V} > x\}$ with possible above-average behavior of the flows $j \notin E$ (manifesting itself in large values of $\mathbf{V}_{\mathcal{I}\setminus E}^{\rho_{\mathcal{I}\setminus E}+\delta}$) leads to the following upper bound for $\mathbb{P}\{\mathbf{V} > x\}$.

Lemma IV.4: Let $E \in \Omega$. Then for any $\delta, \epsilon > 0$ sufficiently small and y,

$$\begin{aligned}\mathbb{P}\{\mathbf{V} > x\} &\leq \mathbb{P}\{\mathbf{V}_E^{c_E-\delta} > x - y\} + \mathbb{P}\{\mathbf{V}_{\mathcal{I}_1}^{\rho_{\mathcal{I}_1}+\epsilon} > x/\mathcal{N}\} \\ &+ \mathbb{P}\{\mathbf{V}_{\mathcal{I}\setminus E}^{\rho_{\mathcal{I}\setminus E}+\delta} > y\}\prod_{j\in E}\mathbb{P}\{\mathbf{V}_j^{\rho_j+\epsilon} > x/\mathcal{N}\} \\ &+ \sum_{E\in\Omega\setminus\Lambda}\prod_{j\in S}\mathbb{P}\{\mathbf{V}_j^{\rho_j+\epsilon} > x/\mathcal{N}\}.\end{aligned}$$

Proof: Sample-path wise,

$$V(t) \leq V_E^{c_E-\delta}(t) + V_{\mathcal{I}\setminus E}^{\rho_{\mathcal{I}\setminus E}+\delta}(t)$$

for any $E \subseteq \mathcal{I}_2$.

In addition, for $\epsilon > 0$ sufficiently small, $V(t) > x$ implies $V_{\mathcal{I}_1}^{\rho_{\mathcal{I}_1}+\epsilon}(t) > x/\mathcal{N}$, or there exists a minimally-critical set $S \in \Omega$ such that $V_j^{\rho_j+\epsilon}(t) > x/\mathcal{N}$ for all $j \in S$, see [35] for details.

This yields, for any $\delta, \epsilon > 0$ sufficiently small and y,

$$\begin{aligned}&\mathbb{P}\{\mathbf{V} > x\} \\ \leq\ &\Pr\{\mathbf{V}_E^{c_E-\delta} + \mathbf{V}_{\mathcal{I}\setminus E}^{\rho_{\mathcal{I}\setminus E}+\delta} > x, \\ &\exists S \in \Omega : \mathbf{V}_{\mathcal{I}_1}^{\rho_{\mathcal{I}_1}+\epsilon} > x/\mathcal{N} \text{ or } \mathbf{V}_j^{\rho_j+\epsilon} > x/\mathcal{N}\ \ \forall j \in S\} \\ \leq\ &\Pr\{\mathbf{V}_E^{c_E-\delta} > x - y \text{ or } \mathbf{V}_{\mathcal{I}\setminus E}^{\rho_{\mathcal{I}\setminus E}+\delta} > y, \\ &\exists S \in \Omega : \mathbf{V}_{\mathcal{I}_1}^{\rho_{\mathcal{I}_1}+\epsilon} > x/\mathcal{N} \text{ or } \mathbf{V}_j^{\rho_j+\epsilon} > x/\mathcal{N}\ \ \forall j \in S\} \\ \leq\ &\mathbb{P}\{\mathbf{V}_E^{c_E-\delta} > x - y\} + \mathbb{P}\{\mathbf{V}_{\mathcal{I}_1}^{\rho_{\mathcal{I}_1}+\epsilon} > x/\mathcal{N}\} \\ +\ &\sum_{S\in\Omega}\mathbb{P}\{\mathbf{V}_j^{\rho_j+\epsilon} > x/\mathcal{N}\ \ \forall j \in S, \mathbf{V}_{\mathcal{I}\setminus E}^{\rho_{\mathcal{I}\setminus E}+\delta} > y\},\end{aligned}$$

which immediately gives the desired result. ■

Lemma IV.5: Let $S \subseteq \mathcal{I}_2$. If $A_j(\cdot) \in \mathcal{R}$ for all $j \in S$ and $c \in (r_S - \min_{j\in S}\{r_j - \rho_j\}, r_S)$, then

$$\lim_{M\to\infty} \limsup_{x\to\infty} \frac{\mathbb{P}\{\sup_{t\geq Mx}\{A_S(0,t) - (c-\epsilon)t\} > x\}}{\mathbb{P}\{\mathbf{V}_S^c > x\}} = 0,$$

for any $\epsilon \in [0, r_S - c)$.

Proof: See [35]. ■

V. Reduced-load equivalence

In this section we give a proof of Theorem III.1. For a proof of Theorem III.2 and other extensions (such as the case with additional heavy-tailed instantaneous input) we refer to [35].

The proofs of the complementing results for the reduced system are presented in Section VI.

Theorem V.1: (Reduced-load equivalence)
Suppose $S^* \in \Omega$ satisfies Assumptions V.1-V.5 as listed below with $c = c_{S^*}$. Then

$$\mathbb{P}\{\mathbf{V} > x\} \sim \mathbb{P}\{\mathbf{V}_{S^*}^{c_{S^*}} > x\}.$$

Assumption V.1: For any y and $\delta > 0$,

$$F_S^c(\delta) := \liminf_{x\to\infty} \frac{\mathbb{P}\{\mathbf{V}_S^{c+\delta} > x + y\}}{\mathbb{P}\{\mathbf{V}_S^c > x\}},$$

is independent of y. In addition, $\lim_{\delta\downarrow 0} F_S^c(\delta) = 1$.

Assumption V.2: For any y and $\delta > 0$,

$$G_S^c(\delta) := \limsup_{x\to\infty} \frac{\mathbb{P}\{\mathbf{V}_S^{c-\delta} > x - y\}}{\mathbb{P}\{\mathbf{V}_S^c > x\}},$$

is independent of y. In addition, $\lim_{\delta\downarrow 0} G_S^c(\delta) = 1$.

Assumption V.3: For any $\epsilon > 0$,

$$\lim_{x\to\infty} \frac{\mathbb{P}\{\mathbf{V}_{\mathcal{I}_1}^{\rho_{\mathcal{I}_1}+\epsilon} > x/\mathcal{N}\}}{\mathbb{P}\{\mathbf{V}_S^c > x\}} = 0.$$

Assumption V.4: For any $\epsilon > 0$,

$$H_S^c(\epsilon) := \limsup_{x\to\infty} \frac{\prod_{j\in S} \mathbb{P}\{\mathbf{V}_j^{\rho_j+\epsilon} > x/\mathcal{N}\}}{\mathbb{P}\{\mathbf{V}_S^c > x\}} < \infty.$$

Assumption V.5: For any $E \in \Omega$, $E \neq S$, for any $\epsilon > 0$,

$$\lim_{x\to\infty} \frac{\prod_{j\in E} \mathbb{P}\{\mathbf{V}_j^{\rho_j+\epsilon} > x/\mathcal{N}\}}{\mathbb{P}\{\mathbf{V}_S^c > x\}} = 0.$$

Proof: The proof consists of a lower bound and an upper bound which asymptotically coincide.

(Lower bound) Combining Lemma IV.3 (take $E = S^*$) with Assumption V.1 yields, for any $\delta > 0$ and y,

$$\liminf_{x\to\infty} \frac{\mathbb{P}\{\mathbf{V} > x\}}{\mathbb{P}\{\mathbf{V}_{S^*}^{c_{S^*}} > x\}} \geq F_{S^*}^{c_{S^*}}(\delta)\mathbb{P}\{\mathbf{Z}_{\mathcal{I}\setminus S^*}^{\rho_{\mathcal{I}\setminus S^*}-\delta} \leq y\}.$$

Letting first $y \to \infty$, and then $\delta \downarrow 0$ completes the proof of the lower bound.

(Upper bound) Combining Lemma IV.4 (take $E = S^*$) with Assumptions V.2-V.5, we obtain for any $\delta, \epsilon > 0$ sufficiently small and y,

$$\begin{aligned} &\limsup_{x\to\infty} \frac{\mathbb{P}\{\mathbf{V} > x\}}{\mathbb{P}\{\mathbf{V}_{S^*}^{c_{S^*}} > x\}} \\ \leq\ & G_{S^*}^{c_{S^*}}(\delta) + H_{S^*}^{c_{S^*}}(\epsilon)\mathbb{P}\{\mathbf{V}_{\mathcal{I}\setminus S^*}^{\rho_{\mathcal{I}\setminus S^*}+\delta} > y\}. \end{aligned}$$

Letting $y \to \infty$, then $\delta \downarrow 0$ completes the proof. ■

In order to complete the proof of the reduced-load equivalence result (3.1), it remains to be shown that a dominant set $S^* \subseteq \mathcal{I}_2$ with $A_j(\cdot) \in \mathcal{R}$ for all $j \in S^*$ satisfies Assumptions V.1-V.5. That is done in the following two propositions for $S = S^*$.

Proposition V.1: Let $S \subseteq \mathcal{I}_2$. If $A_j(\cdot) \in \mathcal{R}$ for all $j \in S$, then Assumptions V.1 and V.2 are satisfied for any $c \in (r_S - \min_{j\in S}\{r_j - \rho_j\}, r_S)$.

Proof: We first prove Assumption V.2. It follows from Theorem VI.3 that if $A_j(\cdot) \in \mathcal{R}$ for all $j \in S$, then $\mathbb{P}\{\mathbf{V}_S^c > x\} \in \mathcal{IR}$. As $\mathcal{R} \subseteq \mathcal{L}$, it thus suffices to prove the property for $y = 0$. Let $\epsilon \in [0, r_S - c)$, and let $\delta \in (0, \epsilon]$. Then

$$\begin{aligned} &\limsup_{x\to\infty} \frac{\mathbb{P}\{\mathbf{V}_S^{c-\delta} > x\}}{\mathbb{P}\{\mathbf{V}_S^c > x\}} \\ \leq\ & \limsup_{x\to\infty} \frac{\mathbb{P}\{\mathbf{V}_S^c > (1-\delta^{1/2})x\}}{\mathbb{P}\{\mathbf{V}_S^c > x\}} \\ +\ & \limsup_{x\to\infty} \frac{\mathbb{P}\{\sup_{t\geq x\delta^{-1/2}}\{A_S(0,t) - (c-\epsilon)t\} > x\}}{\mathbb{P}\{\sup_{t\geq 0}\{A_S(0,t) - ct\} > x\}}. \end{aligned}$$

The fact that $\mathbb{P}\{\mathbf{V}_S^c > x\} \in \mathcal{IR}$ implies that the first term tends to 1 as $\delta \downarrow 0$, while Lemma IV.5 (with $M = \delta^{-1/2}$) shows that the second term then goes to 0.

The proof of Assumption V.1 is exactly the same and therefore omitted. ■

Proposition V.2: Let $S \subseteq \mathcal{I}_2$. If $A_j(\cdot) \in \mathcal{R}$ for all $j \in S$, then Assumptions V.3 and V.4 are satisfied for any $c > \rho_S$. If in addition S is a dominant set, then Assumption V.5 is satisfied as well.

Proof: Assumption V.3 is satisfied by Lemma IV.1, Assumption (II.1) and the assumption that $A_j(\cdot) \in \mathcal{R}$ for all $j \in S$. Assumptions V.4 and V.5 follow from Theorem II.1. ■

VI. Tail asymptotics for the reduced system

In this section we derive the tail asymptotics for the reduced system. In particular, we give a proof of Equations (3.2) and (3.4).

For notational convenience, let c be the capacity of the reduced system, let the set of flows be indexed as $\mathcal{J} = \{1, \ldots, N\}$, and denote $r := r_{\mathcal{J}}$ and $A(0,t) := A_{\mathcal{J}}(0,t)$. By definition, the reduced system satisfies the following two properties:

(i) The On-period distribution of flow i is regularly varying of index $-\nu_i < -1$, i.e., $A_i(\cdot) \in \mathcal{R}_{-\nu_i}$;

(ii) All flows must be On for the drift of the workload process to be positive, i.e., $c \in (r - \min_{i=1,\ldots,N}\{r_i - \rho_i\}, r)$.

We now state our main theorem.

Theorem VI.1: Consider a queue of capacity c fed by N On-Off flows. If $c \in (r - \min_{i=1,\ldots,N}\{r_i - \rho_i\}, r)$ with $r = \sum_{i=1}^N r_i$, and $A_j(\cdot) \in \mathcal{R}$ for all $j = 1, \ldots, N$, then

$$\mathbb{P}\{\mathbf{V}^c > x\} \sim \prod_{j=1}^N p_j \sum_{\mathcal{J}_0 \subseteq \{1,\ldots,N\}} P_{\mathcal{J}_0}(x),$$

where $P_{\mathcal{J}_0}(x)$ is given by (3.3).

An asymptotic characterization of $P_{\mathcal{J}_0}(x)$ which may be useful for further analysis is provided in Subsection VI-D. This characterization also shows that $\mathbb{P}\{\mathbf{V}^c > x\}$ and $P_{\mathcal{J}_0}(x)$ are regularly varying, and gives an expression for the pre-factor in the asymptotic expansion of $\mathbb{P}\{\mathbf{V}^c > x\}$.

The remainder of this section is organized as follows. Detailed heuristic arguments are given in Section VI-A. In Section VI-B, we give some preliminary results on the most probable behavior of the process $\{A(0,t) - ct\}$. The proof of Theorem VI.1 is then completed in Section VI-C. Section VI-D deals with the asymptotic behavior of $P_{\mathcal{J}_0}(x)$.

A. Heuristic arguments

The proof of Theorem VI.1 is quite lengthy. Nevertheless, it is based on a simple intuitive argument: the most likely way for $\mathbf{V}^c \equiv \sup_{t\geq 0}\{A(0,t) - ct\}$ to reach a large value is that all flows have been simultaneously On for a long time. Specifically, each flow is likely to contribute through *exactly one* 'long' On-period; apart from these long On-periods, the flows show typical behavior.

The above heuristic argument may be used for computing $\sup_{t\geq 0}\{A(0,t) - ct\}$. Let's say that the long On-period of flow i begins at time s_i and ends at time $s_i + t_i$. Define

$$t^* := \min_{i=1,\ldots,N}\{s_i + t_i\},$$

as the time epoch at which the first of the long On-periods finishes. One may also interpret t^* as the time epoch at which the process $\{A(0,t) - ct\}$ reaches its largest value. Note that $A_i(0,s_i) \approx \rho_i s_i$, $A_i(s_i, s_i + t_i) = r_i t_i$, and $A_i(s_i + t_i, s_i + t_i + t) \approx \rho_i t$, $t \geq 0$. One thus obtains, using the fact that $c \in (r - \min_{i=1,\ldots,N}\{r_i - \rho_i\}, r)$),

$$\begin{aligned}\sup_{t\geq 0}\{A(0,t) - ct\} &\approx A(0,t^*) - ct^* \\ &\approx \sum_{i=1}^{N}[\rho_i s_i + r_i(t^* - s_i)] - ct^* \\ &= \sum_{i=1}^{N}(\rho_i - r_i)s_i + (r-c)t^*. \quad (6.1)\end{aligned}$$

The problem is thus reduced to calculating

$$\mathbb{P}\{\sum_{i=1}^{N}(\rho_i - r_i)s_i + (r-c)\min_{i=1,\ldots,N}\{s_i + t_i\} > x\}. \quad (6.2)$$

Although the proof is based on the representation $\mathbf{V}^c \equiv \sup_{t\geq 0}\{A(0,t) - ct\}$, it is useful to keep the original workload process $\sup_{0\leq s\leq t}\{A(s,t) - c(t-s)\}$ in mind as well. Figure 1 shows a typical scenario leading to a large workload level (so small fluctuations are ignored) in the case of two On-Off flows.

At a certain time ω_0, the first long On-period begins. Before that time, both flows show average behavior. The queue starts to build (at rate $r_1 + r_2 - c$) at time ω_1 when the second long On-period begins, and reaches its largest level at time ω_3. Level x is crossed at time ω_2.

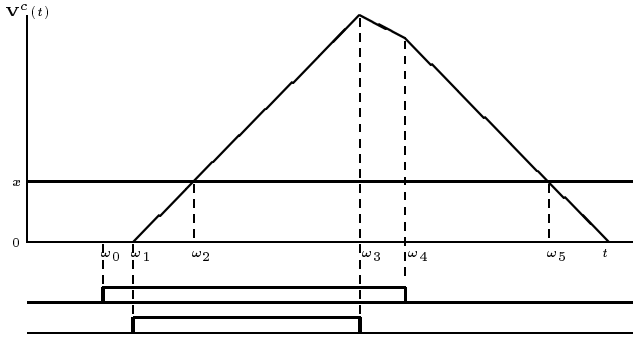

Fig. 1. Typical overflow scenario for two On-Off flows

Between times ω_3 and ω_4, the queue drains at rate $c - r_1 - \rho_2$: flow 1 is still in the middle of its long On-period, and flow 2 shows average behavior (remember small fluctuations are neglected). The process is still above level x between times ω_4 and ω_5. However, here both flows show average behavior again, causing a negative drift $c - \rho_1 - \rho_2$.

The figure illustrates why the analysis of the reduced system is still quite complicated:

- Although the long On-periods must significantly overlap, the difference between the finishing times of these On periods can be quite large (of order x, hence not negligible);
- Given that the observed workload is larger than x, it is not necessarily the case that all flows are in the middle of their long On-periods. In Figure 1, this is only the case in the time interval (ω_2, ω_3). In fact, for any given flow, its long On-period may have finished a long time ago. Consequently, there are 2^N different possibilities (corresponding to which subset of the flows are still in the middle of their long On-periods). Sample-path wise, there are $N+1$ different time intervals in which the workload may be larger than x (depending on how many of the flows are still in the middle of their long On-periods);
- Specifically, given that the observed workload is larger than x, it may still have been even larger at an earlier time epoch. In Figure 1, this is the case in the time intervals (ω_3, ω_4) and (ω_4, ω_5).

These complications do not arise if one considers a related problem, which concerns the overflow probability in a fluid queue with a *finite buffer* of size x. As is shown in a recent paper of Jelenković & Momčilović [18], the analysis of the reduced system is then considerably simpler. It suffices to use bounds which are similar to Lemma IV.1 and Lemma IV.2, and to combine these with the asymptotic results for a single On-Off flow in Jelenković [17] and Zwart [34]. See also [22] for related issues in the fluid queue with $M/G\infty$ input.

B. Characterization of most probable behavior

In this subsection we prove some preliminary results characterizing the most probable behavior of the process $\{A(0,t) - ct\}$ given that it reaches a large value. In particular, we formalize the following two heuristic statements, resulting in a formal version of Equation (6.1).

(i) Each flow contributes to $\sup_{t\geq 0}\{A(0,t) - ct\}$ through exactly one 'long' On-period;

(ii) Apart from these long On-periods, the flows show typical behavior.

An On-period is referred to as 'long' when larger than ϵx, with ϵ some small, but positive constant. In order to formalize the above statements, we need to keep track how many of such long On-periods occur.

With that in mind, we define $\mathcal{N}_i(A,B)$, for intervals $A, B \subseteq [0,\infty)$, as the number of On-periods of flow i of which the length is contained in A and which overlap (in time) with B. For compactness, denote $\mathcal{N}_i(u,t) \equiv \mathcal{N}_i((u,\infty),[0,t])$.

We now proceed with a few preparatory lemmas.

First we show how to obtain an upper bound for the workload process in terms of a simple random walk. As in the proof of Lemma IV.2, we have $V^c(t) \leq V_i^{d_i}(t)$ for all $i = 1,\ldots,N$, with $d_i := c - r_{\mathcal{I}\setminus\{i\}} = c - r + r_i$. Recall that $V_i^{d_i}(t) \stackrel{d}{=} \sup_{0\leq s\leq t}\{A_i(0,s) - d_i s\}$. Now let, for fixed i, $\mathbf{S}_{in} = \mathbf{X}_{i1} + \ldots + \mathbf{X}_{in}$ be a random walk with step sizes $\mathbf{X}_{im} = (r_i - d_i)\mathbf{A}_{im} - d_i\mathbf{U}_{im}$, with $\mathbf{A}_{im}$ and $\mathbf{U}_{im}$ i.i.d. random variables distributed as the On- and Off-periods of flow i, respectively.

Since $c \in (r - \min_{i=1,\ldots,N}\{r_i - \rho_i\}, r)$, we have $\rho_i < d_i$ for all $i = 1,\ldots,N$, so that $\mathbb{E}\{\mathbf{X}_{i1}\} < 0$, i.e., the random walk has negative drift. Because of the saw-tooth nature of the process $A_i(0,s) - d_i s$, we have

$$\sup_{0\leq s\leq t}\{A_i(0,s) - d_i s\} \leq (r_i - d_i)\mathbf{A}_{i0}^r + \sup_{n\leq N_i^A(t)} \mathbf{S}_{in},$$

with $N_i^A(t)$ denoting the number of Off-periods of flow i elapsed during $[0,t]$ which started after time 0. The above observations are summarized in the following auxiliary lemma.

Lemma VI.1: For all $\epsilon > 0$, t and x,

$$\begin{aligned} & \mathbb{P}\{V^c(t) > x, \mathcal{N}_i(\epsilon x, t) = 0\} \\ \leq\ & \mathbb{P}\{\sup_{n\leq N_i^A(t)} \mathbf{S}_{in} > x(1-\epsilon(r_i-d_i)), \mathcal{N}_i(\epsilon x, t) = 0\}. \end{aligned}$$

Proof: See [35]. ■

To obtain upper bounds for probabilities as in Lemma VI.1, we will frequently apply the following key lemma, which is a trivial modification of Lemma 3 in [28].

Lemma VI.2: Let $\mathbf{S}_n = \mathbf{X}_1 + \ldots + \mathbf{X}_n$ be a random walk with i.i.d. step sizes such that $\mathbb{E}\{\mathbf{X}_1\} < 0$ and $\mathbb{E}\{(\mathbf{X}_1^+)^p\} < \infty$ for some $p > 1$. Then, for any $\beta < \infty$, there exists an $\epsilon^* > 0$ and a function $\phi(\cdot) \in \mathcal{R}_{-\beta}$ such that for $\epsilon \in (0,\epsilon^*]$

$$\mathbb{P}\{\mathbf{S}_n > x | \mathbf{X}_j \leq \epsilon x, j = 1,\ldots,n\} \leq \phi(x),$$

for all n and all x.

Note that if $\mathbf{X}_j$ can be represented as the difference of two non-negative independent random variables $\mathbf{X}_j^1$ and $\mathbf{X}_j^2$, then the lemma remains valid if the $\mathbf{X}_j$'s are replaced by $\mathbf{X}_j^1$.

The final preparatory lemma is a simple consequence of Lemma IV.1, and will be used several times in combination with Lemma VI.2 to show that probabilities of certain events are of $o(\mathbb{P}\{\mathbf{V}^c > x\})$. Define $P(x) := \prod_{j=1}^N \mathbb{P}\{\mathbf{A}_j^r > x\} \in \mathcal{R}_{-\mu}$, $\mu := \sum_{j=1}^N (\nu_j - 1)$.

Lemma VI.3: $\limsup_{x\to\infty} \frac{P(x)}{\mathbb{P}\{\mathbf{V}^c > x\}} < \infty$.

We now show that, with overwhelming probability (as $x \to \infty$), the rare event $\{\mathbf{V}^c > x\}$ occurs as follows.
(i) The process $\{A(0,t) - ct\}$ reaches level x before time Mx for some large M;
(ii) Up to time Mx, each flow generates *exactly one* long On-period, i.e., $\mathcal{N}_i(\epsilon x, Mx) = 1$ for $i = 1,\ldots,N$.

Proposition VI.1: $\lim_{M\to\infty} \liminf_{x\to\infty} \frac{\mathbb{P}\{V^c(Mx) > x\}}{\mathbb{P}\{\mathbf{V}^c > x\}} = 1$.

Proof: Follows from Lemma IV.5. ■

Now suppose that the workload reaches level x. By the previous proposition, we may assume that this occurs before time Mx (for M sufficiently large). The next two propositions show that we may restrict the attention to a scenario where each flow initiates *exactly one* long On-period before time Mx.

The first proposition indicates that each flow has *at least* one long On-period.

Proposition VI.2: For all i, there exists an $\epsilon^* > 0$ such that for all $\epsilon \in (0,\epsilon^*]$ and all M,

$$\mathbb{P}\{V^c(Mx) > x, \mathcal{N}_i(\epsilon x, Mx) = 0\} = o(\mathbb{P}\{\mathbf{V}^c > x\}).$$

Proof: Define $N_i^U(t) := \max\{n : \sum_{j=1}^n \mathbf{U}_{ij} \leq t\} + 1$. Note that $N_i^A(t) \leq N_i^U(t)$. Using Lemma VI.1, taking $t = Mx$,

$$\begin{aligned} & \mathbb{P}\{V^c(Mx) > x, \mathcal{N}_i(\epsilon x, Mx) = 0\} \\ \leq\ & \mathbb{P}\{\sup_{n\leq N_i^A(Mx)} \mathbf{S}_n > x(1-\epsilon(r_i-d_i)), \mathcal{N}_i(\epsilon x, Mx) = 0\} \\ \leq\ & \mathbb{P}\{\sup_{n\leq N_i^A(Mx)} \mathbf{S}_n > x(1-\epsilon(r_i-d_i)) | \mathcal{N}_i(\epsilon x, Mx) = 0\} \\ \leq\ & \mathbb{P}\{\sup_{n\leq N_i^U(Mx)} \mathbf{S}_n > x(1-\epsilon(r_i-d_i)) | \mathbf{A}_{ij} < \epsilon x, j \geq 1\} \\ \leq\ & \mathbb{P}\{\sup_{n\leq M_2 x} \mathbf{S}_n > x(1-\epsilon(r_i-d_i)) | \mathbf{A}_{ij} < \epsilon x, j \geq 1\} \\ +\ & \mathbb{P}\{N_i^U(Mx) > M_2 x\}. \end{aligned}$$

The second term decays exponentially fast in x if $M_2 > \lambda M$. The first term can be bounded by

$$\sum_{m=1}^{M_2 x} \mathbb{P}\{\mathbf{S}_m > x(1-\epsilon(r_i-d_i)) | \mathbf{A}_{ij} \leq \epsilon x, j = 1,\ldots,m\}.$$

According to Lemma VI.2, there exists an $\epsilon^* > 0$ and a function $\phi(\cdot) \in \mathcal{R}_{-\beta}$ with $\beta > \mu + 1$, such that for $\epsilon \in (0,\epsilon^*]$ the last quantity is upper bounded by $M_2 x\phi(x)$. The latter function is regularly varying of index $1 - \beta < -\mu$. Invoking Lemma VI.3 then completes the proof. ■

The next proposition shows that each flow has *at most* one long On-period.

Proposition VI.3: For all i, all M and all $\epsilon > 0$,

$$\mathbb{P}\{V^c(Mx) > x, \mathcal{N}_i(\epsilon x, Mx) \geq 2\} = o(\mathbb{P}\{\mathbf{V}^c > x\}).$$

Proof: Without loss of generality we may take $i = 1$. By Proposition VI.2 it suffices to consider

$$\mathbb{P}\{\mathcal{N}_1(\epsilon x, Mx) \geq 2\} \prod_{i=2}^N \mathbb{P}\{\mathcal{N}_i(\epsilon x, Mx) \geq 1\}.$$

Invoking Lemma VI.3 it suffices to show that: (i) $\mathbb{P}\{\mathcal{N}_i(\epsilon x, Mx) \geq 1\}$ is bounded by a function which is regularly varying of index $1-\nu_i$; (ii) $\mathbb{P}\{\mathcal{N}_i(\epsilon x, Mx) \geq 2\} = o(\mathbb{P}\{\mathcal{N}_i(\epsilon x, Mx) \geq 1\})$. The proof of these statements is straightforward, see [35]. ■

We have now shown that, with overwhelming probability, each flow contributes to a large value of $\sup_{t\geq 0}\{A(0,t)-ct\}$ through exactly one long On-period. We thus proceed with the second statement (as indicated at the beginning of this subsection), implying that apart from these long On-periods, the flows show typical behavior. In order to formalize that statement, we need to introduce some notation. Define

$$\tau(y) := \inf\{t \geq 0 : A(0,t) - ct = y\}$$

as the first time at which the process $\{A(0,t)-ct\}$ reaches level y.

For fixed $\epsilon > 0$ and x, let $\tau_{s,i}(\epsilon x)$ and $\tau_{f,i}(\epsilon x)$ be the respective starting and finishing times of the first On-period of flow i exceeding length ϵx. Denote $\tau_s(\epsilon x) := \max_{i=1,\ldots,N} \tau_{s,i}(\epsilon x)$ and $\tau_f(\epsilon x) := \min_{i=1,\ldots,N} \tau_{f,i}(\epsilon x)$.

Note that all flows are in the middle of their long On-periods between times $\tau_s(\epsilon x)$ and $\tau_f(\epsilon x)$. We will show that the fluctuations of the process $\{A(0,t)-ct\}$ away from the mean before time $\tau_s(\epsilon x)$ and after time $\tau_f(\epsilon x)$ can be neglected.

A formal statement is made in the next two propositions (for a proof, see [35]). The first proposition indicates that it is most unlikely that the process $\{A(0,t)-ct\}$ reaches level δx before time $\tau_s(\epsilon x)$.

Proposition VI.4: For any $\delta > 0$, there exists an $\epsilon^* > 0$ such that for all $\epsilon \in (0, \epsilon^*]$,

$$\mathbb{P}\{\tau(\delta x) < \tau_s(\epsilon x)\} = o(\mathbb{P}\{\mathbf{V}^c > x\}).$$

The next proposition shows that, given that the process $\{A(0,t)-ct\}$ reaches level x before time Mx, most probably level $(1-\delta)x$ is crossed before time $\tau_f(\epsilon x)$.

Proposition VI.5: For any $\delta > 0$, there exists an $\epsilon^* > 0$ such that for all $\epsilon \in (0, \epsilon^*)$ and $M < \infty$,

$$\mathbb{P}\{\tau((1-\delta)x) > \tau_f(\epsilon x), V^c(Mx) > x\} = o(\mathbb{P}\{\mathbf{V}^c > x\}).$$

C. Proof of Theorem VI.1

In this subsection we give a sketch of the proof of Theorem VI.1. From the previous subsection we obtain, using Propositions VI.1, VI.4 and VI.5,

Theorem VI.2: For any $\delta > 0$, there exists an $\epsilon^* > 0$ such that for all $\epsilon \in (0, \epsilon^*)$,

$$\begin{aligned}\mathbb{P}\{\mathbf{V}^c > x\} &\geq \mathbb{P}\{A(0,\tau_f(\epsilon x)) - c\tau_f(\epsilon x) > x\}\\ \mathbb{P}\{\mathbf{V}^c > x\} &\lesssim \mathbb{P}\{A(0,\tau_f(\epsilon x)) - c\tau_f(\epsilon x) > (1-\delta)x\}.\end{aligned}$$

In order to obtain tight bounds for the probabilities in Theorem VI.2, we condition upon $\tau_{s,i}$ for all i. Hence, for any $\mathcal{J}_0 \subseteq \mathcal{J}$, define the event $D_{\mathcal{J}_0}(\epsilon x)$ by

$$D_{\mathcal{J}_0}(\epsilon x) := \{\tau_{s,i}(\epsilon x) = 0 \text{ iff } i \in \mathcal{J}_0\}.$$

The event $D_{\mathcal{J}_0}(\epsilon x)$ implies that the flows $i \in \mathcal{J}_0$ started their long On-period before time 0 (remember that we consider the system in stationarity). The flows $i \in \mathcal{J}_1$ start their long On-period at a later time epoch.

Denote $\mathbb{P}_{\mathcal{J}_0}\{\cdot\} = \mathbb{P}\{\cdot | D_{\mathcal{J}_0}(\epsilon x)\}$. The following two lemmas will be useful for providing tight upper and lower bounds for the probabilities in Theorem VI.2.

Lemma VI.4: (Lower bound) There exists an $\epsilon > 0$ such that

$$\begin{aligned}&\mathbb{P}_{\mathcal{J}_0}\{A(0,\tau_f(\epsilon x)) - c\tau_f(\epsilon x) > x\} \prod_{i\in\mathcal{J}_0} \mathbb{P}\{\mathbf{A}_i^r > \epsilon x\}\\ \gtrsim\ & P_{\mathcal{J}_0}(x) \prod_{i\in\mathcal{J}_1} p_i.\end{aligned}$$

Lemma VI.5: (Upper bound) For any $\delta > 0$, there exists an $\epsilon_\delta > 0$ such that for all $\epsilon \in (0, \epsilon_\delta)$

$$\begin{aligned}&\mathbb{P}_{\mathcal{J}_0}\{A(0,\tau_f(\epsilon x)) - c\tau_f(\epsilon x) > (1-\delta)x\} \prod_{i\in\mathcal{J}_0} \mathbb{P}\{\mathbf{A}_i^r > \epsilon x\}\\ \lesssim\ & P_{\mathcal{J}_0}((1-\delta)x) \prod_{i\in\mathcal{J}_1} p_i.\end{aligned}$$

Theorem VI.1 now follows by combining the above two lemmas with Theorem VI.2, see [35] for details.

We conclude with a brief sketch of the proof of Lemmas VI.4 and VI.5. The formal proofs are quite technical and can be found in [35].

Under the event $D_{\mathcal{J}_0}(\epsilon x)$, $A(0,\tau_f) - c\tau_f$ can be represented as

$$A(0,\tau_f) - c\tau_f = \min\{\min_{i\in\mathcal{J}_0} \mathbf{F}_i, \min_{i\in\mathcal{J}_1} \mathbf{G}_i\},$$

where $\mathcal{J}_1 = \mathcal{J} \setminus \mathcal{J}_0$. For a formal definition of the random variables $\mathbf{F}_i$ and $\mathbf{G}_i$ we refer to [35], where it is shown that $\mathbf{F}_i$ and $\mathbf{G}_i$ may be approximated as follows.

$$\begin{aligned}\mathbf{F}_i &\approx (r-c)\bar{\mathbf{A}}_i^r(\epsilon x) + \sum_{k\in\mathcal{J}_1} r_k \mathbb{E}\{\mathbf{U}_k\} N_k(\epsilon x),\\ \mathbf{G}_i &\approx (r-c)\bar{\mathbf{A}}_i(\epsilon x) + [(r-c)\mathbb{E}\{\mathbf{A}_i\} - d_i\mathbb{E}\{\mathbf{U}_i\}]N_i(\epsilon x)\\ &\quad - \sum_{k\in\mathcal{J}_1\setminus\{i\}} r_k \mathbb{E}\{\mathbf{U}_k\} N_k(\epsilon x).\end{aligned}$$

The only random variables appearing in the above expressions are $\bar{\mathbf{A}}_i(\epsilon x)$, $\mathbf{B}_i^r(\epsilon x)$, and $N_i(\epsilon x)$, of which the distributions are known. What thus remains is a lengthy, but straightforward computation.

D. Asymptotic behavior of $P_{\mathcal{J}_0}(x)$ and $\mathbb{P}\{\mathbf{V}^c > x\}$

In this subsection we give an asymptotic characterization of $P_{\mathcal{J}_0}(x)$, which may be useful for further analysis. In particular, we establish that $P_{\mathcal{J}_0}(x)$ and $\mathbb{P}\{\mathbf{V}^c > x\}$ are both regularly varying.

Define $g = \left(\frac{r_j-\rho_j}{r-c}\right)_{j\in\mathcal{J}_1}$, $e = (1,\ldots,1)$. Let $\mathbf{Z}_i, i \in \mathcal{J}$, be i.i.d. random variables with $\mathbb{P}\{\mathbf{Z}_i > y\} = (1+(r-c)y)^{-\nu_i}$, and define $\mathbf{Z}_{\mathcal{J}_1} = (\mathbf{Z}_j)_{j\in\mathcal{J}_1}$.

We have the following theorem (see [35] for a proof).

Theorem VI.3:

$$\begin{aligned}P_{\mathcal{J}_0}(x) &\sim \kappa_{\mathcal{J}_0} \prod_{i=1}^N \mathbb{P}\{\mathbf{A}_i^r > \frac{x}{r-c}\},\\ \mathbb{P}\{\mathbf{V}^c > x\} &\sim \kappa \prod_{i=1}^N p_i \mathbb{P}\{\mathbf{A}_i^r > \frac{x}{r-c}\},\end{aligned}$$

with $\kappa_{\mathcal{J}} = 1$, $\kappa_{\mathcal{J}_0} = \frac{1}{eg-1}\mathbb{P}\{\mathbf{Z}_i \geq \frac{1}{eg-1} g\mathbf{Z}_{\mathcal{J}_1}, i \in \mathcal{J}_0\}$, if $\mathcal{J}_0$ is a proper subset of $\mathcal{J}$, and $\kappa = \sum_{\mathcal{J}_0 \subseteq \{1,\ldots,N\}} \kappa_{\mathcal{J}_0}$. In particular, $P_{\mathcal{J}_0}(x)$ and $\mathbb{P}\{\mathbf{V}^c > x\}$ are both regularly varying of index $-\mu$.

The above theorem is used in proving the reduced-load equivalence (see Section V), and may be potentially useful for computational purposes. In particular, in the case of two On-Off flows, the computation of κ is as difficult as the computation of κ_1 and κ_2. Using the probabilistic interpretation of these constants readily leads to an integral expression, which can be solved explicitly when both ν_1 and ν_2 are integer-valued. We omit the details.

VII. Concluding remarks

We have characterized the asymptotic behavior of the workload distribution in a fluid queue fed by multiple heavy-tailed On-Off flows. The results extend previous work, like the bounds derived in [12], and the exact asymptotics in [9] and [16] which rely on strong peak-rate conditions. As a by-product, the proofs lead to several important insights like the extension of the reduced-load equivalence established in [1] (see Section V), and a detailed understanding of the typical overflow behavior (see Section VI). In the analysis, we excluded the case where the drift may be zero during the path to overflow (see Section III-A for a brief discussion), which appears particularly interesting from a theoretical perspective.

There are several other interesting topics for further research. The methodology of Section VI is also applicable to the fluid queue with $M/G/\infty$ input, as is shown [7]. We expect that other similar problems may also have become more accessible, such as related problems multi-server queues, and Generalized Processor Sharing queues. A further avenue for research is the extension of the results to the case of On-Off flows with more general subexponential On-periods, for example Weibull. Partial results in [1] indicate that the typical overflow behavior may then actually be quite different.

Acknowledgment The authors would like to thank Onno Boxma and Miranda van Uitert for useful comments on an earlier version of the paper.

References

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

[2] Anick, D., Mitra, D., Sondhi, M.M. (1982). Stochastic theory of a data-handling system with multiple sources. *Bell Syst. Techn. J.* **61**, 1871–1894.

[3] Arvidsson, A., Karlsson, P. (1999). On traffic models for TCP/IP. In: *Teletraffic Engineering in a Competitive World, Proc. ITC-16*, Edinburgh, UK, eds. P. Key, D. Smith (North-Holland, Amsterdam), 457–466.

[4] Beran, J., Sherman, R., Taqqu, M.S., Wilinger, W. (1995). Long-range dependence in variable-bit-rate video traffic. IEEE *Trans. Commun.* **43**, 1566–1579.

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

[6] Borst, S.C., Zwart, A.P. (2000). A reduced-peak equivalence for queues with a mixture of light-tailed and heavy-tailed input flows. SPOR-Report 2000-04, Eindhoven University of Technology.

[7] Borst, S.C., Zwart, A.P. (2001). Fluid queues with heavy-tailed $M/G/\infty$ input. SPOR-Report 2001-xx, Eindhoven University of Technology.

[8] Boxma, O.J. (1996). Fluid queues and regular variation. *Perf. Eval.* **27 & 28**, 699–712.

[9] Boxma, O.J., Dumas, V. (1998). Fluid queues with heavy-tailed activity period distributions. *Computer Communications* **21**, 1509–1529.

[10] Choudhury, G. L., Whitt, W. (1997). Long-tail buffer-content distributions in broadband networks. *Perf. Eval.* **30**, 177–190.

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

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

[13] Embrechts, P., Klüppelberg, C., Mikosch, T. (1997). *Modelling Extremal Events.* Springer Verlag, Berlin.

[14] Grossglauser, M., Bolot, J.-C. (1999). On the relevance of long-range dependence in network traffic. IEEE/ACM *Trans. Netw.* **7**, 629–640.

[15] Heyman, D., Lakshman, T.V. (1996). What are the implications of long-range dependence for traffic engineering? IEEE/ACM *Trans. Netw.* **4**, 301–317.

[16] Jelenković, P.R., Lazar, A.A. (1999). Asymptotic results for multiplexing subexponential on-off processes. *Adv. Appl. Prob.* **31**, 394–421.

[17] Jelenković, P.R. (1999). Subexponential loss rates in a GI/GI/1 queue with applications. *Queueing Systems* **33**, 91–123.

[18] Jelenković, P.R., Momčilović, P. (2001). Capacity regions for network multiplexers with heavy-tailed fluid On-Off sources. In these proceedings.

[19] Kosten, L. (1974). Stochastic theory of a multi-entry buffer, part 1. *Delft Progress Report, Series F* **1**, 10–18.

[20] Kosten, L. (1984). Stochastic theory of data-handling systems with groups of multiple sources. In: *Performance of Computer-Communication Systems*, H. Rudin, W. Bux (eds.), Elsevier, Amsterdam, The Netherlands, 321–331.

[21] Leland, W.E., Taqqu, M.S., Willinger, W., Wilson, D.V. (1994). On the self-similar nature of Ethernet traffic (extended version). IEEE/ACM *Trans. Netw.* **2**, 1–15.

[22] Likhanov, N., Mazumdar, R.R. (2000). Cell loss asymptotics in buffers fed by heterogeneous long-tailed sources. In: *Proc. Infocom 2000*, 173–180.

[23] Mandjes, M., Borst, S.C. (1999). Overflow behavior in queues with many long-tailed inputs. *Adv. Appl. Prob.*, to appear.

[24] Mandjes, M., Kim, J.-H. (1999). Large deviations for small buffers: an insensitivity result. *Queueing Systems*, to appear.

[25] Mikosch, T., Resnick, S., Rootzén, H., Stegeman, A.W. (1999). Is network traffic approximated by stable Lévy motion or fractional Brownian motion? Technical Report TR1247, Cornell University. *Ann. Appl. Prob.*, to appear.

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

[27] Paxson, V., Floyd, S. (1995). Wide area traffic: the failure of Poisson modeling. IEEE/ACM *Trans. Netw.* **3**, 226–244.

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

[29] Resnick, S., Samorodnitsky, G. (1999). Steady state distribution of the buffer content for $M/G/\infty$ input fluid queues. Technical Report TR1242, Cornell University.

[30] Rolski, T., Schlegel, S., Schmidt, V. (1999). Asymptotics of Palm-stationary buffer content distributions in fluid flow queues. *Adv. Appl. Prob.* **31**, 235–253.

[31] Ryu, B., Elwalid, A.I. (1996). The importance of long-range dependence of VBR video traffic in ATM traffic engineering: myths and realities. *Comp. Commun. Rev.* **13**, 1017–1027.

[32] Stern, T.E., Elwalid, A.I. (1991). Analysis of separable Markov-modulated rate models for information-handling systems. *Adv. Appl. Prob.* **23**, 105–139.

[33] Willinger, W., Taqqu, M.S., Sherman, R., Wilson, D.V. (1997). Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level. IEEE/ACM *Trans. Netw.* **5**, 71–86.

[34] Zwart, A.P. (2000). A fluid queue with a finite buffer and subexponential input. *Adv. Appl. Prob.* **32**, 221–243.

[35] Zwart, A.P., Borst, S.C., Mandjes, M. (2000). Exact asymptotics for fluid queues fed by multiple heavy-tailed On-Off flows. SPOR-Report 2000-14, Eindhoven University of Technology.