Capacity Regions for Network Multiplexers with Heavy-Tailed Fluid On-Off Sources

Predrag Jelenković and Petar Momčilović
Department of Electrical Engineering
Columbia University
New York, NY 10027
{predrag,petar}@ee.columbia.edu

Abstract— **Consider a network multiplexer with a finite buffer fed by a superposition of independent heterogeneous On-Off sources. An On-Off source consists of a sequence of alternating independent activity and silence periods. During its activity period a source produces fluid with constant rate. For this system, under the assumption that the residual activity periods are intermediately regularly varying, we derive explicit and asymptotically exact formulas for approximating the stationary overflow probability and loss rate.**

The derived asymptotic formulas, in addition to their analytical tractability, exhibit excellent quantitative accuracy, which is illustrated by a number of simulation experiments. We demonstrate through examples how these results can be used for efficient computing of capacity regions for network switching elements. Furthermore, the results provide important insight into qualitative tradeoffs between the overflow probability, offered traffic load, available capacity, and buffer space. Overall, they provide a new set of tools for designing and provisioning of networks with heavy-tailed traffic streams.

Keywords—**Network multiplexer, Finite buffer fluid queue, On-Off process, Heavy-tailed distributions, Subexponential distributions, Long-range dependence**

I. Introduction

Increased utilization in communication networks is achieved through sharing of network resources, e.g. link capacity and buffer space, among different user sessions. The benefits in sharing of common resources are counterbalanced with potential increases in congestion and degradation in Quality of Service (QoS) perceived by individual sessions. Therefore, understanding the tradeoffs between the offered traffic load, perceived QoS measures, link capacity and buffer space is essential for efficient design and provision of network switching elements.

The fundamental switching components used for sharing bandwidth and buffer space are network multiplexers. An established baseline model of a network multiplexer is a single server queue with a constant capacity and finite buffer fed by a superposition of many user sessions. Individual sessions are modeled as On-Off processes, since a session can be either active, in which case it transmits data at a specified rate, or silent. The primary performance measures of this queueing system are the stationary overflow probability and loss rate. The analysis of a related infinite buffer queueing system dates back to [1], [2], [3] (see also [4] for additional references).

Most of the early work on multiplexing focuses on On-Off processes with exponentially distributed On and Off periods (e.g., see [3]). However, repeated empirical measurements in modern networks demonstrate the presence of heavy-tailed/subexponential characteristics in network traffic streams. Early discoveries of the presence of heavy-tails in Ethernet traffic were reported in [5]. Long range dependence and subexponential characteristics of VBR video streams (e.g. MPEG) were explored in [6], [7], [8]. Evidence and possible causes of heavy-tailed characteristics in World Wide Web traffic were discussed in [9]. In this paper, we supply additional confirmation of the existence of heavy tails in network traffic. We have measured the distribution of file sizes on five different file servers in the COMET Lab at Columbia University. The empirical distribution of 350,000 surveyed files is presented on a $\log/\log$ scale in Figure 1. We find that the tail of the measured distribution is almost perfectly matched by a Pareto distribution with parameter $\alpha = 1.44$; see the dashed line in Figure 1. This suggests that the corresponding `ftp` (file transfer protocol) traffic is heavy-tailed.

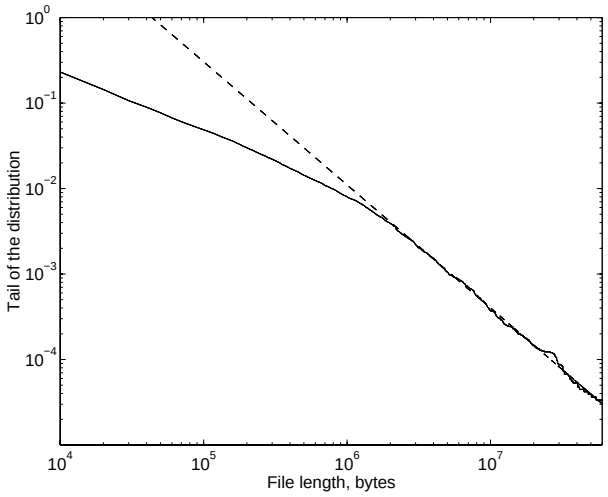

Fig. 1. Log/log plot of the empirical distribution of file sizes on five file servers in COMET laboratory at Columbia University. The tail of the empirical distribution (solid line) is almost perfectly matched by a Pareto distribution with parameter $\alpha = 1.44$ (dashed line).

The analysis of queueing models with multiplexed heavy-tailed renewal arrival sequences, e.g. On-Off processes, is difficult primarily due to the complex dependency structure in the aggregate arrival process [10]. This stems from the well-known fact that the superposition of renewal processes, in general, is not a renewal process. An intermediate case of multiplexing a single heavy-tailed process with exponential streams was investigated in [11], [4], [12]. In [4] it was discovered that

This work is supported by the NSF Presidential Early Career Accomplishment Award No. 9875156.

In Proceedings of IEEE INFOCOM, Anchorage, Alaska, April 2001.

these hybrid queueing systems are asymptotically equivalent to the ones where exponential arrival sequences are replaced with their mean rates. This phenomenon was greatly generalized and termed reduced load equivalence in [13].

An infinite limit of On-Off processes, the so-called M/G/∞ process, represents another instance of an analytically promising model. This is because M/G/∞ processes have both a renewal and Poisson structure. Samples of recent results and additional references on both fluid and discrete time queues with M/G/∞ arrival processes can be found in [11], [4], [14], [15], [16], [17], [18], [19], [20].

However, the understanding of multiplexing a finite number of heavy-tailed On-Off arrival processes is quite limited. General bounds can be found in [21], [22]. In this paper we derive explicit asymptotic results for approximating the stationary overflow probability and loss rate in a finite buffer queue with heterogeneous heavy-tailed On-Off arrival processes. The starting point of our analysis are the results from [23] (see also [24]). During the process of completing this paper we discovered that the complementary results for the infinite buffer model are derived in [25].

Informally, in the case of multiplexing N homogeneous fluid On-Off sources with peak rate r, average rate ρ and probability of being on p_{on} into a queue of capacity c and buffer B our result shows that the fraction of fluid lost is asymptotically, as $B \to \infty$, equal to

$$\frac{R_0 - c}{N\rho} \binom{N}{k_0} p_{on}^{k_0} \mathbb{P}^{k_0} \left[\tau_r^{on} > \frac{B}{R_0 - c} \right], \qquad (1)$$

where $R_0 = k_0 r + (N - k_0)\rho$, τ_r^{on} is the residual On period and k_0 is the smallest integer greater than $(c - N\rho)/(r - \rho)$. Qualitatively, when On periods have Pareto distribution, formula (1) reveals that the fraction of fluid lost decays polynomially in buffer size B and exponentially in capacity c. This insight may prove to be important in designing network switching elements.

Network switching elements can be abstracted by means of capacity regions. Capacity region consists of all combinations of arrival streams that, when fed into a multiplexer of specified capacity, produce required QoS. Let $Q_{\mathbf{n}}^{B,c}, \mathbf{n} = (n_1, \ldots, n_M)$ be the workload in a multiplexer with capacity c, buffer B and arrival sequence which consists of $n_j (1 \le j \le M)$ traffic streams from class j. If we choose the overflow probability as a performance measure and require that this probability is not greater than a specified QoS parameter δ, then the capacity (i.e. admissible) region is defined as

$$\mathcal{C} \equiv \mathcal{C}(c, \delta, M) = \{(n_1, \ldots, n_M) : \mathbb{P}[Q_{\mathbf{n}}^{B,c} = B] \le \delta\}.$$

In the same way loss rate might be chosen to be the performance measure. The obtained results allow for efficient computation of capacity regions.

The rest of the paper is organized as follows. First, in Section II, we present the model description, preliminary results, and necessary extensions of results for a queueing system with a single On-Off arrival process. The main result of this paper, Theorem 2, is presented in Section III. In Section IV, we illustrate the accuracy of this result through simulation experiments. We demonstrate how it can be utilized for efficient computation of capacity regions in network multiplexers. The paper is concluded in Section V.

II. Preliminary Results

Consider a fluid queue model with a constant capacity c, finite buffer B and arrival process $A(t)$. At time t, fluid arrives to this queueing system at rate $A(t)$ and is leaving the system at rate c. When the queue level reaches the buffer limit B fluid arriving in excess of the draining rate c is lost. We use $Q^B(t) \in [0, B]$ to denote the queue content at time t.

In this paper we will only consider arrival processes $A(t)$ that are piece-wise constant and right continuous with almost surely (a.s.) increasing jump times $\{T_0 = 0 < T_1 < T_2 < \cdots\}$. In this case, for any initial value $Q^B(0)$ and $t \in (T_n, T_{n+1}], n \ge 0$, the evolution of $Q^B(t)$ is given by

$$Q^B(t) = (Q^B(T_n) + (t - T_n)(A(T_n) - c))^+ \wedge B, \qquad (2)$$

where $(x)^+ = \max(0, x)$ and $x \wedge y = \min(x, y)$. When necessary, we will use the notation $Q_A^{B,c} \equiv Q^B$ to mark the explicit dependence of $Q^B(t)$ on $A(t)$ and c.

When $A(t)$, i.e. $\{(T_{n+1} - T_n), A(T_n)\}$, is stationary and ergodic, and $\mathbb{E}A(t) < c$, by using Loynes' construction [26], one can show that recursion (2) has a unique stationary and ergodic solution. Furthermore, for all initial conditions $Q^B(0)$, the distribution of $Q^B(t)$ converges to this stationary solution as $t \to \infty$. Unless otherwise indicated we assume throughout the paper that all arrival processes are stationary, ergodic and the corresponding queues are in their stationary regimes. Let Q^B and A be random variables that are equal in distribution to $Q^B(t)$ and $A(t)$, respectively.

Our main objective in this paper is the asymptotic computation, as $B \to \infty$ of the *overflow probability* $\mathbb{P}[Q^B \ge B - K]$, for finite K, and long time average *loss rate* Λ^B given by

$$\Lambda^B \triangleq \lim_{t \to \infty} \frac{L(0,t)}{t},$$

where $L(0,t) = \{\text{amount of fluid lost in } (0,t)\}$. We define the *loss probability* P_{loss}^B as the long time average fraction of fluid that is lost

$$P_{loss}^B \triangleq \lim_{t \to \infty} \frac{L(0,t)}{\int_0^t A(u)du} = \frac{\Lambda^B}{\mathbb{E}A}.$$

The term loss probability stems from the fact that $P_{loss}^B \in [0, 1]$. Since there is a one to one correspondence between the loss rate and loss probability, we use those two terms interchangeably in the paper. An equivalent representation of Λ^B, which will be used for computational purposes, is

$$\Lambda^B = \mathbb{E}\lambda(t), \quad \lambda^B(t) \triangleq (A(t) - c)\,\mathbf{1}\{Q^B(t) = B\};$$

$\lambda^B(t)$ indicates the rate at which the buffer is overflowing at time t. Similarly, the notation $\Lambda_A^{B,c} \equiv \Lambda^B$ will be used to mark the explicit dependence of Λ^B on c and $A(t)$.

Next we prove two useful sample path bounds. The first bound formalizes an intuitively expected notion that multiplexing reduces the aggregate queueing workload.

Proposition 1: Let $A(t) = \sum_{n=1}^N A_n(t)$ and $c = \sum_{n=1}^N c_n$. If $Q_A^{B,c}(t) \le \sum_{n=1}^N Q_{A_n}^{B,c_n}(t)$ for $t = 0$ then the inequality holds for all $t \ge 0$.

Proof: Given in Appendix-B. ■

Now, we consider a stochastic process $\tilde{Q}(t) \equiv \tilde{Q}_A^c(t)$ defined for a right-continuous piece-wise constant arrival processes $A(t)$ with a.s. increasing jump times $\{0 = T_0 < T_1 < T_2 \ldots\}$ by

$$\tilde{Q}(t) = \Big(\tilde{Q}(T_n) + (t - T_n)(c - A(T_n))\Big)^+, \quad t \in (T_n, T_{n+1}], \quad (3)$$

and the initial condition $\tilde{Q}(0)$. Note that $\tilde{Q}(t)$ corresponds to an infinite buffer queueing process with constant arrival rate c and service rate $A(t)$. We use $\tilde{Q}$ to upper bound the amount of free buffer space $B - Q_A^{B,c}(t)$ in the original system.

Proposition 2: If $B - Q_A^{B,c}(t) \leq \tilde{Q}_A^c(t)$ for $t = 0$, then the inequality holds for all $t \geq 0$.

Proof: Given in Appendix-B. ∎

At this point, we turn our attention to a fluid queue with a single On-Off arrival source. The results obtained here will be used for deriving our main theorem in the subsequent section.

First, let us construct an On-Off process. Consider two independent i.i.d. sequences of positive random variables: $\{\tau^{on}, \tau_n^{on}, n \geq 1\}$, $\{\tau^{off}, \tau_n^{off}, n \geq 1\}$. Define a point process $T_n^{off} = \sum_{i=1}^n (\tau_i^{on} + \tau_i^{off})$, $n \geq 1$, $T_0^{off} = 0$; this process represents the beginnings of Off periods in an On-Off process. Next, an On-Off process $A^0(t)$ with rate r is defined as

$$A^0(t) = r \quad \text{if} \quad t \in [T_n^{off} - \tau_n^{on}, T_n^{off}), \quad n \geq 1,$$

and $A^0(t) = 0$, otherwise. We assume that $\mathbb{E}\tau^{on}, \mathbb{E}\tau^{off} < \infty$, and hence, by the Strong Law of Large Numbers, the probability that the source is active in steady state is well defined:

$$p_{on} \triangleq \lim_{t\to\infty} \mathbb{P}[A^0(t) = r] = \frac{\mathbb{E}\tau^{on}}{\mathbb{E}\tau^{on} + \mathbb{E}\tau^{off}}.$$

Process $A^0(t)$ can be extended to a stationary process on the whole real line [22]. We call that process $A(t)$. Note that the expected arrival rate ρ is equal to $\rho \triangleq \mathbb{E}A(t) = p_{on} r$.

In the analysis of renewal processes residual (or excess) random variables and distribution functions play an important role. For a nonnegative random variable X with distribution F and finite mean $\mathbb{E}X$, the residual distribution F_r is defined by

$$F_r(x) = \frac{1}{\mathbb{E}X} \int_0^x (1 - F(u))du, \quad x \geq 0.$$

A random variable X_r with distribution function F_r is called the residual variable of X.

Throughout the paper, for any two real functions $f(x)$ and $g(x)$, we use the standard notation $f(x) \sim g(x)$ as $x \to \infty$ to denote $\lim_{x\to\infty} f(x)/g(x) = 1$ or equivalently $f(x) = g(x)(1 + o(1))$ as $x \to \infty$.

With symbols $\mathcal{S}$ and $\mathcal{IR}$ we denote the classes of subexponential and itermediately regularly varying distributions, respectively. See Appendix-A for the exact definitions.

Proposition 3: If $r > c > \rho$ and $\tau_r^{on} \in \mathcal{S}$, then as $B \to \infty$

$$\mathbb{P}[Q^B = B] \sim p_{on} \mathbb{P}\left[\tau_r^{on} > \frac{B}{r - c}\right].$$

Proof: In [23] it was shown that as $B \to \infty$

$$\Lambda^B \sim \frac{\mathbb{E}[\tau^{on}(r - c) - B]^+}{\mathbb{E}\tau^{on} + \mathbb{E}\tau^{off}}. \quad (4)$$

If $G(0,t)$ defines the amount of time the buffer is full and $L(0,t)$ is the amount of fluid lost in $(0,t)$, then $G(0,t) = L(0,t)/(r-c)$ and by ergodicity of $Q^B(t)$ and (4) as $B \to \infty$

$$\mathbb{P}[Q^B = B] = \lim_{t\to\infty} \frac{L(0,t)}{t(r-c)} \sim p_{on} \mathbb{P}\left[\tau_r^{on} > \frac{B}{r - c}\right].$$

∎

The next result was established in [4]. It provides the asymptotic characterization of the workload in an infinite buffer system.

Theorem 1: If $r > c > \rho$ and $\tau_r^{on} \in \mathcal{S}$, then

$$\mathbb{P}[Q^\infty > B] \sim (1 - p_{on}) \frac{\rho}{c - \rho} \mathbb{P}\left[\tau_r^{on} > \frac{B}{r - c}\right].$$

Note that quantities $\mathbb{P}[Q^B = B]$ and $\mathbb{P}[Q^\infty > B]$ are asymptotically proportional. We use this fact to obtain the following bound.

Proposition 4: If $r > c_i > \rho$, $i = 1, 2$, $\tau_r^{on} \in \mathcal{IR}$ and $\epsilon \in (0,1)$, then for $m > l \geq 0$

$$\lim_{B\to\infty} \frac{\mathbb{P}^m[Q^{B,c_1} \geq (1 - \epsilon)B]}{\mathbb{P}^l[Q^{B,c_2} \geq \epsilon B]} = 0.$$

Proof: Given in Appendix-B. ∎

The proposition below is the main technical result of this section. Due to space limitations the detailed proof is provided in [27].

Proposition 5: If $r > c > \rho$ and $\tau_r^{on} \in \mathcal{IR}$, then

$$\lim_{\epsilon\uparrow 1} \limsup_{B\to\infty} \frac{\mathbb{P}[Q^B \geq \epsilon B]}{\mathbb{P}[Q^B = B]} = 1.$$

III. Main Results

This section contains our main result stated in Theorem 2.

Consider N independent On-Off sources. Without loss of generality, assume that they belong to $M \leq N$ different classes with class i containing n_i identically distributed On-Off sources, $\sum_{i=1}^M n_i = N$. The sources are enumerated as $A_j^{(i)}(t)$, $1 \leq i \leq M$, $1 \leq j \leq n_i$ and the aggregate arrival process is denoted by $A(t) = \sum_{i=1}^M \sum_{j=1}^{n_i} A_j^{(i)}(t)$. $A_j^{(i)}(t)$ is the jth On-Off process of class i with On periods equal in distribution to $\tau_j^{on,(i)}$; peak rate, average rate, and probability of the source being active are equal to $(r_i, \rho_i, p_{on,i})$, respectively. Random variables $\tau^{on,(i)}, \{\tau_j^{on,(i)}\}_{j=1}^{n_i}$ are i.i.d..

Because of the probabilistic sample path techniques that we use in the paper our proofs require the following minor technical assumption. Similar assumptions can be found in [17], [28] and, most recently, in [25].

Assumption 1: The capacity of the queueing system satisfies the following

$$c \notin \left\{ \sum_{i=1}^M [k_i(r_i - \rho_i) + n_i \rho_i] : \mathbf{k} \in \bigotimes_{i=1}^M [0, n_i] \right\},$$

where $\mathbf{k} = (k_1, \ldots, k_M)$ and $\sum_{i=1}^M n_i r_i > c$.

Remark: If this assumption is not satisfied, by choosing an arbitrarily larger or lower capacity one can obtain a lower or upper bound on the queueing performance, respectively.

Before starting and proving our main results we introduce a preparatory lemma. The next lemma derives an asymptotic expression for the overflow probability in the special case when all sources need to be active for a long period of time in order to have a buffer overflow.

Lemma 1: Let $R = \sum_{i=1}^{M} n_i r_i$. If $0 < R - c < r_i - \rho_i$ for all $1 \le i \le M$, then for all $B \ge 0$ and $0 \le \epsilon \le 1$

$$\prod_{i=1}^{M} p_{on,i}^{n_i} \mathbb{P}^{n_i}\left[\tau_r^{on,(i)} > \frac{\epsilon B}{R-c}\right] \le \mathbb{P}[Q_A^{B,c} \ge \epsilon B] \le \prod_{i=1}^{M} \mathbb{P}^{n_i}\left[Q_{A_1^{(i)}}^{B,c-R+r_i} \ge \epsilon B\right].$$

If in addition $\tau_r^{on,(i)} \in \mathcal{S}$ for $1 \le i \le M$, then as $B \to \infty$

$$\mathbb{P}[Q_A^{B,c} = B] \sim \prod_{i=1}^{M} p_{on,i}^{n_i} \mathbb{P}^{n_i}\left[\tau_r^{on,(i)} > \frac{B}{R-c}\right].$$

Proof: Let $c_i \triangleq c - R + r_i$. Assume that at time $t = 0$ all considered queues are empty. For all $1 \le i \le M$, $1 \le j \le n_i$, Proposition 1 yields

$$\begin{aligned} Q_A^{B,c}(t) &\le Q_{A_j^{(i)}}^{B,c_i}(t) + Q_{A-A_j^{(i)}}^{B,c-c_i}(t) \\ &= Q_{A_j^{(i)}}^{B,c_i}(t), \end{aligned} \tag{5}$$

where the equality follows from the fact that $Q_{A-A_j^{(i)}}^{B,c-c_i}(t) \equiv 0,\ t \ge 0$. Since (5) holds for all i, j, then

$$Q_A^{B,c}(t) \le \min_{i,j} Q_{A_j^{(i)}}^{B,c_i}(t),$$

which, by applying the operator $\mathbb{P}[\cdot \ge \epsilon B]$, using independence of $A_j^{(i)}$, and passing $t \to \infty$, yields in stationarity

$$\mathbb{P}[Q_A^{B,c} \ge \epsilon B] \le \prod_{i=1}^{M} \mathbb{P}^{n_i}\left[Q_{A_1^{(i)}}^{B,c_i} \ge \epsilon B\right].$$

Obtaining the lower bound is straightforward from evaluating the system in stationarity at $t = 0$; for simplicity the time index is omitted. Let $\alpha_j^{(i)} = \left\{A_j^{(i)} = r_i, \tau_{j,r}^{on,(i)} > \epsilon B/(R-c)\right\}$, then

$$\begin{aligned} \mathbb{P}[Q_A^{B,c} \ge \epsilon B] &\ge \mathbb{P}\left[Q_A^{B,c} \ge \epsilon B, \bigcap_{i=1}^{M}\bigcap_{j=1}^{n_i} \alpha_j^{(i)}\right] \\ &= \prod_{i=1}^{M} p_{on,i}^{n_i} \mathbb{P}^{n_i}\left[\tau_r^{on,(i)} > \frac{\epsilon B}{R-c}\right]. \end{aligned}$$

Setting $\epsilon = 1$ in the preceding upper and lower bounds and combining it with Proposition 3 yields the second statement of the proposition. ∎

In order to state our main result, we need to introduce some additional notations. Let $\mathcal{E} = \bigotimes_{i=1}^{M}[0, n_i]$ and $\mathcal{E}^* = \bigotimes_{i=1}^{M}[0,1]^{n_i}$. An element $\mathbf{e} \in \mathcal{E}^*$ is of the form $\mathbf{e} = (\mathbf{e}_1, \ldots, \mathbf{e}_M)$, where $\mathbf{e}_i = (e_1^{(i)}, \ldots, e_{n_i}^{(i)}) \in [0,1]^{n_i}$, for all i. In order to distinguish between scalar and vector quantities, vectors are denoted with bold letters. Let $|\mathbf{e}_i| = \sum_{j=1}^{n_j} e_j^{(i)}$ and $r_{\mathbf{m}} = \sum_{i=1}^{M}[m_i(r_i - \rho_i) + n_i\rho_i]$.

Definition 1: Define the minimum overflow set

$$\mathcal{O} \triangleq \{\mathbf{k} \in \mathcal{E} : 0 < r_{\mathbf{k}} - c < r_j - \rho_j, \forall j : k_j > 0\}$$

and the detailed minimum overflow set

$$\mathcal{O}^* \triangleq \{\mathbf{e} \in \mathcal{E}^* : (|\mathbf{e}_1|, \ldots, |\mathbf{e}_M|) \in \mathcal{O}\}.$$

Remarks: (i) Informally, the motivation behind this definition comes from the fact that only a few On-Off processes with very long On periods are causing the most likely buffer overflows, while the remaining processes behave on average. Hence, an element of $\mathcal{O}$ indicates the minimum number of processes from each class that need to have very long On periods in order for a buffer overflow to occur. Similarly, a more detailed set $\mathcal{O}^*$ contains binary vectors which denote particular overflow scenarios. *(ii)* The definition of $\mathbf{e} \in \mathcal{O}^*$ is similar to the definition of the minimal set in [22].

Finally, we are ready to state our main result that derives the exact asymptotic characterization of the loss rate and the buffer overflow probability. To simplify the exposition we define

$$\hat{P}(B) \triangleq \sum_{\mathbf{m}\in\mathcal{O}} \prod_{i=1}^{M} \binom{n_i}{m_i} p_{on,i}^{m_i} \mathbb{P}^{m_i}\left[\tau_r^{on,(i)} > \frac{B}{r_{\mathbf{m}} - c}\right].$$

Theorem 2: If $\sum \rho_i n_i < c$ and $\tau_r^{on,(i)} \in \mathcal{IR}$ for $1 \le i \le M$, then under Assumption 1 as $B \to \infty$

$$\Lambda^B \sim \sum_{\mathbf{m}\in\mathcal{O}} (r_{\mathbf{m}} - c) \prod_{i=1}^{M} \binom{n_i}{m_i} p_{on,i}^{m_i} \mathbb{P}^{m_i}\left[\tau_r^{on,(i)} > \frac{B}{r_{\mathbf{m}} - c}\right]$$

and for any $\eta > 0$ there exists K_η such that for all $K \ge K_\eta$

$$1 - \eta \le \liminf_{B\to\infty} \frac{\mathbb{P}[Q^B \ge B - K]}{\hat{P}(B)} \le \limsup_{B\to\infty} \frac{\mathbb{P}[Q^B \ge B - K]}{\hat{P}(B)} = 1.$$

If, in addition, we assume that $\sum_{i=1}^{M} m_i r_i > c$ for all $\mathbf{m} \in \mathcal{O}$, then for any $K \ge 0$

$$\mathbb{P}[Q_A^{B,c} \ge B - K] \sim \mathbb{P}[Q_A^{B,c} = B] \sim \hat{P}(B) \quad \text{as} \quad B \to \infty.$$

Remarks: (i) Recall that the loss probability is immediately computable from $P_{loss}^B = \Lambda^B / \mathbb{E}A$.
(ii) Informally, for large K and B much larger than K

$$\mathbb{P}[Q_A^{B,c} \ge B - K] \approx \hat{P}(B).$$

Hence, the result states that the fraction of time during which the buffer is effectively 100% full is asymptotically equal to $\hat{P}(B)$. *(iii)* The heuristic for this result can be easily explained by the following simple example. Consider two i.i.d. On-Off processes with On periods in $\mathcal{IR}$ and $r < c < r + \rho$; these assumptions result in the overflow set being a single number $m = 1$. In this case, the most probable way the buffer overflows is when one of the processes (say the first one) has a very long On period and the other behaves on average $\int_0^t A_2^{(1)}(u)du \approx \rho t$. During that long On period, the average amount of arriving fluid will be

higher than the service rate $r+\rho > c$ and the buffer will tend to fill. After the buffer fills ($Q^B(t) = B$), its content will stay close to the buffer boundary; when $r < c$, the queueing content will make small excursions away from the boundary during the Off periods in the second On-Off process, see Figure 2. In the proof we show that these excursions are almost surely finite and uniformly bounded for all B.

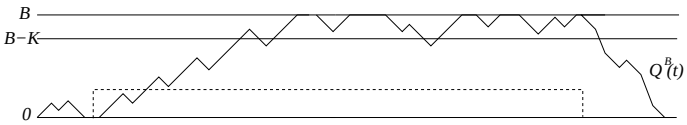

Fig. 2. Illustration for Remark (iii) after Theorem 2. The long On period is shown with a dashed line.

(iv) In the last statement of the theorem the values of ρ_i-s do not affect the computation of the minimal overflow set. Hence, during the most likely overflow event the arrival rate is always higher than the capacity and, therefore, the buffer content Q^B remains on the boundary B. This fact makes the asymptotic computation of the probability that the buffer is full $\mathbb{P}[Q^B = B]$ feasible. Also, due to the fluid nature of the model, $\mathbb{P}[Q^B = B]$ represents the fraction of time that the fluid is being lost.

(v) With additional assumptions on the ratios of tails of $\tau_r^{on,(i)}$ the minimum overflow set $\mathcal{O}$ in the statement of the theorem can be replaced by a smaller, most probable overflow set $\mathcal{O}_0$. For example, if $\tau^{on,(i)} \in \mathcal{R}_{\alpha_i}$, then

$$\mathcal{O}_0 = \left\{ \mathbf{k} \in \mathcal{O} : \sum_{i=1}^{M} \alpha_i k_i = \min_{\mathbf{l}\in\mathcal{O}} \sum_{i=1}^{M} \alpha_i l_i \right\},$$

with $\mathbf{k} = (k_1, \ldots, k_M)$ and $\mathbf{l} = (l_1, \ldots, l_M)$.

(vi) Complementary results for the infinite buffer model are recently obtained in [25]. A related result for a discrete time finite buffer queue loaded by a Pareto-like M/G/∞ arrival process can be found in [28]; in their proofs the authors exploit Poisson decomposition property of the arrival processes, which does not hold for the multiplexed On-Off processes. In addition, in [28] it is assumed that the buffer overflows in a unique way, i.e., the overflow set $\mathcal{O}$ contains a single element.

Before proving Theorem 2, we introduce a combinatorial lemma that will be used in the proof of the main theorem. Let $\{X_j^{(i)}, 1 \le j \le n_i\}_{i=1}^M$ be a set of independent random variables. Random variables with the same superscript are equal in distribution. For every element $\mathbf{e} \in \mathcal{E}^*$, let us define the following quantity

$$S_{\mathbf{e}} \triangleq \sum_{i=1}^{M} \sum_{j=1}^{n_i} (1 - e_j^{(i)}) X_j^{(i)}.$$

At this point, we are ready to state our last preparatory lemma. For two vectors $\mathbf{m}$ and $\mathbf{k}$ we say $\mathbf{m} > \mathbf{k}$ if $m_i \ge k_i$ for all i and $m_j > k_j$ for at least one j.

Lemma 2: There exists a finite set $\overline{\mathcal{O}}$ with a feature that for each $\mathbf{m} \in \overline{\mathcal{O}}$ exist $\mathbf{k} \in \mathcal{O}$ such that $\mathbf{m} > \mathbf{k}$, and

$$\mathbb{P}\left[\min_{\mathbf{e}\in\mathcal{O}^*} S_{\mathbf{e}} > y\right] \le \sum_{\mathbf{m}\in\overline{\mathcal{O}}} \prod_{i=1}^{M} \binom{n_i}{m_i} \mathbb{P}^{m_i}\left[X_1^{(i)} > \frac{y}{N}\right].$$

Proof: Here we prove the result for the case $M = 1$. An inductive proof for general M is given in [27].

In this case $N = n_1$ and, by Assumption 1, there exists an integer k_1 such that for all $\mathbf{e} \in \mathcal{O}^*$ we have $\sum_{j=1}^{N} e_j = k_1$. Define $D \equiv D(y)$ as the number of events $\{X_j^{(1)} > y/N\}$:

$$D \triangleq \sum_{j=1}^{n_1} \mathbf{1}\{X_j^{(1)} > y/N\}.$$

Next, observe that

$$\{D \le k_1\} \cap \left\{\min_{\mathbf{e}\in\mathcal{O}^*} S_{\mathbf{e}} > y\right\} = \emptyset$$

due to the fact that there is at least one $\mathbf{e} \in \mathcal{O}^*$ such that $S_{\mathbf{e}} \le \frac{N-k_1}{N} y \le y$ on $\{D \le k_1\}$. Thus,

$$\begin{aligned}
\mathbb{P}\left[\min_{\mathbf{e}\in\mathcal{O}^*} S_{\mathbf{e}} > y\right] &\le \mathbb{P}[D \ge k_1 + 1] \\
&= \mathbb{P}\left[\bigcup_{\substack{\mathbf{d}\in[0,1]^{n_1} \\ |\mathbf{d}|=k_1+1}} \bigcap_{j\,:\,d_j=1} \left\{X_j^{(1)} > \frac{y}{N}\right\}\right] \\
&\le \binom{N}{k_1+1} \mathbb{P}^{k_1+1}\left[X_1^{(1)} > \frac{y}{N}\right],
\end{aligned}$$

where $\binom{N}{k+1} \equiv 0$ if $N < k+1$. Therefore, with the choice of $\overline{\mathcal{O}} = \{k_1+1\}$, the case $M = 1$ is proven. ■

Below we present the proof of our main result, Theorem 2.

Proof of Theorem 2: Due to the space limitation we provide the proof only for the overflow probability. The proof for the loss rate can be completed in the same spirit and is provided in [27]. The proof consists of a lower and an upper bound.

Upper bound. Let $c_{\mathbf{e}} \triangleq c - \sum_{i=1}^{M}\sum_{j=1}^{n_i}(1 - e_j^{(i)})\rho_i$ and $A_{\mathbf{e}} \triangleq \sum_{i=1}^{M}\sum_{j=1}^{n_i} e_j^{(i)} A_j^{(i)}$. For $\delta > 0$ consider the queues $Q_{A_{\mathbf{e}}}^{B,c_{\mathbf{e}}-\delta}$, $\mathbf{e} \in \mathcal{O}^*$, $Q_{A_j^{(i)}}^{B,\rho_i+\delta/N}$; assume that these queues are empty at time $t = 0$. For any $\mathbf{e} \in \mathcal{O}^*$ and sufficiently small $\delta > 0$ such that all considered queues have their capacity greater than the average arrival rate Proposition 1 yields ($Q_A^{B,c} \equiv Q^B$) for $t \ge 0$

$$Q_A^{B,c}(t) \le Q_{A_{\mathbf{e}}}^{B,c_{\mathbf{e}}-\delta}(t) + \sum_{i=1}^{M}\sum_{j=1}^{n_i}(1 - e_j^{(i)}) Q_{A_j^{(i)}}^{B,\rho_i+\delta/N}(t),$$

and, thus

$$Q_A^{B,c}(t) \le \min_{\mathbf{e}\in\mathcal{O}^*}\left(Q_{A_{\mathbf{e}}}^{B,c_{\mathbf{e}}-\delta}(t) + \sum_{i=1}^{M}\sum_{j=1}^{n_i}(1 - e_j^{(i)}) Q_{A_j^{(i)}}^{B,\rho_i+\frac{\delta}{N}}(t)\right).$$

Next, by applying the operator $\mathbb{P}[\cdot \ge B - K]$ in the preceding inequality and then passing $t \to \infty$, we derive in stationarity

$$\begin{aligned}
&\mathbb{P}[Q_A^{B,c} \ge B - K] \le \\
&\mathbb{P}\left[\min_{\mathbf{e}\in\mathcal{O}^*}\left(Q_{A_{\mathbf{e}}}^{B,c_{\mathbf{e}}-\delta} + \sum_{i=1}^{M}\sum_{j=1}^{n_i}(1 - e_j^{(i)}) Q_{A_j^{(i)}}^{B,\rho_i+\delta/N}\right) \ge B - K\right].
\end{aligned}$$

Therefore, the above inequality, union bound and Lemmas 1, 2 yield for any $\epsilon \in (0,1)$

$$\begin{aligned}
&\mathbb{P}[Q_A^{B,c} \geq B-K] \\
&\leq \mathbb{P}\left[\bigcup_{\mathbf{e}\in\mathcal{O}^*}\{Q_{A_{\mathbf{e}}}^{B,c_{\mathbf{e}}-\delta} \geq \epsilon(B-K)\}\right] \\
&\quad + \mathbb{P}\left[\min_{\mathbf{e}\in\mathcal{O}^*}\sum_{i=1}^{M}\sum_{j=1}^{n_i}(1-e_j^{(i)})Q_{A_j^{(i)}}^{B,\rho_i+\frac{\delta}{N}} \geq (1-\epsilon)(B-K)\right] \\
&\leq \sum_{\mathbf{e}\in\mathcal{O}^*}\mathbb{P}\left[Q_{A_{\mathbf{e}}}^{B,c_{\mathbf{e}}-\delta} \geq \epsilon(B-K)\right] \\
&\quad + \sum_{\mathbf{m}\in\overline{\mathcal{O}}}\prod_{i=1}^{M}\binom{n_i}{m_i}\mathbb{P}^{m_i}\left[Q_{A_1^{(i)}}^{B,\rho_i+\frac{\delta}{N}} \geq \frac{1-\epsilon}{N}(B-K)\right] \\
&\leq \sum_{\mathbf{m}\in\mathcal{O}}\prod_{i=1}^{M}\binom{n_i}{m_i}\mathbb{P}^{m_i}\left[Q_{A_1^{(i)}}^{B,c_{\mathbf{m}}^i-\delta} \geq \epsilon(B-K)\right] \\
&\quad + \sum_{\mathbf{m}\in\overline{\mathcal{O}}}\prod_{i=1}^{M}\binom{n_i}{m_i}\mathbb{P}^{m_i}\left[Q_{A_1^{(i)}}^{B,\rho_i+\frac{\delta}{N}} \geq \frac{1-\epsilon}{N}(B-K)\right], \qquad (6)
\end{aligned}$$

where $c_{\mathbf{m}}^i \triangleq c_{\mathbf{e}} - \sum_{j=1}^{M} m_j r_j + r_i$. Now, (6), in conjunction with Proposition 4 and Lemma 2, results in

$$\begin{aligned}
&\limsup_{B\to\infty}\frac{\mathbb{P}[Q_A^{B,c} \geq B-K]}{\hat{P}(B)} \leq \\
&\limsup_{B\to\infty}\frac{\sum_{\mathbf{m}\in\mathcal{O}}\prod_{i=1}^{M}\binom{n_i}{m_i}\mathbb{P}^{m_i}\left[Q_{A_1^{(i)}}^{B,c_{\mathbf{m}}^i-\delta} \geq \epsilon(B-K)\right]}{\sum_{\mathbf{m}\in\mathcal{O}}\prod_{i=1}^{M}\binom{n_i}{m_i}p_{on,i}^{m_i}\mathbb{P}^{m_i}\left[\tau_r^{on,(i)} > \frac{B}{r_i-c_{\mathbf{m}}^i}\right]}. \qquad (7)
\end{aligned}$$

Here, recall that Proposition 5 implies for all $\mathbf{m}$ and i

$$\lim_{\delta\downarrow 0}\lim_{\epsilon\uparrow 1}\limsup_{B\to\infty}\frac{\mathbb{P}\left[Q_{A_1^{(i)}}^{B,c_{\mathbf{m}}^i-\delta} \geq \epsilon(B-K)\right]}{\mathbb{P}\left[Q_{A_1^{(i)}}^{B,c_{\mathbf{m}}^i} = B\right]} = 1,$$

which, by Proposition 3 and Lemma 4, yields

$$\lim_{\delta\downarrow 0}\lim_{\epsilon\uparrow 1}\limsup_{B\to\infty}\frac{\sum_{\mathbf{m}\in\mathcal{O}}\prod_{i=1}^{M}\binom{n_i}{m_i}\mathbb{P}^{m_i}\left[Q_{A_1^{(i)}}^{B,c_{\mathbf{m}}^i-\delta} \geq \epsilon(B-K)\right]}{\sum_{\mathbf{m}\in\mathcal{O}}\prod_{i=1}^{M}\binom{n_i}{m_i}\mathbb{P}^{m_i}\left[Q_{A_1^{(i)}}^{B,c_{\mathbf{m}}^i} = B\right]} = 1.$$

Finally, by using the last limit and letting $\delta \downarrow 0$ and $\epsilon \uparrow 1$ in (7) we derive the upper bound.

Lower bound. The lower bound is obtained by estimating the queueing system in stationarity at (say) time $t = 0$. Let $r_{\mathbf{e}} \triangleq \sum_{i=1}^{M}\sum_{j=1}^{n_i}[e_j^{(i)}(r_i-\rho_i)+\rho_i]$ and, for any $\epsilon > 0$, define a family of events indicating that the jth process of type i is active at time $t = 0$ and its On period has lasted for an amount of time larger then $t_{\mathbf{e}} \triangleq B(1+\epsilon)/(r_{\mathbf{e}}-c)$

$$\alpha_j^{(i)} \triangleq \left\{A_j^{(i)}(0) = r_i,\ \inf\{t>0 : A_j^{(i)}(-t) = 0\} > t_{\mathbf{e}}\right\},$$

with $\alpha_j^{(i),c}$ being the complement of $\alpha_j^{(i)}$. We point out that $\inf\left\{t>0 : A_j^{(i)}(-t) = 0\right\}$ equals in distribution to $\tau_{j,r}^{on,(i)}$ on event $\{A_j^{(i)}(0) = r_i\}$. Next,

$$\begin{aligned}
&\mathbb{P}[Q_A^{B,c}(0) \geq B-K] \\
&\geq \sum_{\mathbf{e}\in\mathcal{O}^*}\mathbb{P}\left[Q_A^{B,c}(0) \geq B-K,\ \bigcap_{i,j\,:\,e_j^{(i)}=1}\alpha_j^{(i)},\ \bigcap_{i,j\,:\,e_j^{(i)}=0}\alpha_j^{(i),c}\right] \\
&\triangleq \sum_{\mathbf{e}\in\mathcal{O}^*}P_{\mathbf{e}}. \qquad (8)
\end{aligned}$$

For all arrival processes we define the following sample path averages

$$\overline{A}_j^{(i)} \equiv \overline{A}_j^{(i)}(\epsilon,B) \triangleq \frac{1}{t_{\mathbf{e}}}\int_{-t_{\mathbf{e}}}^{0}A_j^{(i)}(u)du$$

and two collections of disjoint events

$$\begin{aligned}
\gamma_{\mathbf{e}}(B) &\triangleq \left\{\bigcap_{i,j\,:\,e_j^{(i)}=0}\left\{\alpha_j^{(i),c}, \overline{A}_j^{(i)} > \rho_i - \frac{\epsilon(r_{\mathbf{e}}-c)}{N(1+\epsilon)}\right\}\right\}, \\
\xi_{\mathbf{e}}(B) &\triangleq \left\{\bigcap_{i,j\,:\,e_j^{(i)}=1}\alpha_j^{(i)},\ \gamma_{\mathbf{e}}(B)\right\}. \qquad (9)
\end{aligned}$$

Then, $P_{\mathbf{e}}$ is bounded from below by

$$P_{\mathbf{e}} \geq \mathbb{P}\left[Q_A^{B,c}(0) \geq B-K, \xi_{\mathbf{e}}(B)\right]. \qquad (10)$$

Next, we estimate the probability of event $\{Q_A^{B,c}(0) \geq B-K,\ \xi_{\mathbf{e}}(B)\}$. For notational purposes define $\tilde{c}_{\mathbf{e}} \triangleq c - \sum_{i=}^{M} r_i|\mathbf{e}_i|$, $\tilde{A}_{\mathbf{e}} \triangleq A - A_{\mathbf{e}}$ and let $\underline{Q}_{\tilde{A}_{\mathbf{e}}}^{B,\tilde{c}_{\mathbf{e}}}(t)$, $\underline{\tilde{Q}}_{\tilde{A}_{\mathbf{e}}}^{\tilde{c}_{\mathbf{e}}}(t)$ be defined by recursions (2), (3) and initial conditions $\underline{Q}_{\tilde{A}_{\mathbf{e}}}^{B,\tilde{c}_{\mathbf{e}}}(-t_{\mathbf{e}}) = 0$, $\underline{\tilde{Q}}_{\tilde{A}_{\mathbf{e}}}^{\tilde{c}_{\mathbf{e}}}(-t_{\mathbf{e}}) = B$, respectively. From (10) and inequality $Q_A^{B,c}(0) \geq \underline{Q}_{\tilde{A}_{\mathbf{e}}}^{B,\tilde{c}_{\mathbf{e}}}(0)$ we derive

$$\begin{aligned}
P_{\mathbf{e}} &\geq \mathbb{P}\left[B - \underline{Q}_{\tilde{A}_{\mathbf{e}}}^{B,\tilde{c}_{\mathbf{e}}}(0) \leq K, \xi_{\mathbf{e}}(B)\right] \\
&\geq \mathbb{P}\left[\underline{\tilde{Q}}_{\tilde{A}_{\mathbf{e}}}^{\tilde{c}_{\mathbf{e}}}(0) \leq K,\ \xi_{\mathbf{e}}(B)\right] \\
&= \mathbb{P}\left[\underline{\tilde{Q}}_{\tilde{A}_{\mathbf{e}}}^{\tilde{c}_{\mathbf{e}}}(0) \leq K,\ \gamma_{\mathbf{e}}(B)\right]\prod_{i=1}^{M}\left(\mathbb{P}[\alpha_1^{(i)}]\right)^{|\mathbf{e}_i|}, \qquad (11)
\end{aligned}$$

where the second inequality follows from Proposition 2 and the last equality from the independence of processes $A_{\mathbf{e}}$ and $\tilde{A}_{\mathbf{e}} = A - A_{\mathbf{e}}$. Using the standard queueing reflection mapping argument quantity $\underline{\tilde{Q}}_{\tilde{A}_{\mathbf{e}}}^{\tilde{c}_{\mathbf{e}}}(0)$ can be represented as

$$\underline{\tilde{Q}}_{\tilde{A}_{\mathbf{e}}}^{\tilde{c}_{\mathbf{e}}}(0) = \sup_{-t_{\mathbf{e}}\leq s\leq 0}\left\{\tilde{c}_{\mathbf{e}}|s| - \int_s^0\tilde{A}_{\mathbf{e}}du\right\} \vee \left(B + \tilde{c}_{\mathbf{e}}t_{\mathbf{e}} - \int_{-t_{\mathbf{e}}}^{0}\tilde{A}_{\mathbf{e}}du\right).$$

By noting that $B + \tilde{c}_{\mathbf{e}}t_{\mathbf{e}} - \int_{-t_{\mathbf{e}}}^{0}\tilde{A}_{\mathbf{e}}du < 0$ on event $\xi_{\mathbf{e}}(B)$ and $\tilde{c}_{\mathbf{e}} < \mathbb{E}\tilde{A}_{\mathbf{e}}$ we conclude that on $\xi_{\mathbf{e}}(B)$

$$\underline{\tilde{Q}}_{\tilde{A}_{\mathbf{e}}}^{\tilde{c}_{\mathbf{e}}}(0) \leq \sup_{s\leq 0}\left\{\tilde{c}_{\mathbf{e}}|s| - \int_s^0\tilde{A}_{\mathbf{e}}du\right\} \stackrel{d}{=} \tilde{Q}_{\tilde{A}_{\mathbf{e}}}^{\tilde{c}_{\mathbf{e}}}(0) < \infty \quad \text{a.s.} \qquad (12)$$

Next, the stationarity and ergodicity of the arrival processes and the fact that the residual On periods are a.s. finite result in

$$\lim_{B\to\infty} \mathbb{P}\left[\gamma_{\mathbf{e}}(B)\right] = 1. \tag{13}$$

Now, (8), in conjunction with (11), (12) and (13), leads to

$$\liminf_{B\to\infty} \frac{\mathbb{P}[Q_A^{B,c} \geq B-K]}{\hat{P}(B)} \geq$$

$$\min_{\mathbf{e}\in\mathcal{O}^*} \mathbb{P}[\tilde{Q}_{\bar{A}_{\mathbf{e}}}^{\bar{c}_{\mathbf{e}}}(0) \leq K] \ \liminf_{B\to\infty} \frac{\sum_{\mathbf{e}\in\mathcal{O}^*} \prod_{i=1}^{M} \left(\mathbb{P}[\alpha_1^{(i)}(\epsilon,B)]\right)^{|\mathbf{e}_i|}}{\hat{P}(B)}.$$

At this point, by counting the number of identical elements in the above sum, using the fact that $\tau_r^{on,(i)} \in \mathcal{IR}$, Lemma 4 and passing $\epsilon \downarrow 0$ in the preceding inequality, we obtain

$$\liminf_{B\to\infty} \frac{\mathbb{P}[Q_A^{B,c} \geq B-K]}{\hat{P}(B)} \geq \min_{\mathbf{e}\in\mathcal{O}^*} \mathbb{P}[\tilde{Q}_{\bar{A}_{\mathbf{e}}}^{\bar{c}_{\mathbf{e}}}(0) \leq K].$$

Finally, the last inequality and (12) yield the lower bound and the statement of the theorem. ∎

For the case of homogeneous sources ($M = 1$), the expressions for the loss rate and overflow probability admit simpler forms.

Corollary 1: Homogeneous sources ($M = 1$). Let $r_0 = k_0 r + (N-k_0)\rho$ and

$$\hat{P}(B) \triangleq \binom{N}{k_0} p_{on}^{k_0} \mathbb{P}^{k_0}\left[\tau_r^{on} > \frac{B}{r_0 - c}\right],$$

If $\rho N < c < Nr$, $\tau_r^{on} \in \mathcal{IR}$, and there is an integer k_0 such that $0 < r_0 - c < r - \rho$, then as $B \to \infty$

$$\Lambda^B \sim (r_0 - c)\hat{P}(B)$$

and for any $\eta > 0$ there exists K_η such that for all $K \geq K_\eta$

$$1-\eta \leq \liminf_{B\to\infty} \frac{\mathbb{P}[Q_A^{B,c} \geq B-K]}{\hat{P}(B)} \leq \limsup_{B\to\infty} \frac{\mathbb{P}[Q_A^{B,c} \geq B-K]}{\hat{P}(B)} = 1.$$

If, in addition, $k_0 r > c$ then $\mathbb{P}[Q_A^{B,c} = B] \sim \hat{P}(B)$ as $B \to \infty$.

Next, we allow for some of the multiplexed arrival processes to have lighter then polynomial tails; we term these processes subpolynomial. A stationary, ergodic and right-continuous with left limits process $A(t)$ is subpolynomial ($A \in \mathcal{SP}$) if for all $c > \mathbb{E}A(t)$ and $\beta > 0$ the stationary workload of the corresponding infinite buffer queue $Q_A^{\infty,c}$ satisfies

$$\lim_{B\to\infty} B^\beta \mathbb{P}[Q_A^{\infty,c} \geq B] = 0.$$

The above condition is satisfied for a general class of exponentially bounded arrival processes (see [29], [30]). It can be seen that it also holds for some heavy-tailed processes, e.g., On-Off processes with Weibull On periods, $\mathbb{P}[\tau^{on} > x] = e^{-x^b}, 0 < b < 1, x \geq 0$ (see Theorem 1). Note that if $A_1, A_2 \in \mathcal{SP}$ then $A_1 + A_2 \in \mathcal{SP}$. This easily follows from the well known fact that $Q_{A_1+A_2}^{\infty,c_1+c_2}$ is stochastically dominated by $Q_{A_1}^{\infty,c_1} + Q_{A_2}^{\infty,c_2}$, $c_i < \mathbb{E}A_i$ (an infinite buffer equivalent of Proposition 1). Thus, we will use $A_{\mathcal{SP}}$ to denote the aggregate process of all arriving subpolynomial processes. Assume that all considered queues are stationary and ergodic; in case of $A_{\mathcal{SP}}$ being piece-wise constant this follows from the discussion in Section 2. The following corollary yields the reduce load equivalence results for multiplexing subpolynomial and intermediately regularly varying processes.

Corollary 2: Suppose that $A_{\mathcal{SP}} \in \mathcal{SP}$ and Assumption 1 is satisfied with $(c - \mathbb{E}A_{\mathcal{SP}})$ in place of c. If $\sum_{i=1}^{M} n_i\rho_i < c - \mathbb{E}A_{\mathcal{SP}}$ and $\tau_r^{on,(i)} \in \mathcal{IR}$ for $1 \leq i \leq M$, then as for any $\eta > 0$ there exists K_η such that for all $K \geq K_\eta$

$$1-\eta \leq \liminf_{B\to\infty} \frac{\mathbb{P}\left[Q_{A+A_{\mathcal{SP}}}^{B,c} \geq B-K\right]}{\mathbb{P}\left[Q_A^{B,c-\mathbb{E}A_{\mathcal{SP}}} \geq B-K\right]}$$
$$\leq \limsup_{B\to\infty} \frac{\mathbb{P}\left[Q_{A+A_{\mathcal{SP}}}^{B,c} \geq B-K\right]}{\mathbb{P}\left[Q_A^{B,c-\mathbb{E}A_{\mathcal{SP}}} \geq B-K\right]} \leq 1+\eta$$

and, if for some $\delta > 0$, $\mathbb{E}A_{\mathcal{SP}}^{1+\delta} < \infty$,

$$\Lambda_{A+A_{\mathcal{SP}}}^{B,c} \sim \Lambda_A^{B,c-\mathbb{E}A_{\mathcal{SP}}} \quad \text{as} \quad B \to \infty.$$

Proof: Provided in [27]. ∎

Remarks on the discrete time model: Here, we show that our results extend to the discrete time model as well. In fact, the discrete time and fluid models are asymptotically equivalent. Often, discrete time models appear convenient for simulation experiments and numerical computations and are commonly used in the telecommunication literature (e.g., see [28], [14], [15]).

Consider a nonnegative discrete time arrival process $a[T]$, $T \in \mathbb{N}_0$ with a bounded peak rate r_{max}. Let $q^B[T]$ be the workload at time T in a discrete time queue with capacity c, buffer size B and arrival process $a[T]$. The evolution of the process $q^B[T]$ is governed by

$$q^B[T+1] = (q^B[T] + a[T+1] - c)^+ \wedge B,$$

with $q^B[0] = 0$. Now, define a right-continuous process $\tilde{a}(t) \triangleq a[\lfloor t \rfloor + 1]$, where $\lfloor t \rfloor$ denotes the integer part of t, and the corresponding fluid queue process $Q^B(t) \equiv Q_{\tilde{a}}^{B,c}(t)$. Then, simple sample path arguments yield for all $t \geq 0$

$$q^B[\lfloor t \rfloor] - c \leq Q^B(t) \leq q^B[\lfloor t \rfloor] + r_{max}.$$

Thus, from the preceding inequality and long-tailed nature of $Q^B(t)$ the extension of Theorem 2 to the discrete time model is immediate.

IV. Numerical Examples

In this section we illustrate, through simulation experiments, the precision of our asymptotic results in approximating the overflow probabilities for finite buffers sizes. Then, we demonstrate how these results can be used for real time computation of capacity regions in network multiplexing elements. Efficient and accurate estimation of the available capacity is of outmost importance both for network provisioning and admission control.

The first example demonstrate the accuracy of Theorem 2, or more precisely Corollary 1.

Example 1: Consider a multiplexer with buffer size B and an output link with speed $c = 35$ kbits/s. Let $N = 10$ users with access speed $r = 20$ kbits/s share the system. The users are sending data files to the buffer. The probability that a particular user has something to send at a given moment is $p_{on} = 0.1$. The distribution of file sizes is chosen to be the one reported in Section I. Since the asymptotic results are insensitive to the distribution of Off periods we choose the distribution of time between two file transfers to be exponential $\mathbb{P}[\tau^{off} > x] = e^{-\lambda x}, x \geq 0$ for all users. Now, the asymptotic approximation from Corollary 1 computes explicitly to

$$\hat{P}(B) = 4.5\ 10^{-5}\ \left(1 - F_r(3B)\right)^2,$$

where F_r is the residual distribution of file sizes and B is in bytes. To ensure an increased accuracy of our experiment we selected the length of the simulated sample path to be $t = 10^{12}$ s. The experiment took approximately six hours on a Pentium III PC.

We simulated the overflow probability and loss rate for buffer sizes $B = 0.5 \times i$ Mbytes, $i = 1, \ldots, 10$. The results of the simulation are presented in Figure 3 and Figure 4 with symbols "+" for the overflow probability and loss rate, respectively. The approximations are plotted on the same figures with solid lines. Its striking accuracy is apparent. It is interesting to note that on average 2 kbits/s is sent to the buffer and that the buffer is overflowing at an average rate between 0.5-5 bits/s for the given range of buffer sizes (see Figure 4).

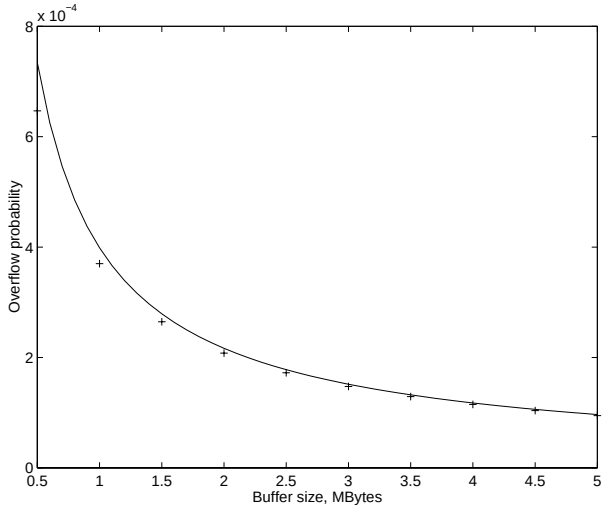

Fig. 3. Illustration for Example 1: Overflow probability

Next, we describe how our result can be used for efficient computation of capacity regions for network multiplexers. For our simulation experiments we choose two traffic classes $M = 2$ of On-Off sources. The sources are of the same type as described in the previous two examples; they are completely characterized by triples $(\lambda_j, \alpha_j, r_j), j = 1, 2$. The approximation $\hat{P}(B) \equiv \hat{P}$, as defined in Theorem 2, is easily computable. Our implementation in Matlab produced real-time answers. Here, we provide two simulation studies in which we check the correctness of the asymptotic method in computing capacity region $\mathcal{C}$ defined in Section I. These simulation studies were much lengthier than in the preceding examples and, therefore, we had

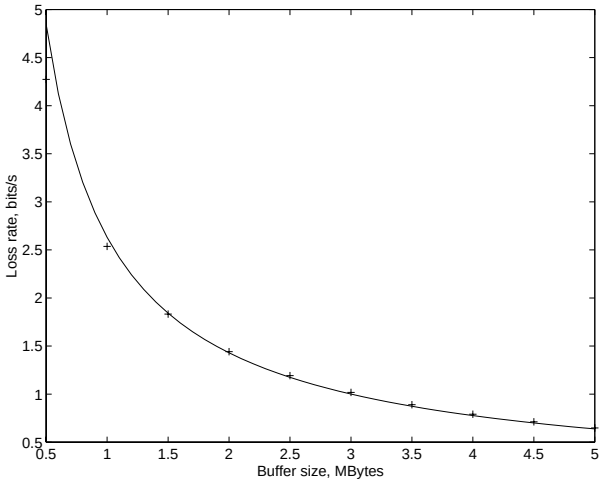

Fig. 4. Illustration for Example 1: Loss rate

to optimize the simulation time; our measurements of the overflow probabilities had sufficiently low variance for simulation runs equal to 10^7 time units.

Example 2: We set the triplets $(\lambda_i, \alpha_i, r_i)$ to be (0.041,1.9,13) for class I and (0.176,1.7,5) for class II. This results in $p_{on,1} = 0.08$, $p_{on,2} = 0.3$, and $\rho_1 = 1.04$, $\rho_1 = 1.5$. The simulation experiment was conducted for the choice of $c = 23.02$ and $B = 600$. The capacity of the system is chosen in such a way that Assumption 1 is satisfied a priori for all possible choices of n_1 and n_2. The QoS parameter δ is set to 10^{-5}. The outcome of the experiment is presented on Figure 5 with "+" symbols. The experiment took seven hours on a Pentium III PC. On the same figure with symbols "o" we indicated the approximation of $\mathcal{C}$ obtained with Theorem 2. Both the simulation and the analytical approximation produced the same capacity region. In order to provide the reader with the information on system utilization we plotted with a dashed line the border of the stability region defined by $n_1\rho_1 + n_2\rho_2 < c$.

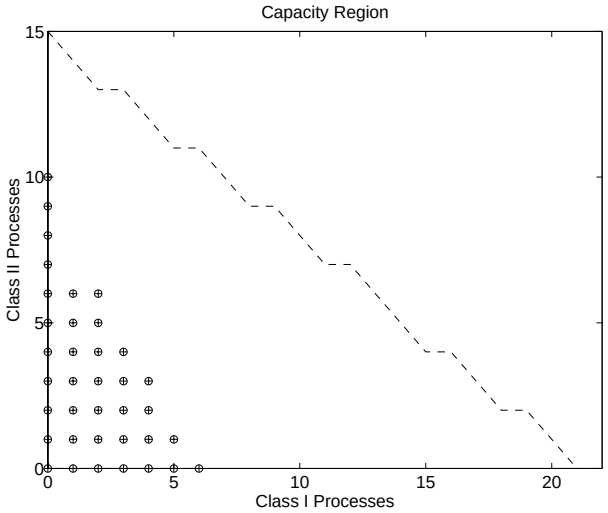

Fig. 5. Illustration for Example 2

Example 3: Consider the previous example with overflow probability requirement $\delta = 10^{-6}$. Let the queue parameters

be $c = 95.105$ and $B = 1000$. The reason for not making the parameter c a round number is the fact that Assumption 1 needs to be satisfied. Class I and II On-Off sources are determined by triplets (0.158,1.9,20) and (0.349,2.1,10), respectively. This yields $p_{on,1} = 0.25$, $p_{on,2} = 0.4$, and $\rho_1 = 5$, $\rho_1 = 4$. The capacity region for this case is presented in Figure 6. Symbols "+" indicate the results of the simulation experiment and symbols "o" denote our analytic approximation. On the same graph the border of the stability region is plotted with a dashed line. The experiment took two days to complete. It is evident from the figure that the capacity regions computed by lengthy simulation and readily computable analytic approximation are almost the same. This exemplifies the importance of having analytical tools for computing these regions.

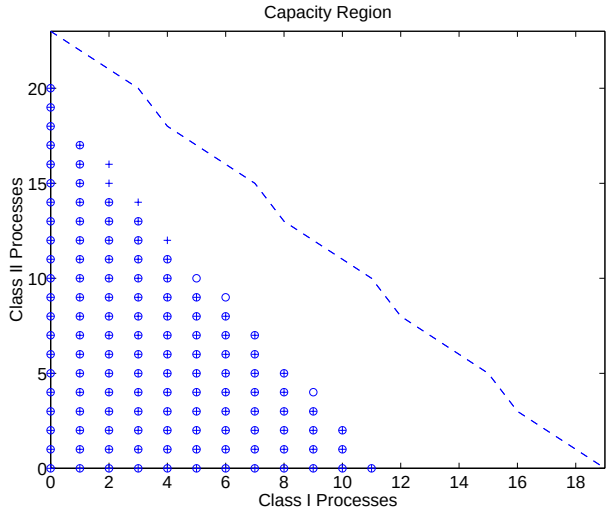

Fig. 6. Illustration for Example 3

V. Conclusion

In this paper we considered a finite buffer fluid queue with a superposition of heterogeneous heavy-tailed On-Off sources. Explicit and asymptotically exact formulas were derived for overflow probability and loss rate in the case when residual on periods are intermediately regularly varying. The formulas were further validated with simulation experiments. Their accuracy and low computational complexity makes them valuable tools for efficient computation of capacity regions. In addition, the results provide important insight into qualitative tradeoffs between the overflow probability, offered traffic load, available capacity and buffer space. Overall, they render a new set of tools for designing and provisioning of networks that will carry heavy-tailed traffic streams.

Appendices

A. Heavy-tailed distributions

The appendix contains a brief introduction to heavy-tailed and subexponential distributions.

First, we introduce a family of long-tailed distribution functions. This is the largest operational class of heavy-tailed distributions. Let X be a random variable with distribution function (d.f.) F.

Definition 2: A nonnegative random variable X (or d.f. F) is called long-tailed $X \in \mathcal{L}$ ($F \in \mathcal{L}$) if

$$\lim_{x\to\infty} \frac{1-F(x-y)}{1-F(x)} = 1, \quad \forall y \in \mathbb{R}.$$

The following class of heavy-tailed distributions was introduced by Chistyakov [31].

Definition 3: A nonnegative random variable X (or d.f. F) is called subexponential $X \in \mathcal{S}$ ($F \in \mathcal{S}$) if

$$\lim_{x\to\infty} \frac{1-F^{2*}(x)}{1-F(x)} = 2,$$

where F^{2*} denotes the 2-nd convolution of F with itself, i.e., $F^{2*}(x) = \int_{[0,\infty)} F(x-y)F(dy)$.

It is well known that $\mathcal{S} \subset \mathcal{L}$ [32]. A recent survey on subexponential distributions can be found in [33]. The class of intermediately regularly varying distributions $\mathcal{IR}$ is a subclass of $\mathcal{S}$.

Definition 4: A nonnegative random variable X (or d.f. F) is called intermediately regularly varying $X \in \mathcal{IR}$ ($F \in \mathcal{IR}$) if

$$\lim_{\eta\uparrow 1} \limsup_{x\to\infty} \frac{1-F(\eta x)}{1-F(x)} = 1.$$

Regularly varying distributions $\mathcal{R}_\alpha$, which contain Pareto distribution, are the best known examples from $\mathcal{IR}$ ($\mathcal{R}_\alpha \subset \mathcal{IR}$).

Definition 5: A nonnegative random variable X (or its d.f. F) is called regularly varying with index α, $X \in \mathcal{R}_\alpha$ ($F \in \mathcal{R}_\alpha$) if

$$F(x) = 1 - \frac{l(x)}{x^\alpha}, \quad \alpha \geq 0,$$

where $l(x) : \mathbb{R}_+ \to \mathbb{R}_+$ is a function of slow variation, i.e., $\lim_{x\to\infty} l(\eta x)/l(x) = 1, \ \eta > 1$.

The following two basic lemmas on $\mathcal{IR}$ distributions are useful for analysis.

Lemma 3: Let $F \in \mathcal{IR}$ and $\eta \in (0,1)$, then

$$\sup_{x\in(0,\infty]} \frac{1-F(\eta x)}{1-F(x)} < \infty.$$

For any bounded nondecreasing function F we say that $F \in \mathcal{IR}$ if it satisfies Definition 4. Then, the following lemma follows directly from the definition.

Lemma 4: If $F_1, F_2 \in \mathcal{IR}$, then
(i) $F_1 F_2 \in \mathcal{IR}$,
(ii) $w_1 F_1 + w_2 F_2 \in \mathcal{IR}$, for $w_1, w_2 > 0$.

B. Proofs

Proof of Proposition 1: Let $0 = T_0 < T_1 < T_2 \cdots < T_m < T_{m+1} \cdots$ (a.s.) be the jump points in $A(t)$. Then, by assumption on the initial conditions and (2), the statement holds for any $t \in [0, T_1]$

$$Q_A^{B,c}(t) \leq \left(\sum_{n=1}^N Q_{A_n}^{B,c_n}(0) + t(A_n(0) - c_n) \right)^+ \wedge B$$
$$\leq \sum_{n=1}^N \left(Q_{A_n}^{B,c_n}(0) + t(A_n(0) - c_n) \right)^+ \wedge B = \sum_{n=1}^N Q_{A_n}^{B,c}(t),$$

where the last inequality follows from

$$\left(\sum_{n=1}^{N} x_n\right)^+ \wedge B \le \left(\sum_{n=1}^{N} x_n^+\right) \wedge B \le \sum_{n=1}^{N} x_n^+ \wedge B. \quad (14)$$

Now, assume that the proposition holds for any $t \in [0, T_m]$, $m \ge 1$. Hence, by this inductive assumption, (2) and (14), for any $t \in (T_m, T_{m+1}]$

$$\begin{aligned} Q_A^{B,c}(t) &\le \left(\sum_{n=1}^{N}\Big(Q_{A_n}^{B,c_n}(T_m) + (t-T_m)(A_n(T_m)-c_n)\Big)\right)^+ \wedge B \\ &\le \sum_{n=1}^{N} Q_{A_n}^{B,c_n}(t) \end{aligned}$$

and, therefore, it holds for all t. This concludes the proof. ■

Proof of Proposition 2: The proof is by induction and very similar to the proof of Proposition 1. ¿From (2) for all $t \in (T_n, T_{n+1}]$

$$Q_A^{B,c}(t) \ge \min\left(\Big(Q_A^{B,c}(T_n) + (t-T_n)(A(T_n)-c)\Big), B\right),$$

and, therefore,

$$B - Q_A^{B,c}(t) \le \Big(B - Q_A^{B,c}(T_n) + (t-T_n)(c - A(T_n))\Big)^+.$$

Hence, the preceding inequality and the same arguments as in the proof of Proposition 1 imply the statement of the lemma. ■

Proof of Proposition 4: Using sample path arguments it is easy to show that $Q^{B,c}$ is stochastically dominated by $Q^{\infty,c}$, and therefore

$$0 \le \frac{\mathbb{P}^m[Q^{B,c_1} \ge (1-\epsilon)B]}{\mathbb{P}^l[Q^{B,c_2} \ge \epsilon B]} \le \frac{\mathbb{P}^m[Q^{\infty,c_1} \ge (1-\epsilon)B]}{\mathbb{P}^l[Q^{B,c_2} = B]}.$$

Next, Proposition 3 and Theorem 1 yield

$$\begin{aligned} \limsup_{B\to\infty} \frac{\mathbb{P}^m[Q^{\infty,c_1} \ge (1-\epsilon)B]}{\mathbb{P}^l[Q^{B,c_2} = B]} &\le \limsup_{B\to\infty} \frac{K^m \mathbb{P}^m[\tau_r^{on} > \frac{(1-\epsilon)B}{r-c}]}{p_{on}^l \mathbb{P}^l[\tau_r^{on} > \frac{B}{r-c}]} \\ &\le M \limsup_{B\to\infty} \mathbb{P}^{m-l}[\tau_r^{on} > B] = 0, \end{aligned}$$

where $M < \infty$; the last inequality is implied by Lemma 3. ■

REFERENCES

[1] M. Rubinovitch, "The output of a buffered data communication system," *Stochastic Processes and their Applications*, vol. 1, pp. 375–380, 1973.

[2] J. W. Cohen, "Superimposed renewal processes and storage with gradual input," *Stochastic Processes and their Applications*, vol. 2, pp. 31–58, 1974.

[3] D. Anick, D. Mitra, and M. M. Sondhi, "Stochastic theory of a data handling system with multiple sources," *Bell Syst. Techn. J.*, vol. 61, pp. 1871–1894, 1982.

[4] P. R. Jelenković and A. A. Lazar, "Asymptotic results for multiplexing subexponential on-off processes," *Advances in Applied Probability*, vol. 31, no. 2, June 1999.

[5] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of Ethernet traffic," in *SIGCOMM'93*, 1993, pp. 183–193.

[6] D. P. Heyman and T. V. Lakshman, "Source models for VBR broadcast-video traffic," *IEEE/ACM Transactions on Networking*, vol. 4, pp. 40–48, February 1996.

[7] P. R. Jelenković, A. A. Lazar, and N. Semret, "The effect of multiple time scales and subexponentiality of MPEG video streams on queueing behavior," *IEEE Journal on Selected Areas in Communications*, vol. 15, no. 6, pp. 1052–1071, Aug. 1997.

[8] K. R. Krishnan and G. Meempat, "Long-range dependence in VBR video streams and atm traffic engineering," *Performance Evaluation*, vol. 30, pp. 46–56, 1997.

[9] M. Crovella and A. Bestavros, "Self-similarity in world wide web traffic: Evidence and possible causes," *IEEE/ACM Transactions on Networking*, vol. 5, no. 6, pp. 835–846, 1997.

[10] D. Heath, S. Resnick, and G. Samorodnitsky, "Heavy tails and long range dependence in on/off processes and associated fluid models," *Mathematics of Operations Research*, vol. 23, pp. 145–165, 1998.

[11] O. J. Boxma, "Regular variation in a multi-source fluid queue," in *ITC 15*, Washington, D.C., USA, June 1997, pp. 391–402.

[12] T. Rolski, S. Schlegel, and V. Schmidt, "Asymptotics of palm-stationary buffer content distribution in fluid flow queues," *Advances in Applied Probability*, vol. 31, no. 1, pp. 235–253, 1999.

[13] R. Agrawal, A. M. Makowski, and P. Nain, "On a reduced load equivalence for fluid queues under subexponentiality," *Queueing Systems*, vol. 33, no. 1-3, pp. 5–41, 1999.

[14] M. Parulekar and A. M. Makowski, "Tail probabilitites for M/G/∞ input processes (I): preliminary asymptotices," *Queueing Systems*, vol. 27, pp. 271–296, 1997.

[15] S. Vamvakos and V. Anantharam, "On the depature process of a leaky bucket system with long-range dependent input traffic," *Queueing Systems*, vol. 28, pp. 191–214, 1998.

[16] N. G. Duffield, "Queueing at large resources driven by long-tailed M/G/∞-modulated processes," *Queueing Systems*, vol. 28, no. 1-3, pp. 245–266, 1998.

[17] D. Heath, S. Resnick, and G. Samorodnitsky, "How system performance is affected by the interplay of averages in a fluid queue with long range dependence induced by heavy tails," *Annals of Applied Probability*, pp. 352–375, 1999.

[18] S. Resnick and G. Samorodnitsky, "Steady state distribution of the buffer content for M/G/∞ input fluid queues," preprint, 1999.

[19] P. Jelenković, "On the asymptotic behavior of a fluid queue with a heavy-tailed M/G/∞ arrival process," preprint, June 2000.

[20] Z. Liu, P. Nain, D. Towsley, and Z.-L. Zhang, "Asymptotic behavior of a multiplexer fed by a long-range dependent process," *Journal of Applied Probability*, vol. 36, no. 1, pp. 105–118, March 1999.

[21] G. L. Choudhury and W. Whitt, "Long-tail buffer-content distributions in broadband networks," *Performance Evaluation*, vol. 30, pp. 177–190, 1997.

[22] V. Dumas and A. Simonian, "Asymptotic bounds for the fluid queue fed by subexponential on/off sources," *Advances in Applied Probability*, vol. 32, pp. 224–255, 2000.

[23] P. R. Jelenković, "Subexponential loss rates in a GI/GI/1 queue with applications," *Queueing Systems*, vol. 33, pp. 91–123, 1999.

[24] B. Zwart, "A fluid queue with a finite buffer and subexponential input," *Advances in Applied Probability*, vol. 32, no. 1, pp. 221–243, 2000.

[25] B. Zwart, S. Borst, and M. Mandjes, "Exact asymptotics for fluid queues fed by multiple heavy-tailed on-off flows," Preprint, June 2000.

[26] R. M. Loynes, "The stability of a queue with non-independent inter-arrival and service times," *Proc. Cambridge Philos Soc.*, vol. 58, pp. 497–520, 1962.

[27] P. Jelenković and P. Momčilović, "Asymptotic loss probability in a finite buffer fluid queue with hetrogeneous heavy-tailed on-off processes," submitted for publication, 2000.

[28] N. Likhanov and R. Mazumdar, "Cell loss asymptotics in buffers fed by heterogeneous long-tailed sources," in *INFOCOM 2000*, 2000, pp. 173–180.

[29] P. V. Glynn and W. Whitt, "Logarithmic asymptotics for steady-state tail probabilities in a single-server queue," in *Studies in Applied Probability*, J. Galambos and J. Gani, Eds., vol. 31A (special issue of *J. of Appl. Prob.*), pp. 131–156. Applied Probability Trust, Sheffield, England, 1994.

[30] A. Weiss and A. Shwartz, *Large Deviations for Performance Analysis: Queues, Communications, and Computing*, New York: Chapman & Hall, 1995.

[31] V. P. Chistyakov, "A theorem on sums of independent positive random variables and its application to branching random processes," *Theor. Probab. Appl.*, vol. 9, pp. 640–648, 1964.

[32] K. B. Athreya and P. E. Ney, *Branching Processes*, Springer-Verlag, 1972.

[33] C. M. Goldie and C. Klüppelberg, "Subexponential distributions," in *A Practical Guide to Heavy Tails: Statistical Techniques for Analysing Heavy Tailed Distributions*, M.S. Taqqu R. Adler, R. Feldman, Ed., pp. 435–459. Birkhäuser, Boston, 1998.