The Annals of Applied Probability
2003, Vol. 13, No. 2, 576–603

ASYMPTOTIC LOSS PROBABILITY IN A FINITE BUFFER FLUID QUEUE WITH HETEROGENEOUS HEAVY-TAILED ON–OFF PROCESSES[1]

BY PREDRAG JELENKOVIĆ AND PETAR MOMČILOVIĆ

Columbia University

Consider a fluid queue with a finite buffer B and capacity c fed by a superposition of N independent On–Off processes. An On–Off process consists of a sequence of alternating independent periods of activity and silence. Successive periods of activity, as well as silence, are identically distributed. The process is active with probability p and during its activity period produces fluid at constant rate r. For this queueing system, under the assumption that the excess activity periods are intermediately regularly varying, we derive explicit and asymptotically exact formulas for approximating the stationary overflow probability and loss rate. In the case of homogeneous processes with excess activity periods equal in distribution to τ^e, the queue loss rate is asymptotically, as $B \to \infty$, equal to

$$\Lambda^B = (r_0 - c)\binom{N}{m}\left(p\mathbb{P}\left[\tau^e > \frac{B}{r_0 - c}\right]\right)^m (1 + o(1)),$$

where m is the smallest integer greater than $(c - N\rho)/(r - \rho)$, $r_0 = mr + (N - m)\rho$, $\rho = rp$ and $N\rho < c$; the results require a mild technical assumption that $(c - N\rho)/(r - \rho)$ is not an integer. The analyzed queueing system represents a standard model of resource sharing in telecommunication networks. The derived asymptotic results are shown to provide accurate approximations to simulation experiments. Furthermore, the results offer insight into qualitative tradeoffs between the overflow probability, offered traffic load, capacity and buffer space.

1. 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 of sharing common resources are counterbalanced with potential increase 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 the 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

Received September 2000; revised January 2002.

[1] Supported by NSF PECASE Award 9875156.

AMS 2000 subject classifications. Primary 60K25; secondary 60F99.

Key words and phrases. Fluid queue, On–Off processes, subexponential distributions, statistical multiplexing, long-range dependence.

multiplexer is a single server queue with a constant capacity and finite buffer fed by a superposition of 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 [2, 7, 32].

Most of the early work on the multiplexing problem focused on On–Off processes with exponentially distributed On and Off periods (e.g., see [2]). However, repeated empirical measurements in modern telecommunications networks demonstrate the presence of heavy-tailed subexponential characteristics in network traffic streams. Early discoveries of the self-similar nature of Ethernet traffic were reported in [23]. Long-range dependence and subexponential properties of variable bit rate (VBR) video streams (e.g., MPEG) were explored in [16, 20, 22]. Evidence and possible causes of heavy-tailed characteristics in World Wide Web traffic were presented in [9]. In this paper, we provide an additional confirmation of the existence of heavy tails in network traffic. We have measured the distribution of file sizes on five file servers in COMET laboratory 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 well 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.

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 [14]. This stems from the fact that a superposition of renewal processes, in general, is not a renewal process. An intermediate case of multiplexing a single long-tailed arrival sequence with

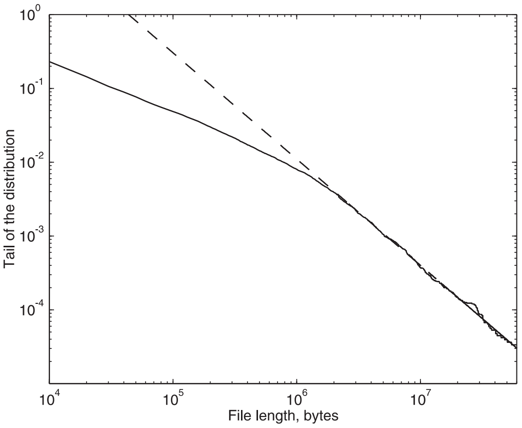

FIG. 1. *Log–log plot of the empirical distribution of file sizes on five file servers in COMET laboratory at Columbia University: the tail of the empirical distribution (solid line) is well matched by a Pareto distribution $ax^{-\alpha}$ with $\alpha = 1.44$ (dashed line).*

exponential processes was investigated in [1, 4, 19, 31]. An infinite limit of On–Off processes, the so-called $M/G/\infty$ process, represents an instance of an analytically tractable model since it has both a renewal and Poisson structure. Recent results and additional references on both fluid and discrete time queues with $M/G/\infty$ arrival processes can be found in [4, 10, 15, 18, 19, 25, 28, 30, 33].

On the other hand, the understanding of multiplexing a finite number of heavy-tailed On–Off arrival processes is quite limited; for general bounds see [6, 11]. In this paper we derive explicit and asymptotically exact results for 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 [17]. Very recently the complementary results for the infinite buffer model were derived in [35].

The rest of the paper is organized as follows. In Section 2 we define the fluid model and introduce the preliminary results. The main results of this paper, Theorems 3.1 and 3.2, are presented in Section 3. In Section 4 we illustrate the accuracy of our results with simulation experiments. Concluding remarks are stated in Section 5. The last section contains some of the more technical proofs.

2. Preliminary results.

2.1. *Fluid queue definition and sample path bounds.* Consider a fluid queue with a constant capacity c, finite buffer B and arrival process $A(t)$. Informally, at time t, fluid is arriving 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 only consider right-continuous piecewise constant processes $A(t)$ with almost surely (a.s.) increasing jump times $T_0 = 0 < T_1 < T_2 < \cdots$. Then, for any initial value $Q^B(0)$ the evolution of $Q^B(t)$ is given by

$$Q^B(t) = \big(Q^B(T_n) + (t - T_n)\big(A(T_n) - c\big)\big)^+ \wedge B, \qquad t \in (T_n, T_{n+1}],\ n \geq 0, \tag{2.1}$$

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

In the case of $A(t)$, that is, $\{(T_{n+1} - T_n), A(T_n)\}$, being stationary and ergodic, and $\mathbb{E}A(t) < c$, by using Loynes's construction [26], one can show that recursion (2.1) has a unique stationary and ergodic solution. Furthermore, for all initial conditions $Q^B(0)$, the distribution of $Q^B(t)$ converges to that stationary solution as $t \to \infty$. Unless otherwise indicated, we assume throughout the paper that all arrival processes are stationary, ergodic and that 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.

The main focus of this paper is the asymptotic evaluation, as $B \to \infty$, of the *overflow probability* $\mathbb{P}[Q^B \geq B - K]$, for finite K, and long time average *loss* *rate* Λ^B defined by

$$\Lambda^B \triangleq \lim_{t\to\infty} \frac{1}{t}\int_0^t \lambda^B(u)\,du,$$

where $\lambda^B(t) \triangleq (A(t)-c)\,\mathbf{1}\{Q^B(t) = B\}$ indicates the rate at which the buffer is overflowing at time t. We define the *loss probability* $P^B = \Lambda^B/\mathbb{E}A$ as the long time average fraction of fluid that is lost. Since there is a one-to-one correspondence between the loss rate and loss probability, we use those two terms interchangeably. An equivalent representation of Λ^B, which will be used for computational purposes, is $\Lambda^B = \mathbb{E}\lambda^B(t)$. Similarly, the notation $\Lambda_A^{B,c} \equiv \Lambda^B$ will be used to mark the explicit dependence of Λ^B on $A(t)$ and c.

Next, we prove two useful sample path bounds. The first bound formalizes an intuitively expected notion that multiplexing reduces the aggregate queueing workload. Let $A_n(t)$ and c_n, $1 \le n \le N$, be arrival processes and service rates, respectively, with $A(t) = \sum_{n=1}^N A_n(t)$ and $c = \sum_{n=1}^N c_n$.

PROPOSITION 2.1. *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. Let $0 = T_0 < T_1 < T_2 < \cdots$ a.s. be the jump points in $A(t)$. Then, by the assumption and (2.1), the statement of the proposition holds for any $t \in [0, T_1]$:

$$\begin{aligned} Q_A^{B,c}(t) &\le \left(\sum_{n=1}^N \bigl(Q_{A_n}^{B,c_n}(0) + t\bigl(A_n(0)-c_n\bigr)\bigr)\right)^+ \wedge B \\ &\le \sum_{n=1}^N \bigl(Q_{A_n}^{B,c_n}(0) + t\bigl(A_n(0)-c_n\bigr)\bigr)^+ \wedge B = \sum_{n=1}^N Q_{A_n}^{B,c_n}(t), \end{aligned}$$

where the last inequality follows from

$$\left(\sum_{n=1}^N x_n\right)^+ \wedge B \le \left(\sum_{n=1}^N x_n^+\right) \wedge B \le \sum_{n=1}^N x_n^+ \wedge B. \tag{2.2}$$

Now, assume that the proposition holds for any $t \in [0, T_k]$, $k \ge 1$. Hence, by the inductive assumption, (2.1) and (2.2), for any $t \in (T_k, T_{k+1}]$,

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

and, therefore, the result holds for all $t \ge 0$. □

Next, we consider a stochastic process $Q_c^{\infty,A}(t)$ defined by the initial condition $Q_c^{\infty,A}(0)$ and

$$Q_c^{\infty,A}(t) = \big(Q_c^{\infty,A}(T_n) + (t - T_n)\big(c - A(T_n)\big)\big)^+, \qquad t \in (T_n, T_{n+1}],\ n \geq 0. \tag{2.3}$$

Note that $Q_c^{\infty,A}(t)$ corresponds to an infinite buffer queueing process with constant arrival rate c and service rate $A(t)$. We use $Q_c^{\infty,A}(t)$ to upper bound the amount of free buffer space, $B - Q^B(t)$, in the original system defined by (2.1).

LEMMA 2.1. *If $B - Q_A^{B,c}(t) \leq Q_c^{\infty,A}(t)$ for $t = 0$, then the inequality holds for all $t \geq 0$.*

PROOF. The proof is by induction and very similar to the proof of Proposition 2.1. From (2.1) for all $t \in (T_n, T_{n+1}]$, $n \geq 0$,

$$Q_A^{B,c}(t) \geq \big(Q_A^{B,c}(T_n) + (t - T_n)\big(A(T_n) - c\big)\big) \wedge B,$$

and, therefore,

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

The preceding inequality and the same arguments used in the proof of Proposition 2.1 imply the statement of the lemma. □

2.2. *Fluid queue with a single On–Off arrival process.* The results of this subsection characterize the asymptotic behavior of a finite buffer fluid queue fed by a single On–Off process. These results will be used for deriving our main theorems in the subsequent section.

A stationary On–Off process $A(t)$ consists of a sequence of alternating independent On and Off periods. During the corresponding On and Off periods the process is equal to $A(t) = r$ and $A(t) = 0$. Successive On as well as Off periods are identically distributed and equal in distribution to τ and ν, respectively. Random variables τ and ν have finite first moments and the process is in On state with probability $p = \mathbb{P}[A(t) = r] = \mathbb{E}\tau/(\mathbb{E}\tau + \mathbb{E}\nu)$. The average rate of the process is $\rho = rp$. For a detailed construction of such a process see, for example, [11].

Excess (or residual) random variables play an important role in the analysis of renewal processes. For a nonnegative random variable X with finite mean, the excess distribution F^e is defined by $F^e(x) = (\mathbb{E}X)^{-1} \int_0^x \mathbb{P}[X > u]\,du$, $x \geq 0$. A random variable X^e with distribution F^e is called the excess variable of X.

Throughout the paper, for any two real functions $f(x)$ and $g(x)$, we use the customary notation $f(x) \sim g(x)$ as $x \to \infty$ to denote $\lim_{x\to\infty} f(x)/g(x) = 1$. Definitions of the classes $\mathcal{L}$, $\mathcal{S}$, $\mathcal{IR}$ and $\mathcal{R}_\alpha$ of heavy-tailed distributions can be found in the Appendix.

The following proposition provides the asymptotic characterization of the overflow probability when the excess On periods are subexponential.

PROPOSITION 2.2. *If $r > c > \rho$ and $\tau^e \in \mathcal{S}$, then as $B \to \infty$,*

$$\mathbb{P}[Q^B = B] \sim p\,\mathbb{P}\left[\tau^e > \frac{B}{r-c}\right].$$

PROOF. In [17] it was shown that $\Lambda^B \sim p(r-c)\mathbb{P}[\tau^e > B/(r-c)]$ as $B \to \infty$. Since $\Lambda^B = \mathbb{E}[(r-c)\mathbf{1}\{Q^B = B\}] = (r-c)\mathbb{P}[Q^B = B]$ the statement holds. □

The next result characterizes the workload $Q^\infty \equiv Q_A^{\infty,c}$ in an infinite buffer system.

THEOREM 2.1 [19]. *If $r > c > \rho$ and $\tau^e \in \mathcal{S}$, then as $B \to \infty$,*

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

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 2.3. *If $r > c_i > \rho$, $i = 1, 2$, and $\tau^e \in \mathcal{IR}$, then for $1 > \varepsilon > 0$,*

$$\limsup_{B\to\infty} \frac{\mathbb{P}[Q_A^{B,c_1} \ge (1-\varepsilon)B]}{\mathbb{P}[Q_A^{B,c_2} \ge \varepsilon B]} < \infty.$$

PROOF. Using sample path arguments it is easy to show that $Q_A^{B,c}$ is stochastically dominated by $Q_A^{\infty,c}$, and therefore

$$\frac{\mathbb{P}[Q_A^{B,c_1} \ge (1-\varepsilon)B]}{\mathbb{P}[Q_A^{B,c_2} \ge \varepsilon B]} \le \frac{\mathbb{P}[Q_A^{\infty,c_1} \ge (1-\varepsilon)B]}{\mathbb{P}[Q_A^{B,c_2} = B]}.$$

Next, Proposition 2.2 and Theorem 2.1 yield

$$\begin{aligned}&\limsup_{B\to\infty} \frac{\mathbb{P}[Q_A^{\infty,c_1} \ge (1-\varepsilon)B]}{\mathbb{P}[Q_A^{B,c_2} = B]}\\ &\qquad \le \frac{(1-p)\rho}{(c-\rho)p}\limsup_{B\to\infty}\frac{\mathbb{P}[\tau^e > (1-\varepsilon)B/(r-c)]}{\mathbb{P}[\tau^e > B/(r-c)]} < \infty,\end{aligned}$$

where the last inequality is implied by Lemma A.1 of the Appendix. □

The last proposition is the main technical result of this section. In order to alleviate the reading process we postpone the proof until Section 6.

PROPOSITION 2.4. *If $r > c > \rho$ and $\tau^e \in \mathcal{IR}$, then*

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

3. Main results. This section contains the main results of this paper stated in Theorems 3.1 and 3.2. The theorems describe the asymptotic behavior of a finite buffer fluid queue fed by N independent On–Off processes. Without loss of generality, assume that they belong to $M \le N$ different classes with class i containing n_i statistically identical On–Off processes, $\sum_{i=1}^{M} n_i = N$. The processes are enumerated as $A_{ij}(t)$, $1 \le i \le M$, $1 \le j \le n_i$, and the aggregate arrival process is denoted by $A(t) = \sum_{i=1}^{M} \sum_{j=1}^{n_i} A_{ij}(t)$. Process $A_{ij}(t)$ is the jth process of class i with On periods equal in distribution to τ_{ij}. Its peak rate, average rate and probability of being On are equal to r_i, ρ_i and p_i, respectively. Random variables τ_{ij}, $1 \le j \le n_i$, are equal in distribution to τ_i. For convenience we define vectors $\mathbf{r} = (r_1, \dots, r_M)$, $\boldsymbol{\rho} = (\rho_1, \dots, \rho_M)$ and $\mathbf{n} = (n_1, \dots, n_M)$. To distinguish between scalar and vector quantities, vectors are denoted with bold letters.

Our proofs require the following minor technical assumption. Similar assumptions can be found in [15, 24] and, most recently, in [35]. For vectors $\mathbf{x}$ and $\mathbf{y}$ by $\mathbf{x} \cdot \mathbf{y} = x_1 y_1 + \cdots + x_M y_M$ we denote their scalar product.

ASSUMPTION 3.1. The capacity of the queueing system satisfies $\mathbf{n} \cdot \mathbf{r} > c > \mathbf{n} \cdot \boldsymbol{\rho}$ and

$$c \notin \left\{ \mathbf{m} \cdot \mathbf{r} + (\mathbf{n} - \mathbf{m}) \cdot \boldsymbol{\rho} : \mathbf{m} \in \bigotimes_{i=1}^{M} [0, n_i] \right\}.$$

REMARK 3.1. (i) The first part of the assumption states that the queue is stable and that overflows are possible.

(ii) If the second part of the 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. The assumption ensures that the queue is not critically stable during periods of time when some of the processes have long On periods.

Before stating and proving our main results we introduce two preparatory lemmas. The first lemma derives an asymptotic expression for the overflow probability in the case when all processes need to be in the active state for a long period of time in order to have a buffer overflow. Throughout the paper we use $\mathbb{P}^m[\cdot]$ to denote $(\mathbb{P}[\cdot])^m$.

LEMMA 3.1. *If* $\mathbf{n} \cdot \mathbf{r} - r_i + \rho_i < c < \mathbf{n} \cdot \mathbf{r}$ *for all* $1 \le i \le M$, *then for all* $B \ge 0$ *and* $0 \le \varepsilon \le 1$,

$$\prod_{i=1}^{M} p_i^{n_i} \mathbb{P}^{n_i}\left[\tau_i^e > \frac{\varepsilon B}{\mathbf{n} \cdot \mathbf{r} - c}\right] \le \mathbb{P}[Q_A^{B,c} \ge \varepsilon B] \le \prod_{i=1}^{M} \mathbb{P}^{n_i}[Q_{A_{i1}}^{B, c - \mathbf{n} \cdot \mathbf{r} + r_i} \ge \varepsilon B].$$

If in addition $\tau_i^e \in \mathcal{S}$ for $1 \le i \le M$, then as $B \to \infty$,

$$\mathbb{P}[Q_A^{B,c} = B] \sim \prod_{i=1}^{M} p_i^{n_i} \mathbb{P}^{n_i}\left[\tau_i^e > \frac{B}{\mathbf{n}\cdot\mathbf{r} - c}\right].$$

PROOF. Assume that at time $t = 0$ all the considered queues are empty. For all $1 \le i \le M$, $1 \le j \le n_i$, Proposition 2.1 yields

$$\begin{aligned} Q_A^{B,c}(t) &\le Q_{A_{ij}}^{B,c-\mathbf{n}\cdot\mathbf{r}+r_i}(t) + Q_{A-A_{ij}}^{B,\mathbf{n}\cdot\mathbf{r}-r_i}(t) \\ &= Q_{A_{ij}}^{B,c-\mathbf{n}\cdot\mathbf{r}+r_i}(t), \end{aligned} \tag{3.1}$$

where the equality follows from the fact that $A(t) - A_{ij}(t) \le \mathbf{n}\cdot\mathbf{r} - r_i$ for all t and, therefore, $Q_{A-A_{ij}}^{B,\mathbf{n}\cdot\mathbf{r}-r_i}(t) \equiv 0$, $t \ge 0$. Since (3.1) holds for all i, j, then

$$Q_A^{B,c}(t) \le \bigwedge_{i,j} Q_{A_{ij}}^{B,c-\mathbf{n}\cdot\mathbf{r}+r_i}(t),$$

which, by applying the operator $\mathbb{P}[\cdot \ge \varepsilon B]$, using the independence of A_{ij} and passing $t \to \infty$, yields in stationarity

$$\mathbb{P}[Q_A^{B,c} \ge \varepsilon B] \le \prod_{i=1}^{M} \mathbb{P}^{n_i}[Q_{A_{i1}}^{B,c-\mathbf{n}\cdot\mathbf{r}+r_i} \ge \varepsilon B].$$

Obtaining the lower bound is straightforward from evaluating the system in stationarity at (say) $t = 0$; for simplicity the time index is omitted:

$$\begin{aligned} \mathbb{P}[Q_A^{B,c} \ge \varepsilon B] &\ge \mathbb{P}\left[Q_A^{B,c} \ge \varepsilon B, \bigcap_{i=1}^{M}\bigcap_{j=1}^{n_i}\left\{A_{ij} = r_i, \tau_{ij}^e > \frac{\varepsilon B}{\mathbf{n}\cdot\mathbf{r} - c}\right\}\right] \\ &= \mathbb{P}\left[\bigcap_{i=1}^{M}\bigcap_{j=1}^{n_i}\left\{A_{ij} = r_i, \tau_{ij}^e > \frac{\varepsilon B}{\mathbf{n}\cdot\mathbf{r} - c}\right\}\right] \\ &= \prod_{i=1}^{M} p_i^{n_i} \mathbb{P}^{n_i}\left[\tau_i^e > \frac{\varepsilon B}{\mathbf{n}\cdot\mathbf{r} - c}\right]. \end{aligned}$$

By setting $\varepsilon = 1$ in the preceding upper and lower bounds and combining it with Proposition 2.2, we obtain the second statement of the proposition. □

To state our second preliminary lemma and the main results, we need to introduce additional notation. 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 $\mathbf{e} = (\mathbf{e}_1, \dots, \mathbf{e}_M)$, where $\mathbf{e}_i = (e_{i1}, \dots, e_{in_i}) \in [0, 1]^{n_i}$. Let $|\mathbf{e}_i| = \sum_{j=1}^{n_j} e_{ij}$ and $|\mathbf{e}| = (|\mathbf{e}_1|, \dots, |\mathbf{e}_M|)$. Note that if $\mathbf{e} \in \mathcal{E}_*$, then $|\mathbf{e}| \in \mathcal{E}$.

DEFINITION 3.1. The minimum overflow set is defined as

$$\mathcal{O} \triangleq \{\mathbf{m} \in \mathcal{E} : c < \mathbf{m} \cdot \mathbf{r} + (\mathbf{n} - \mathbf{m}) \cdot \boldsymbol{\rho} < c + r_i - \rho_i,\ \forall\, i : m_i > 0\},$$

and the detailed minimum overflow set $\mathcal{O}_* \triangleq \{\mathbf{e} \in \mathcal{E}_* : |\mathbf{e}| \in \mathcal{O}\}$.

REMARK 3.2. (i) Informally, the motivation behind this definition comes from the fact that only a few On–Off processes with long On periods are causing the most likely buffer overflows, while the remaining processes exhibit their average behavior. Hence, an element of $\mathcal{O}$ indicates how many processes from each class need to have long On periods in order for a buffer overflow to occur. Correspondingly, the detailed set $\mathcal{O}_*$ contains binary vectors which denote particular overflow scenarios.

(ii) The definition of $\mathcal{O}_*$ is analogous to the definition of the minimal set in [11].

Similarly, we define an underflow set $\mathcal{U}$ of the combinations that do not cause an overflow

$$\mathcal{U} \triangleq \{\mathbf{m} \in \mathcal{E} : \mathbf{m} \cdot \mathbf{r} + (\mathbf{n} - \mathbf{m}) \cdot \boldsymbol{\rho} < c\}$$

and the corresponding detailed underflow set $\mathcal{U}_* \triangleq \{\mathbf{e} \in \mathcal{E}_* : |\mathbf{e}| \in \mathcal{U}\}$.

Next, let $\{X_{ij}\}$ be a set of independent random variables. The variables with the same first subscript are equal in distribution. For every element $\mathbf{e} \in \mathcal{E}_*$, let us define the following partial sum:

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

At this point, we are ready to state our last preparatory lemma.

LEMMA 3.2. *If $X_{i1} \in \mathcal{IR}$ for all $1 \le i \le M$, then as $x \to \infty$,*

$$\mathbb{P}\left[\bigwedge_{\mathbf{e} \in \mathcal{O}_* \cup \mathcal{U}_*} S_{\mathbf{e}} > x\right] = o\left(\sum_{\mathbf{m} \in \mathcal{O}} \prod_{i=1}^{M} \mathbb{P}^{m_i}[X_{i1} > x]\right).$$

PROOF. Define $D_i(x)$ for $1 \le i \le M$ as the number of indices j satisfying $X_{ij} > x/N$; that is,

$$D_i(x) \triangleq \sum_{j=1}^{n_i} \mathbf{1}\left\{X_{ij} > \frac{x}{N}\right\}.$$

Observe that, for all $\mathbf{m} \in \mathcal{O} \cup \mathcal{U}$,

$$\left\{\bigcap_{i=1}^{M} \{D_i(x) = m_i\}\right\} \bigcap \left\{\bigwedge_{\mathbf{e} \in \mathcal{O}_* \cup \mathcal{U}_*} S_{\mathbf{e}} > x\right\} = \varnothing, \tag{3.2}$$

since, on the preceding events, we can define $e_{ij} = \mathbf{1}\{X_{ij} > x/N\}$ such that $|\mathbf{e}_i| = m_i$ and $\mathbf{e} \in \mathcal{O}_* \cup \mathcal{U}_*$, which, by the definition of $S_{\mathbf{e}}$, yields, for $x \geq 0$,

$$S_{\mathbf{e}} \leq \sum_{i=1}^{M} \frac{n_i - m_i}{N} x \leq x.$$

Next, (3.2) implies

$$\mathbb{P}\Bigg[\bigwedge_{\mathbf{e}\in\mathcal{O}_*\cup\mathcal{U}_*} S_{\mathbf{e}} > x\Bigg] \leq \mathbb{P}\Bigg[\bigcup_{\mathbf{m}\notin\mathcal{O}\cup\mathcal{U}} \bigcap_{i=1}^{M} \{D_i(x) = m_i\}\Bigg],$$

which, in conjunction with

$$\mathbb{P}[D_i(x) \geq m_i] = \mathbb{P}\Bigg[\bigcup_{\mathbf{d}\in[0,1]^{n_i}\,:\,|\mathbf{d}|=m_i} \; \bigcap_{j\,:\,d_j=1} \Big\{X_{ij} > \frac{x}{N}\Big\}\Bigg] \leq \binom{n_i}{m_i} \mathbb{P}^{m_i}\Big[X_{i1} > \frac{x}{N}\Big],$$

yields

$$\mathbb{P}\Bigg[\bigwedge_{\mathbf{e}\in\mathcal{O}_*\cup\mathcal{U}_*} S_{\mathbf{e}} > x\Bigg] \leq \sum_{\mathbf{m}\notin\mathcal{O}\cup\mathcal{U}} \prod_{i=1}^{M} \binom{n_i}{m_i} \mathbb{P}^{m_i}\Big[X_{i1} > \frac{x}{N}\Big].$$

The lemma follows from the preceding inequality, Lemma A.1 of the Appendix and the definitions of $\mathcal{O}$ and $\mathcal{U}$, which imply that for every $\mathbf{m} \notin \mathcal{O} \cup \mathcal{U}$ there exists $\mathbf{k} \in \mathcal{O}$ such that $m_i \geq k_i$ for all i and $m_j > k_j$ for at least one j. □

At last, we arrive at our first main result.

THEOREM 3.1. *Let* $\tau_i^e \in \mathcal{IR}$ *for* $1 \leq i \leq M$ *and*

$$\hat{P}(B) \triangleq \sum_{\mathbf{m}\in\mathcal{O}} \prod_{i=1}^{M} \binom{n_i}{m_i} p_i^{m_i} \mathbb{P}^{m_i}\Big[\tau_i^e > \frac{B}{\mathbf{m}\cdot\mathbf{r} + (\mathbf{n}-\mathbf{m})\cdot\boldsymbol{\rho} - c}\Big].$$

Then, *under Assumption* 3.1,

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

If, *in addition*, *we assume* $\mathbf{m}\cdot\mathbf{r} > c$ *for all* $\mathbf{m} \in \mathcal{O}$, *then for any* $K \geq 0$,

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

REMARK 3.3. (i) Informally, the double limit implies that $\mathbb{P}[Q_A^{B,c} \geq B-K] \approx \hat{P}(B)$ for large K and B much larger than K. Hence, the result states that the fraction of time during which the buffer is effectively 100% full is asymptotically equal to $\hat{P}(B)$.

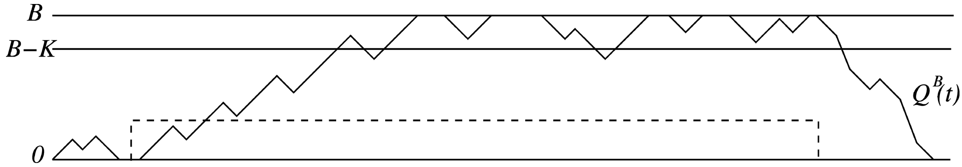

FIG. 2. *Illustration for Remark* 3.3(ii): *the long On period is shown with a dashed line.*

(ii) The heuristic for this result can be explained by the following straightforward example. Consider two i.i.d. On–Off processes with excess On periods in $\mathcal{IR}$ and $r_1 < c < r_1 + \rho_1$. 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, that is, $\int_0^t A_{12}(u)\,du \approx \rho_1 t$. During that long On period, the average amount of arriving fluid will be higher than the service rate, $r_1 + \rho_1 > c$, and the buffer will tend to fill. After the buffer fills, its content will stay close to the buffer boundary. When $r_1 < c$, 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.

(iii) 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 approximation 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 fluid is being lost.

(iv) With additional assumptions on the ratios of tails of τ_i^e the minimum overflow set $\mathcal{O}$ in the statement of the theorem can be replaced by a smaller overflow set $\mathcal{O}_0$, which asymptotically yields the same value for $\hat{P}(B)$. For example, if $\tau_i^e \in \mathcal{R}_{\alpha_i}$ (see the Appendix), then $\mathcal{O}_0 = \{\mathbf{m} \in \mathcal{O} : \boldsymbol{\alpha} \cdot \mathbf{m} = \bigwedge_{\mathbf{l}\in\mathcal{O}} \boldsymbol{\alpha} \cdot \mathbf{l}\}$.

(v) Complementary results for the infinite buffer model were recently obtained in [35].

PROOF OF THEOREM 3.1.

Upper bound. Let $A_{\mathbf{e}}$, $\mathbf{e} \in \mathcal{E}_*$, denote the sum of arrival processes A_{ij} such that $e_{ij} = 0$,

$$A_{\mathbf{e}}(t) \triangleq \sum_{i=1}^{M} \sum_{j=1}^{n_i} (1 - e_{ij}) A_{ij}(t) \tag{3.3}$$

and for $\delta > 0$ consider queues $Q^{B,c-\mathbb{E}A_{\mathbf{e}}-\delta}_{A-A_{\mathbf{e}}}$, $Q^{B,\rho_i+\delta/N}_{A_{ij}}$ assuming that they are empty at time $t = 0$. For any $\mathbf{e} \in \mathcal{O}_* \cup \mathcal{U}_*$, Proposition 2.1 yields

$$Q_A^{B,c}(t) \le Q_{A-A_{\mathbf{e}}}^{B,c-\mathbb{E}A_{\mathbf{e}}-\delta}(t) + \sum_{i=1}^{M} \sum_{j=1}^{n_i} (1-e_{ij}) Q_{A_{ij}}^{B,\rho_i+\delta/N}(t),$$

and thus

$$Q_A^{B,c}(t) \le \bigwedge_{\mathbf{e}\in\mathcal{O}_*\cup\mathcal{U}_*} \left(Q_{A-A_{\mathbf{e}}}^{B,c-\mathbb{E}A_{\mathbf{e}}-\delta}(t) + \sum_{i=1}^{M} \sum_{j=1}^{n_i} (1-e_{ij}) Q_{A_{ij}}^{B,\rho_i+\delta/N}(t) \right). \tag{3.4}$$

Next, by selecting sufficiently small δ, such that all the queues in the preceding inequality have their capacities greater than the average arrival rate, applying the operator $\mathbb{P}[\cdot \ge B - K]$ in (3.4) and then passing $t \to \infty$, we derive in stationarity

$$\begin{aligned}
&\mathbb{P}[Q_A^{B,c} \ge B - K] \\
&\quad \le \mathbb{P}\left[\bigwedge_{\mathbf{e}\in\mathcal{O}_*\cup\mathcal{U}_*} \left(Q_{A-A_{\mathbf{e}}}^{B,c-\mathbb{E}A_{\mathbf{e}}-\delta} + \sum_{i=1}^{M} \sum_{j=1}^{n_i} (1-e_{ij}) Q_{A_{ij}}^{B,\rho_i+\delta/N} \right) \ge B - K \right].
\end{aligned}$$

Now, let us select $\delta < \bigwedge_{\mathbf{e}\in\mathcal{U}_*}(c - |\mathbf{e}| \cdot \mathbf{r} - (\mathbf{n} - |\mathbf{e}|) \cdot \boldsymbol{\rho})$, such that $Q_{A-A_{\mathbf{e}}}^{B,c-\mathbb{E}A_{\mathbf{e}}-\delta} \equiv 0$ for all $\mathbf{e} \in \mathcal{U}_*$. Then the preceding inequality and union bound yield, for $0 < \varepsilon < 1$,

$$\begin{aligned}
&\mathbb{P}[Q_A^{B,c} \ge B - K] \\
&\quad \le \mathbb{P}\left[\bigcup_{\mathbf{e}\in\mathcal{O}_*} \{ Q_{A-A_{\mathbf{e}}}^{B,c-\mathbb{E}A_{\mathbf{e}}-\delta} \ge \varepsilon(B-K) \} \right] \\
&\qquad + \mathbb{P}\left[\bigwedge_{\mathbf{e}\in\mathcal{O}_*\cup\mathcal{U}_*} \sum_{i=1}^{M} \sum_{j=1}^{n_i} (1-e_{ij}) Q_{A_{ij}}^{B,\rho_i+\delta/N} \ge (1-\varepsilon)(B-K) \right] \\
&\quad \le \sum_{\mathbf{e}\in\mathcal{O}_*} \mathbb{P}[Q_{A-A_{\mathbf{e}}}^{B,c-\mathbb{E}A_{\mathbf{e}}-\delta} \ge \varepsilon(B-K)] \\
&\qquad + \mathbb{P}\left[\bigwedge_{\mathbf{e}\in\mathcal{O}_*\cup\mathcal{U}_*} \sum_{i=1}^{M} \sum_{j=1}^{n_i} (1-e_{ij}) Q_{A_{ij}}^{B,\rho_i+\delta/N} \ge (1-\varepsilon)(B-K) \right] \\
&\quad \le (1+o(1)) \sum_{\mathbf{m}\in\mathcal{O}} \prod_{i=1}^{M} \binom{n_i}{m_i} \mathbb{P}^{m_i}[Q_{A_{i1}}^{B,c_{\mathbf{m}}+r_i-\delta} \ge \varepsilon(B-K)],
\end{aligned} \tag{3.5}$$

as $B \to \infty$, where $c_{\mathbf{m}} \triangleq c - \mathbf{m} \cdot \mathbf{r} - (\mathbf{n} - \mathbf{m}) \cdot \boldsymbol{\rho}$ and the last inequality is due to Lemmas 3.1, 3.2 and Proposition 2.3. Here, by recalling that $\tau_i^e \in \mathcal{IR}$, one obtains from Propositions 2.2 and 2.4 for all $\mathbf{m} \in \mathcal{O}$ and i, such that $m_i > 0$,

$$\lim_{\delta\downarrow 0} \lim_{\varepsilon\uparrow 1} \limsup_{B\to\infty} \frac{\mathbb{P}[Q_{A_{i1}}^{B,c_{\mathbf{m}}+r_i-\delta} \ge \varepsilon(B-K)]}{\mathbb{P}[Q_{A_{i1}}^{B,c_{\mathbf{m}}+r_i} = B]} = 1,$$

which, by (3.5), Proposition 2.2 and Lemma A.3 of the Appendix yields, for any $K \geq 0$,

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

Lower bound for the first statement. The lower bound is obtained by observing the queueing system in stationarity at (say) time $t = 0$. For any $\varepsilon > 0$ and all $\mathbf{e} \in \mathcal{O}_*$, define an event indicating that all the processes $A_{ij}(t)$ with $e_{ij} = 1$ are in the active state at time $t = 0$ and their On periods have lasted for an amount of time larger than $t_\mathbf{e} \triangleq (1+\varepsilon)B/(|\mathbf{e}|\cdot\mathbf{r} + (\mathbf{n} - |\mathbf{e}|)\cdot\boldsymbol{\rho} - c)$, that is,

$$\Psi_\mathbf{e} \triangleq \bigcap_{i,j\,:\,e_{ij}=1} \{A_{ij}(0) = r_i,\ \inf\{t > 0 : A_{ij}(-t) = 0\} > t_\mathbf{e}\}. \tag{3.6}$$

We point out that $\inf\{t > 0 : A_{ij}(-t) = 0\}$ is equal in distribution to τ_i^e on event $\{A_{ij}(0) = r_i\}$. In a similar way, we define a corresponding event $\Upsilon_\mathbf{e}$ indicating that processes $A_{ij}(t)$ with $e_{ij} = 0$ do not have long On periods at time $t = 0$, that is,

$$\Upsilon_\mathbf{e} \triangleq \bigcap_{i,j\,:\,e_{ij}=0} \overline{\{A_{ij}(0) = r_i,\ \inf\{t > 0 : A_{ij}(-t) = 0\} > t_\mathbf{e}\}}.$$

Now, $Q_A^{B,c}(-t_\mathbf{e})$ being nonnegative and Proposition 2.1 imply $Q_A^{B,c}(0) \geq Q_{A_\mathbf{e}}^{B,c-|\mathbf{e}|\cdot\mathbf{r}}(0)$ on event $\Psi_\mathbf{e}$, where $Q_{A_\mathbf{e}}^{B,c-|\mathbf{e}|\cdot\mathbf{r}}(-t_\mathbf{e}) = 0$ and $A_\mathbf{e}$ is defined by (3.3). Then, since for all different $\mathbf{e} \in \mathcal{O}_*$ events $\{\Psi_\mathbf{e} \cap \Upsilon_\mathbf{e}\}$ are disjoint, by Lemma 2.1 one obtains

$$\begin{aligned}
&\mathbb{P}[Q_A^{B,c}(0) \geq B - K] \\
&\qquad \geq \sum_{\mathbf{e}\in\mathcal{O}_*} \mathbb{P}[Q_A^{B,c}(0) \geq B - K,\ \Upsilon_\mathbf{e},\ \Psi_\mathbf{e}] \\
&\qquad \geq \sum_{\mathbf{e}\in\mathcal{O}_*} \mathbb{P}[Q_{c-|\mathbf{e}|\cdot\mathbf{r}}^{\infty,A_\mathbf{e}}(0) \leq K,\ \Upsilon_\mathbf{e},\ \Psi_\mathbf{e}] \\
&\qquad \geq \left(\bigwedge_{\mathbf{e}\in\mathcal{O}_*} \mathbb{P}[Q_{c-|\mathbf{e}|\cdot\mathbf{r}}^{\infty,A_\mathbf{e}}(0) \leq K,\ \Upsilon_\mathbf{e}]\right) \sum_{\mathbf{e}\in\mathcal{O}_*} \mathbb{P}[\Psi_\mathbf{e}],
\end{aligned} \tag{3.7}$$

where $Q_{c-|\mathbf{e}|\cdot\mathbf{r}}^{\infty,A_\mathbf{e}}$ is defined by recursion (2.3) and the initial condition $Q_{c-|\mathbf{e}|\cdot\mathbf{r}}^{\infty,A_\mathbf{e}}(-t_\mathbf{e}) = B$; the last inequality follows from the independence of $A_\mathbf{e}$ and $A - A_\mathbf{e}$. The preceding inequality and $\mathbb{P}[\Upsilon_\mathbf{e}] \to 1$ as $B \to \infty$ lead to

$$\liminf_{B\to\infty} \frac{\mathbb{P}[Q_A^{B,c} \geq B - K]}{\hat{P}(B)} \geq \liminf_{B\to\infty} \bigwedge_{\mathbf{e}\in\mathcal{O}_*} \mathbb{P}[Q_{c-|\mathbf{e}|\cdot\mathbf{r}}^{\infty,A_\mathbf{e}}(0) \leq K] \liminf_{B\to\infty} \frac{\sum_{\mathbf{e}\in\mathcal{O}_*} \mathbb{P}[\Psi_\mathbf{e}]}{\hat{P}(B)}.$$

At this point, by recalling the definition of $\Psi_{\mathbf{e}}$, counting the number of identical elements in the above sum, using the fact that $\tau_i^e \in \mathcal{IR}$, Lemma A.3 and passing $\varepsilon \downarrow 0$ in the preceding inequality, we obtain

$$\liminf_{B\to\infty} \frac{\mathbb{P}[Q_A^{B,c} \geq B - K]}{\hat{P}(B)} \geq \liminf_{\varepsilon\downarrow 0} \liminf_{B\to\infty} \bigwedge_{\mathbf{e}\in\mathcal{O}_*} \mathbb{P}[Q_{c-|\mathbf{e}|\cdot\mathbf{r}}^{\infty,A_{\mathbf{e}}}(0) \leq K]. \tag{3.8}$$

Finally, using the standard queueing reflection mapping argument, quantity $Q_{c-|\mathbf{e}|\cdot\mathbf{r}}^{\infty,A_{\mathbf{e}}}(0)$ can be represented as

$$\begin{aligned} Q_{c-|\mathbf{e}|\cdot\mathbf{r}}^{\infty,A_{\mathbf{e}}}(0) = \sup_{-t_{\mathbf{e}}\leq s\leq 0} &\left\{(c - |\mathbf{e}|\cdot\mathbf{r})|s| - \int_s^0 A_{\mathbf{e}}(u)\,du\right\} \\ &\vee \left(B + (c - |\mathbf{e}|\cdot\mathbf{r})t_{\mathbf{e}} - \int_{-t_{\mathbf{e}}}^0 A_{\mathbf{e}}(u)\,du\right), \end{aligned}$$

where $\vee$ denotes the maximum. Then the stationarity leads to

$$\begin{aligned} &\mathbb{P}[Q_{c-|\mathbf{e}|\cdot\mathbf{r}}^{\infty,A_{\mathbf{e}}}(0) \leq K] \\ &\quad \geq \mathbb{P}\left[\sup_{s\leq 0}\left\{(c - |\mathbf{e}|\cdot\mathbf{r})|s| - \int_s^0 A_{\mathbf{e}}(u)\,du\right\} \leq K\right] \\ &\qquad - \mathbb{P}\left[t_{\mathbf{e}}^{-1}\int_0^{t_{\mathbf{e}}} A_{\mathbf{e}}(u)\,du - \mathbb{E}A_{\mathbf{e}} \leq -\frac{\varepsilon}{1+\varepsilon}(c - |\mathbf{e}|\cdot\mathbf{r} + (\mathbf{n} - |\mathbf{e}|)\cdot\boldsymbol{\rho})\right] \end{aligned} \tag{3.9}$$

and by the facts that the supremum in the first term is equal to the workload in a stable queue and that process $A_{\mathbf{e}}$ is stationary and ergodic one obtains

$$\liminf_{K\to\infty} \liminf_{\varepsilon\downarrow 0} \liminf_{B\to\infty} \mathbb{P}[Q_{c-|\mathbf{e}|\cdot\mathbf{r}}^{\infty,A_{\mathbf{e}}}(0) \leq K] = 1. \tag{3.10}$$

The preceding limit and (3.8) yield the lower bound for the first statement.

Lower bound for the second statement. Note that in this case $\Psi_{\mathbf{e}} \subseteq \{Q_{A(0)}^{B,c} = B\}$ and, therefore, (3.7) simplifies to

$$\mathbb{P}[Q_A^{B,c}(0) = B] \geq \sum_{\mathbf{e}\in\mathcal{O}_*} \mathbb{P}[\Psi_{\mathbf{e}}].$$

By dividing the preceding inequality with $\hat{P}(B)$, taking $\liminf$ as $B \to \infty$, recalling (3.6), using $\tau_i^e \in \mathcal{IR}$ and letting $\varepsilon \downarrow 0$ we obtain the lower bound for the second statement and conclude the proof of the theorem. □

Our second primary result characterizes the asymptotic behavior of the average loss rate.

THEOREM 3.2. *Let* $r_{\mathbf{m}} = \mathbf{m}\cdot\mathbf{r} + (\mathbf{n}-\mathbf{m})\cdot\boldsymbol{\rho}$. *If* $\tau_i^e \in \mathcal{IR}$ *for* $1 \le i \le M$, *then under Assumption* 3.1, *as* $B \to \infty$,

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

REMARK 3.4. (i) Recall that the loss probability is computable from $P^B = \Lambda^B/\mathbb{E}A$.

(ii) A related result for a discrete time finite buffer queue with a Pareto-like $M/G/\infty$ arrival process can be found in [24]. In their proofs the authors exploit the Poisson decomposition property of the arrival processes, which does not hold for the multiplexed On–Off processes. In addition, in [24] it is assumed that the buffer overflows in a unique way.

PROOF OF THEOREM 3.2. Since the proof is very similar to the proof of Theorem 3.1, we omit some details.

Upper bound. Let $\delta > 0$ be sufficiently small, such that the queues $Q^{B,c-\mathbb{E}A_{\mathbf{e}}-\delta}_{A-A_{\mathbf{e}}}$, $\mathbf{e} \in \mathcal{O}_* \cup \mathcal{U}_*$, $Q^{B,\rho_i+\delta/N}_{A_{ij}}$ have their service rates greater than the mean arrival rates, and $Q^{B,c-\mathbb{E}A_{\mathbf{e}}-\delta}_{A-A_{\mathbf{e}}} \equiv 0$ for all $\mathbf{e} \in \mathcal{U}_*$. Then, recalling the definition of $c_{\mathbf{m}} = c - \mathbf{m}\cdot\mathbf{r} - (\mathbf{n}-\mathbf{m})\cdot\boldsymbol{\rho}$, (3.4) yields, for $0 < \varepsilon < 1$,

$$\begin{aligned}
\Lambda^B &= \mathbb{E}[(A-c)\mathbf{1}\{Q^{B,c}_A = B\}] \\
&\le \sum_{\mathbf{e}\in\mathcal{O}_*} \mathbb{E}[(A-c)\mathbf{1}\{Q^{B,c-\mathbb{E}A_{\mathbf{e}}-\delta}_{A-A_{\mathbf{e}}} \ge \varepsilon B\}] \\
&\quad + (\mathbf{n}\cdot\mathbf{r} - c)\mathbb{P}\left[\bigwedge_{\mathbf{e}\in\mathcal{O}_*\cup\mathcal{U}_*} Q^{B,\mathbb{E}A_{\mathbf{e}}+\delta}_{A_{\mathbf{e}}} \ge (1-\varepsilon)B\right] \\
&\le (1+o(1)) \sum_{\mathbf{m}\in\mathcal{O}} (\mathbf{m}\cdot\mathbf{r} + (\mathbf{n}-\mathbf{m})\cdot\boldsymbol{\rho} - c) \prod_{i=1}^{M} \binom{n_i}{m_i} \mathbb{P}^{m_i}[Q^{B,c_{\mathbf{m}}+r_i-\delta}_{A_{i1}} \ge \varepsilon B],
\end{aligned}$$

where the last inequality follows from the independence of arrival processes, Lemmas 3.1, 3.2 and Proposition 2.3. By dividing both sides of the preceding expression with $\hat{\Lambda}(B)$, taking lim sup as $B \to \infty$, and then passing $\varepsilon \uparrow 1$ and $\delta \downarrow 0$, we obtain the upper bound.

Lower bound. Assume that all processes are in their stationary regimes. Note that, for all $T > 0$,

$$\Lambda^B = \mathbb{E}\lambda^B(0) = T^{-1}\mathbb{E}\int_0^T \lambda^B(u)\,du$$

and recall the definitions of events $\Psi_{\mathbf{e}}$, $\Upsilon_{\mathbf{e}}$ and initial condition for $Q^{\infty,A_{\mathbf{e}}}_{c-|\mathbf{e}|\cdot\mathbf{r}}$ from the proof of the lower bound in Theorem 3.1. Then, by using $B - Q_A^{B,c}(0) \leq Q_{c-|\mathbf{e}|\cdot\mathbf{r}}^{\infty,A_{\mathbf{e}}}(0)$ on event $\Psi_{\mathbf{e}}$, we derive

$$\begin{aligned}\Lambda^B &\geq T^{-1}\sum_{\mathbf{e}\in\mathcal{O}_*}\mathbb{E}\biggl[\int_0^T \lambda^B(u)\,du\,\mathbf{1}\{Q_A^{B,c}(0)\geq B-K,\ \Upsilon_{\mathbf{e}},\ \Psi_{\mathbf{e}}\}\biggr]\\ &\geq T^{-1}\sum_{\mathbf{e}\in\mathcal{O}_*}\mathbb{E}\biggl[\biggl(-K+\int_0^T A(u)\,du - cT\biggr)\\ &\qquad\qquad\times\mathbf{1}\{Q_{c-|\mathbf{e}|\cdot\mathbf{r}}^{\infty,A_{\mathbf{e}}}(0)\leq K,\ \Upsilon_{\mathbf{e}},\ \Psi_{\mathbf{e}}\}\biggr]\\ &\triangleq \sum_{\mathbf{e}\in\mathcal{O}_*} L_{\mathbf{e}}.\end{aligned}\tag{3.11}$$

Next, let $\tau_{\mathbf{e}} \triangleq \wedge_{e_{ij}=1}\inf\{t>0 : A_{ij}(t)=0\}$, that is, $\tau_{\mathbf{e}}$ is the first time after $t=0$ that one of the processes with a large On period is equal to zero. By the independence of $A_{\mathbf{e}}$ and $A - A_{\mathbf{e}}$ we lower bound $L_{\mathbf{e}}$ for every $\mathbf{e}\in\mathcal{O}_*$ as follows:

$$\begin{aligned}L_{\mathbf{e}} &\geq -(KT^{-1}+c)\mathbb{P}[Q_{c-|\mathbf{e}|\cdot\mathbf{r}}^{\infty,A_{\mathbf{e}}}(0)\leq K,\ \Upsilon_{\mathbf{e}}]\mathbb{P}[\Psi_{\mathbf{e}}]\\ &\quad + T^{-1}\mathbb{E}\biggl[\int_0^T A_{\mathbf{e}}(u)\,du\,\mathbf{1}\{Q_{c-|\mathbf{e}|\cdot\mathbf{r}}^{\infty,A_{\mathbf{e}}}(0)\leq K,\ \Upsilon_{\mathbf{e}}\}\biggr]\mathbb{P}[\Psi_{\mathbf{e}}]\\ &\quad + |\mathbf{e}|\cdot\mathbf{r}\,T^{-1}\mathbb{E}[(T\wedge\tau_{\mathbf{e}})\mathbf{1}\{\Psi_{\mathbf{e}}\}]\,\mathbb{P}[Q_{c-|\mathbf{e}|\cdot\mathbf{r}}^{\infty,A_{\mathbf{e}}}(0)\leq K,\ \Upsilon_{\mathbf{e}}],\end{aligned}\tag{3.12}$$

where in the last term we used $\int_0^T (A(u)-A_{\mathbf{e}}(u))\,du \geq (T\wedge\tau_{\mathbf{e}})|\mathbf{e}|\cdot\mathbf{r}$. Now, for all finite T, due to $\tau_i^e \in \mathcal{IR}\subset\mathcal{L}$ for all i and the independence of arrival processes, it follows that

$$\liminf_{B\to\infty}\frac{\mathbb{E}[(T\wedge\tau_{\mathbf{e}})\mathbf{1}\{\Psi_{\mathbf{e}}\}]}{\mathbb{P}[\Psi_{\mathbf{e}}]}\geq T\liminf_{B\to\infty}\mathbb{P}[T<\tau_{\mathbf{e}}|\Psi_{\mathbf{e}}]=T.$$

Inequality (3.12), together with the preceding limit and $\mathbb{P}[\Upsilon_{\mathbf{e}}]\to 1$ as $B\to\infty$, implies

$$\begin{aligned}\liminf_{B\to\infty}\frac{L_{\mathbf{e}}}{\mathbb{P}[\Psi_{\mathbf{e}}]} &\geq \bigl(-KT^{-1}-c+|\mathbf{e}|\cdot\mathbf{r}+(\mathbf{n}-|\mathbf{e}|)\cdot\boldsymbol{\rho}\bigr)\liminf_{B\to\infty}\mathbb{P}[Q_{c-|\mathbf{e}|\cdot\mathbf{r}}^{\infty,A_{\mathbf{e}}}(0)\leq K]\\ &\qquad - \mathbf{n}\cdot\mathbf{r}\limsup_{B\to\infty}\mathbb{P}[Q_{c-|\mathbf{e}|\cdot\mathbf{r}}^{\infty,A_{\mathbf{e}}}(0)>K].\end{aligned}$$

Hence, by setting $T=K^2$, in view of (3.9), for any $\delta>0$ there exist large enough K and B_δ such that, uniformly for all $B\geq B_\delta$ and $\mathbf{e}\in\mathcal{O}_*$,

$$L_{\mathbf{e}}\geq(1-\delta)\mathbb{P}[Q_{c-|\mathbf{e}|\cdot\mathbf{r}}^{\infty,A_{\mathbf{e}}}(0)\leq K](r_{|\mathbf{e}|}-c)\mathbb{P}[\Psi_{\mathbf{e}}],$$

which, when replaced in (3.11), yields

$$\Lambda^B\geq(1-\delta)\biggl(\bigwedge_{\mathbf{e}\in\mathcal{O}_*}\mathbb{P}[Q_{c-|\mathbf{e}|\cdot\mathbf{r}}^{\infty,A_{\mathbf{e}}}(0)\leq K]\biggr)\sum_{\mathbf{e}\in\mathcal{O}_*}(r_{|\mathbf{e}|}-c)\mathbb{P}[\Psi_{\mathbf{e}}].$$

By dividing the preceding equation with $\hat{\Lambda}(B)$ and taking lim inf as $B \to \infty$ we derive

$$\liminf_{B\to\infty} \frac{\Lambda^B}{\hat{\Lambda}(B)} \geq (1-\delta) \liminf_{B\to\infty} \biggl(\bigwedge_{\mathbf{e}\in\mathcal{O}_*} \mathbb{P}\bigl[Q^{\infty, A_{\mathbf{e}}}_{c-|\mathbf{e}|\cdot\mathbf{r}}(0) \leq K\bigr] \biggr)$$

$$\times \liminf_{B\to\infty} \frac{\sum_{\mathbf{e}\in\mathcal{O}_*} (r_{|\mathbf{e}|} - c)\mathbb{P}[\Psi_{\mathbf{e}}]}{\hat{\Lambda}(B)},$$

which after passing $\varepsilon \downarrow 0$ and using Lemma A.3 results in

$$\liminf_{B\to\infty} \frac{\Lambda^B}{\hat{\Lambda}(B)} \geq (1-\delta) \liminf_{\varepsilon\downarrow 0} \liminf_{B\to\infty} \biggl(\bigwedge_{\mathbf{e}\in\mathcal{O}_*} \mathbb{P}\bigl[Q^{\infty, A_{\mathbf{e}}}_{c-|\mathbf{e}|\cdot\mathbf{r}}(0) \leq K\bigr] \biggr).$$

Finally, by setting first $\delta \downarrow 0$, recalling (3.10) and then setting $K \to \infty$ the lower bound follows. This concludes the proof of the theorem. □

For the case of homogeneous arrival processes ($M = 1$), the expressions for the loss rate and overflow probability admit the following simple forms.

COROLLARY 3.1. *Homogeneous sources* ($M = 1$). *Let*

$$\hat{P}(B) \triangleq \binom{N}{m} \biggl(p\,\mathbb{P}\biggl[\tau^e > \frac{B}{mr + (N-m)\rho - c}\biggr]\biggr)^m.$$

If $\rho N < c < rN$, $\tau^e \in \mathcal{IR}$ *and there is an integer* $m \geq 1$ *such that* $0 < mr + (N-m)\rho - c < r - \rho$, *then*

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

and, as $B \to \infty$,

$$\Lambda^B \sim \bigl(mr + (N-m)\rho - c\bigr)\hat{P}(B).$$

If in addition $mr > c$, *then* $\mathbb{P}[Q_A^{B,c} \geq B - K] \sim \mathbb{P}[Q_A^{B,c} = B] \sim \hat{P}(B)$ *as* $B \to \infty$ *for all* $K \geq 0$.

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 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

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

This is satisfied for a general class of exponentially bounded arrival processes (see [12, 34]) as well as for some heavy-tailed processes, for example, On–Off processes with Weibull On periods, $\mathbb{P}[\tau > x] = e^{-x^b}$, $0 < b < 1$, $x \geq 0$ (see Theorem 2.1). 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 $Q^{\infty,c_1+c_2}_{A_1+A_2}$ is stochastically dominated by $Q^{\infty,c_1}_{A_1} + Q^{\infty,c_2}_{A_2}$, $c_i > \mathbb{E}A_i$ (an infinite buffer equivalent of Proposition 2.1). Thus, we will use $A_{\mathcal{SP}}$ to denote the aggregate process of all arriving subpolynomial processes. The following corollary yields the reduce load equivalence results for multiplexing subpolynomial and intermediately regularly varying processes.

COROLLARY 3.2. *Suppose that $A_{\mathcal{SP}} \in \mathcal{SP}$ and Assumption* 3.1 *is satisfied with* $(c - \mathbb{E}A_{\mathcal{SP}})$ *in place of* c. *If* $\tau_i^e \in \mathcal{IR}$ *for* $1 \leq i \leq M$, *then*

$$\lim_{K\to\infty} \liminf_{B\to\infty} \frac{\mathbb{P}[Q^{B,c}_{A+A_{\mathcal{SP}}} \geq B - K]}{\mathbb{P}[Q^{B,c-\mathbb{E}A_{\mathcal{SP}}}_{A} \geq B - K]}$$

$$= \lim_{K\to\infty} \limsup_{B\to\infty} \frac{\mathbb{P}[Q^{B,c}_{A+A_{\mathcal{SP}}} \geq B - K]}{\mathbb{P}[Q^{B,c-\mathbb{E}A_{\mathcal{SP}}}_{A} \geq B - K]} = 1,$$

and if, for some $\delta > 0$, $\mathbb{E}A^{1+\delta}_{\mathcal{SP}} < \infty$, *then as* $B \to \infty$,

$$\Lambda^{B,c}_{A+A_{\mathcal{SP}}} \sim \Lambda^{B,c-\mathbb{E}A_{\mathcal{SP}}}_{A}.$$

PROOF. First, by stochastic dominance, for any $c > \mathbb{E}A_{\mathcal{SP}}$, $\mathbb{P}[Q^{\infty,c}_{A_{\mathcal{SP}}} \geq B - K] \geq \mathbb{P}[Q^{B,c}_{A_{\mathcal{SP}}} \geq B - K]$, and therefore, for any $\beta > 0$,

$$\lim_{B\to\infty} B^{\beta}\mathbb{P}[Q^{B,c}_{A_{\mathcal{SP}}} \geq B - K] = 0. \tag{3.13}$$

Then, by Proposition 2.1, for any $0 < \delta < c - \mathbb{E}A_{\mathcal{SP}}$ and $0 < \varepsilon < 1$,

$$\begin{aligned} \mathbb{P}[Q^{B,c}_{A+A_{\mathcal{SP}}} \geq B - K] &\leq \mathbb{P}[Q^{B,c-\mathbb{E}A_{\mathcal{SP}}-\delta}_{A} \geq \varepsilon(B - K)] \\ &\quad + \mathbb{P}[Q^{B,\mathbb{E}A_{\mathcal{SP}}+\delta}_{A_{\mathcal{SP}}} \geq (1-\varepsilon)(B - K)]. \end{aligned} \tag{3.14}$$

Next, recall the definition of $\hat{P}(B) \equiv \hat{P}(B, c')$, $c' = c - \mathbb{E}A_{\mathcal{SP}}$ from Theorem 3.1. Clearly, $\hat{P}(B)$ belongs to $\mathcal{IR}$, and thus, there exists a finite α such that, for all sufficiently large B,

$$\hat{P}(B) \geq \frac{1}{B^{\alpha}}; \tag{3.15}$$

see equation (1.6) of [29]. Now, by dividing (3.14) with $\hat{P}(B)$, taking lim sup as $B \to \infty$, using Theorem 3.1, (3.13) and (3.15), and then passing $\varepsilon \uparrow 1$, $\delta \downarrow 0$ we complete the proof of the upper bound.

The upper bound for the loss rate is obtained by using the same approach and, instead of (3.14),

$$\begin{aligned}
\Lambda^{B,c}_{A+A_{\mathcal{SP}}} &= \mathbb{E}\big[(A + A_{\mathcal{SP}} - c)\mathbf{1}\big\{Q^{B,c}_{A+A_{\mathcal{SP}}} = B\big\}\big] \\
&\le \mathbb{E}\big[(A + A_{\mathcal{SP}} - c)\mathbf{1}\big\{Q_A^{B,c-\mathbb{E}A_{\mathcal{SP}}-\delta} \ge \varepsilon B\big\}\big] \\
&\quad + \mathbb{E}\big[(A + A_{\mathcal{SP}} - c)\mathbf{1}\big\{Q_{A_{\mathcal{SP}}}^{B,\mathbb{E}A_{\mathcal{SP}}+\delta} \ge (1-\varepsilon)B\big\}\big] \\
&\le \mathbb{E}\big[(A + \mathbb{E}A_{\mathcal{SP}} - c)\mathbf{1}\big\{Q_A^{B,c-\mathbb{E}A_{\mathcal{SP}}-\delta} \ge \varepsilon B\big\}\big] \\
&\quad + (\mathbb{E}A - c)\,\mathbb{P}\big[Q_{A_{\mathcal{SP}}}^{B,\mathbb{E}A_{\mathcal{SP}}+\delta} \ge (1-\varepsilon)B\big] \\
&\quad + \big(\mathbb{E}A_{\mathcal{SP}}^{1+\delta}\big)^{1/(1+\delta)}\big(\mathbb{P}\big[Q_{A_{\mathcal{SP}}}^{B,\mathbb{E}A_{\mathcal{SP}}+\delta} \ge (1-\varepsilon)B\big]\big)^{\delta/(1+\delta)},
\end{aligned}$$

where the last inequality follows from the independence of A and $A_{\mathcal{SP}}$, and Hölder's inequality.

The proofs of the lower bounds can be done in the same spirit as in Theorems 3.1 and 3.2. These proofs only require that $A_{\mathcal{SP}}$ satisfies the strong law of large numbers, which follows from the stationarity and ergodicity. To avoid repetition we omit the details. □

4. Numerical experiments.

In this section we illustrate through simulation experiments the accuracy of Theorems 3.1 and 3.2, or more precisely Corollary 3.1, in approximating the overflow probabilities and loss rates for finite buffer sizes. Since the asymptotic results are insensitive to the distribution of Off periods we choose their distribution to be exponential $\mathbb{P}[\nu > x] = e^{-\lambda x}$, $x \ge 0$; On periods are selected from Pareto family $\mathbb{P}[\tau > x] = x^{-\alpha}$, $x \ge 1$, $\alpha > 1$. We select α in the range of recently measured file sizes ($\alpha = 1.44$, see Figure 1). The asymptotic approximation from Corollary 3.1 computes explicitly to

$$\hat{P}(B) = \binom{N}{m}\left[\frac{p}{\alpha B^{\alpha-1}}\big(mr + (N-m)\rho - c\big)^{\alpha-1}\right]^m, \tag{4.1}$$

where $p = \lambda\alpha/(\lambda\alpha + \alpha - 1)$ and $\rho = rp$. To ensure the accuracy of our simulation experiments we select the length of the simulated sample path to be $t = 10^{10}$.

EXAMPLE 1. Here, we select $N = 10$ i.i.d. On–Off processes with parameters $\lambda = 0.012$, $\alpha = 1.3$ and $r = 2$, which yield $p = 0.05$ and $\rho = 0.1$. For the choice of capacity $c = 5$ we simulated the overflow probabilities for buffer sizes $B = 100i$, $i = 1, \dots, 10$. The results of the simulation are presented in Figure 3 with "+" symbols. The selected parameters render $m = 3$ for the asymptotic approximation $\hat{P}(B)$, as defined in (4.1). Note that $mr > c$ and, therefore, we can use the last statement of Corollary 3.1 to approximate $\mathbb{P}[Q^B = B]$. The accuracy of the approximation, plotted on the same figure with dashed lines, is apparent.

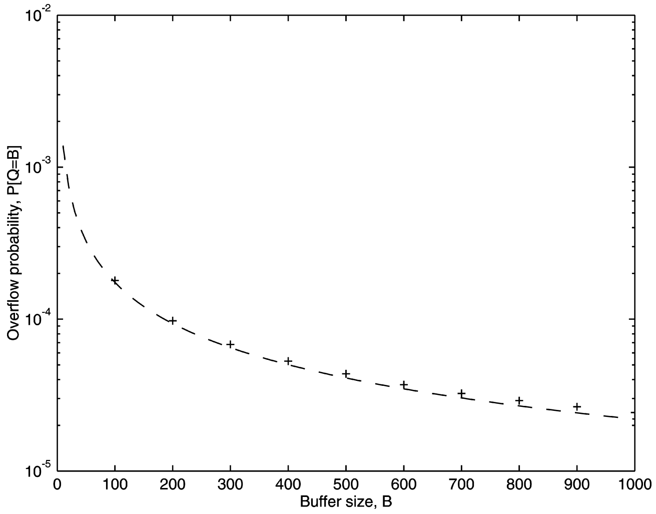

FIG. 3. *Illustration for Example* 1.

EXAMPLE 2. In this example we choose $N = 50$, $\lambda = 3.37 \cdot 10^{-3}$, $\alpha = 1.5$, $r = 3$, which imply $p = 0.01$ and $\rho = 0.03$. Now, for the same capacity $c = 5$, we readily compute $m = 2$, the asymptotic formula $\hat{\Lambda}(B) = (mr + (N - m)\rho - c) \times \hat{P}(B)$ and repeat the same simulation procedure as in Example 1. The results of the simulation and approximation are plotted with "+" symbols and dashed lines, respectively. Again, the approximation matches well the simulated estimates even for relatively high probabilities.

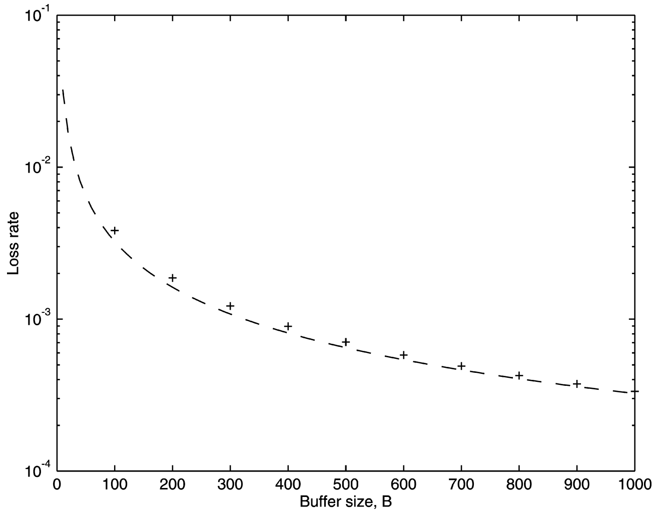

FIG. 4. *Illustration for Example* 2.

Additional experimental validation of our asymptotic results can be found in [21].

5. Concluding remarks. We considered a finite buffer fluid queue fed by a superposition of heterogeneous heavy-tailed On–Off processes. Explicit and asymptotically exact results were derived for approximating the overflow probability and loss rate in the case when excess On periods are intermediately regularly varying. The potential of using the obtained asymptotic formulas in the nonlimiting regime was demonstrated with simulation experiments. The results provide important insight into qualitative tradeoffs between the performance measures and system parameters.

6. Proofs. This section contains the proof of Proposition 2.4. The proof is based on the subsequent three lemmas which derive preliminary results on a discrete time finite buffer queue, defined as follows. Consider two i.i.d. sequences of positive random variables $\{X, X_n\}$ and $\{Y, Y_n\}$, $n \in \mathbb{N}$. Let $W_0^B = 0$ and

$$W_{n+1}^B = \big((W_n^B + X_{n+1}) \wedge B - Y_{n+1}\big)^+. \tag{6.1}$$

Assuming that $\mathbb{P}[X_n = Y_n] < 1$, in Chapter III.5 of [8] it was shown that W_n^B has a unique stationary distribution and that, for any initial condition W_0^B, W_n^B converges to that stationary distribution. Let W^B be a random variable that is equal in distribution to W_n^B in stationarity.

LEMMA 6.1. *If $X^e \in \mathcal{IR}$ and $\mathbb{E}X < \mathbb{E}Y$, then*

$$\lim_{\delta \uparrow 1} \limsup_{B \to \infty} \frac{\mathbb{P}[W^B \geq \delta B]}{\mathbb{P}[X^e \geq B]} = 0.$$

PROOF. Instead of proving the statement for W^B directly, we will consider an easier-to-analyze queueing process V^B that stochastically upper bounds W^B. Let $\{V_n^B\}_{n=0}^{\infty}$ be defined by $V_0^B = 0$ and

$$V_{n+1}^B = \big(V_n^B + X_{n+1} - Y_{n+1}\big)^+ \wedge B.$$

The above recursion is similar to (6.1) and under the nontriviality condition $\mathbb{P}[X_n = Y_n] < 1$, V_n^B converges to a stationary distribution (see [8], Chapter III.4); again, let V^B be a random variable that is equal in distribution to V_n^B in stationarity. Now, we show that $W_n^B \leq V_n^B$ for all $n \geq 0$. Clearly, this is implied by $W_0^B = V_0^B = 0$ and the next inductive step

$$\begin{aligned} W_{n+1}^B &= \big((W_n^B + X_{n+1}) \wedge B - Y_{n+1}\big)^+ \\ &\leq \big(V_n^B + X_{n+1} - Y_{n+1}\big)^+ \wedge B = V_{n+1}^B, \end{aligned}$$

where the inequality follows from

$$(x \wedge B - y)^+ \le (x \wedge B - y \wedge B)^+ \le (x - y)^+ \wedge B,$$

for $x, y \ge 0$. Therefore, it will be enough to prove the statement of the lemma with W^B being replaced by V^B.

First, we restrict our attention to the case of $\{X_n\}_{n=1}^\infty$, $\{Y_n\}_{n=1}^\infty$ being lattice valued and Y bounded. Without loss of generality we assume that X_n and Y_n are integer valued. Let $q_i^B \triangleq \mathbb{P}[V^B = i]$ and $\eta_i^B \triangleq q_i^B / q_0^B$ for $0 \le i \le B \le \infty$, $B \in \mathbb{N}$ ($B = \infty$ represents the infinite buffer case). Lemma 1 from [17] yields, for all $B \ge 0$,

$$\begin{aligned}\mathbb{P}[V^B \ge \delta B] &\le \sum_{i=\lfloor \delta B \rfloor}^{B} q_0^B \eta_i^B \le \sum_{i=\lfloor \delta B \rfloor}^{B} \eta_i^B \\ &\le \sum_{i=\lfloor \delta B \rfloor}^{B} \left(\eta_i^\infty + K_2 \mathbb{P}[V^B + X - Y > B]\mu^{B-i}\right),\end{aligned}$$

where K_2 is a positive constant, $0 < \mu < 1$ if $\mathbb{P}[Y_n = 0] + \mathbb{P}[Y_n = 1] < 1$, and $\mu = 0$, otherwise. Next, the preceding inequality and Lemma 2 from [17] imply

$$\mathbb{P}[V^B \ge \delta B] \le \frac{1}{q_0^\infty} \mathbb{P}[\delta B \le V^\infty \le B] + \mathbf{1}\{\mu \ne 0\} \frac{K_2}{1 - \mu} o(\mathbb{P}[X^e > B]).$$

Thus,

$$\limsup_{B \to \infty} \frac{\mathbb{P}[V^B \ge \delta B]}{\mathbb{P}[X^e \ge B]} \le \frac{1}{q_0^\infty}\left(\limsup_{B \to \infty} \frac{\mathbb{P}[V^\infty \ge \delta B]}{\mathbb{P}[X^e \ge B]} - \liminf_{B \to \infty} \frac{\mathbb{P}[V^\infty \ge B]}{\mathbb{P}[X^e \ge B]}\right),$$

which, by Pakes's theorem [27] and $X^e \in \mathcal{IR}$ results in

$$\lim_{\delta \uparrow 1} \limsup_{B \to \infty} \frac{\mathbb{P}[V^B \ge \delta B]}{\mathbb{P}[X^e \ge B]} = 0.$$

This proves the result for lattice-valued X and Y, with Y being bounded. Next, we use the technique from [17], pages 98–99, to extend the result to the general case of nonlattice-valued X, Y with Y unbounded.

If Y is unbounded we can always choose a truncated service variable $Y_n^d = Y_n \wedge d$, with d being sufficiently large to satisfy the stability condition $\mathbb{E}X_n < \mathbb{E}Y_n^d$. Let $W_n^{B,d}$ be a process satisfying recursion (6.1) with the process Y_n^d in place of Y_n. It is clear that the stationary workload for this queue $W^{B,d}$ is stochastically larger than the original workload W^B, that is, $\mathbb{P}[W^B \ge \delta B] \le \mathbb{P}[W^{B,d} \ge \delta B]$. Therefore, since the lemma holds for process $W^{B,d}$, it is true for W^B as well. When X and Y are nonlattice, we can approximate them with lattice-valued random variables X' and Y'. For any $\Delta > 0$ such that $\mathbb{E}Y - \mathbb{E}X + 2\Delta < 0$, define distribution functions for Y' and X' as

$$\begin{aligned} \mathbb{P}[Y' = i\Delta] &= \mathbb{P}[i\Delta \le Y < (i+1)\Delta], \qquad i \ge 0, \\ \mathbb{P}[X' = i\Delta] &= \mathbb{P}[(i-1)\Delta \le X < i\Delta], \qquad i \ge 1. \end{aligned}$$

From these definitions it follows that, for all $x \ge 0$,

$$\begin{aligned} \mathbb{P}[Y > x + \Delta] \le \mathbb{P}[Y' > x] &\le \mathbb{P}[Y > x], \\ \mathbb{P}[X > x] \le \mathbb{P}[X' > x] &\le \mathbb{P}[X > x - \Delta], \end{aligned}$$

which implies that $X' - Y'$ is stochastically larger that $X - Y$, $\mathbb{E}X' \le \mathbb{E}X + \Delta < \mathbb{E}Y - \Delta \le \mathbb{E}Y'$, and

$$\int_B^\infty \mathbb{P}[X' > u]\, du \sim \int_B^\infty \mathbb{P}[X > u]\, du \qquad \text{as } B \to \infty. \tag{6.2}$$

Next, let $\{X'_n\}_{n=1}^\infty$ and $\{Y'_n\}_{n=1}^\infty$ be two independent i.i.d. sequences with probability distributions equal to distributions of X' and Y', respectively. Consider a queue W'^B with buffer B that corresponds to sequences $\{X'{}_n\}_{n=1}^\infty$ and $\{Y'{}_n\}_{n=1}^\infty$. Then the newly constructed queueing process, W'^B, dominates the original queueing process in distribution

$$\mathbb{P}[W^B \ge \delta B] \le \mathbb{P}\big[W'^B \ge \delta B\big].$$

Finally, the preceding inequality and (6.2) imply

$$\limsup_{B\to\infty} \frac{\mathbb{P}[W^B \ge \delta B]}{\mathbb{P}[X^e \ge B]} \le \limsup_{B\to\infty} \frac{\mathbb{P}[W'^B \ge \delta B]}{\mathbb{P}[X'^e \ge B]},$$

which, by using the already proved lattice case and letting $\delta \uparrow 1$ yields the desired result. This concludes the proof. □

LEMMA 6.2. *If X^e is independent of W^B, $X^e \in \mathcal{IR}$ and $\mathbb{E}X < \mathbb{E}Y$, then*

$$\lim_{\varepsilon\uparrow 1} \limsup_{B\to\infty} \frac{\mathbb{P}[X^e + W^B \ge \varepsilon B]}{\mathbb{P}[X^e \ge B]} = 1.$$

PROOF. Let $\varepsilon \in (0,1)$. For all $\delta \in (0, \varepsilon/2)$ a simple argument leads to

$$\begin{aligned} \mathbb{P}[X^e + W^B \ge \varepsilon B] &\le \mathbb{P}[X^e \ge (\varepsilon - \delta)B] \\ &\qquad + \mathbb{P}[X^e + W^B \ge \varepsilon B, X^e < (\varepsilon - \delta)B] \\ &\le \mathbb{P}[X^e \ge (\varepsilon - \delta)B] \\ &\qquad + \frac{1}{\mathbb{E}X} \int_0^{(\varepsilon-\delta)B} \mathbb{P}[W^B \ge \varepsilon B - x]\mathbb{P}[X \ge x]\, dx. \end{aligned} \tag{6.3}$$

Next, we bound the second term in (6.3) as follows:

$$\int_0^{\delta B} \mathbb{P}[W^B \geq \varepsilon B - x]\mathbb{P}[X \geq x]\,dx \leq \mathbb{E}X\mathbb{P}[W^B \geq (\varepsilon - \delta)B]$$

and

$$\int_{\delta B}^{(\varepsilon-\delta)B} \mathbb{P}[W^B \geq \varepsilon B - x]\mathbb{P}[X \geq x]\,dx \leq \mathbb{P}[W^B \geq \delta B](\varepsilon - 2\delta)B\mathbb{P}[X \geq \delta B],$$

which together with Lemma A.2 results in

$$\begin{aligned} &\limsup_{B\to\infty} \frac{\int_0^{(\varepsilon-\delta)B} \mathbb{P}[W^B \geq \varepsilon B - x]\mathbb{P}[X \geq x]\,dx}{\mathbb{P}[X^e \geq B]} \\ &\qquad \leq \mathbb{E}X \limsup_{B\to\infty} \frac{\mathbb{P}[W^B \geq (\varepsilon-\delta)B]}{\mathbb{P}[X^e \geq B]}. \end{aligned} \tag{6.4}$$

By dividing (6.3) with $\mathbb{P}[X^e \geq B]$, taking lim sup with respect to B and using (6.4), we obtain

$$\limsup_{B\to\infty} \frac{\mathbb{P}[X^e + W^B \geq \varepsilon B]}{\mathbb{P}[X^e \geq B]} \leq \limsup_{B\to\infty} \frac{\mathbb{P}[X^e \geq (\varepsilon-\delta)B] + \mathbb{P}[W^B \geq (\varepsilon-\delta)B]}{\mathbb{P}[X^e \geq B]}.$$

Hence, by letting $\delta \downarrow 0$, $\varepsilon \uparrow 1$ and invoking Lemma 6.1 in the preceding inequality the desired statement follows. □

LEMMA 6.3. *Let X, Y^e and W^B be mutually independent. If $X^e \in \mathcal{IR}$ and $\mathbb{E}X < \mathbb{E}Y$, then*

$$\lim_{\varepsilon\uparrow 1} \limsup_{B\to\infty} \frac{\mathbb{P}[(W^B + X) \wedge B - Y^e \geq \varepsilon B]}{\mathbb{P}[X^e \geq B]} = 0.$$

PROOF. Let $\varepsilon \in (0, 1)$. For all $\delta \in (0, \varepsilon)$ we write

$$\begin{aligned} \mathbb{P}[(W^B + X) \wedge B - Y^e \geq \varepsilon B] &\leq \mathbb{P}[W^B + X \geq \varepsilon B] \\ &\leq \mathbb{P}[W^B \geq (\varepsilon-\delta)B] + \mathbb{P}[X \geq \delta B], \end{aligned}$$

which, jointly with Lemmas A.2 and 6.1, leads to

$$\begin{aligned} &\lim_{\varepsilon\uparrow 1} \limsup_{B\to\infty} \frac{\mathbb{P}[X^e + W^B \geq \varepsilon B]}{\mathbb{P}[X^e \geq B]} \\ &\qquad \leq \lim_{\varepsilon\uparrow 1} \lim_{\delta\downarrow 0} \limsup_{B\to\infty} \frac{\mathbb{P}[X \geq \delta B] + \mathbb{P}[W^B \geq (\varepsilon-\delta)B]}{\mathbb{P}[X^e \geq B]} = 0. \end{aligned}$$ □

At this point we are able to provide a proof of Proposition 2.4.

PROOF OF PROPOSITION 2.4. Let $X_{n+1} = (r-c)\tau_{n+1}$, $Y_{n+1} = c\nu_{n+1}$ in (6.1) and assume that W_n^B is in its stationary regime. Then W_n^B represents the amount of fluid in a queue with a single On–Off arrival process observed at the beginnings of On periods. Thus, by evaluating $Q^B(t)$ in stationarity at time (say) $t = 0$ we derive (for simplicity of notation we leave out the time index)

$$\begin{aligned}\mathbb{P}[Q^B \geq \varepsilon B] &= \mathbb{P}\big[A = 0,\ \big(W^B + (r-c)\tau\big) \wedge B - c\nu^e \geq \varepsilon B\big]\\ &\quad + \mathbb{P}[A = r,\ W^B + (r-c)\tau^e \geq \varepsilon B]\\ &= (1-p)\mathbb{P}\big[\big(W^B + (r-c)\tau\big) \wedge B - c\nu^e \geq \varepsilon B\big]\\ &\quad + p\mathbb{P}[W^B + (r-c)\tau^e \geq \varepsilon B].\end{aligned}$$

Then, by dividing the preceding equality with $\mathbb{P}[(r-c)\tau^e \geq B]$, applying the operator $\lim_{\varepsilon\uparrow 1}\limsup_{B\to\infty}$, and using Lemmas 6.2, 6.3 and Proposition 2.2 we complete the proof. □

APPENDIX: HEAVY-TAILED DISTRIBUTIONS

This appendix contains the definitions and some of the basic properties of heavy-tailed 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 F.

DEFINITION A.1. A nonnegative r.v. X is called long-tailed, $X \in \mathcal{L}$, if

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

The following class of heavy-tailed distributions was introduced by Chistyakov [5].

DEFINITION A.2. A nonnegative r.v. X is called subexponential, $X \in \mathcal{S}$, if

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

where F^{2*} denotes the twofold convolution of F with itself, that is, $F^{2*}(x) = \int_0^\infty F(x-y)F(dy)$.

It is well known that $\mathcal{S} \subset \mathcal{L}$ [3]. A survey on subexponential distributions can be found in [13]. The class of intermediately regularly varying distributions $\mathcal{IR}$ is a subclass of $\mathcal{S}$.

DEFINITION A.3. A nonnegative r.v. X is called intermediately regularly varying, $X \in \mathcal{IR}$, 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 A.4. A nonnegative r.v. X is called regularly varying with index α, $X \in \mathcal{R}_\alpha$, if

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

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

We conclude the Appendix with three basic lemmas on $\mathcal{IR}$ distributions.

LEMMA A.1. *Let $X \in \mathcal{IR}$ and $\eta \in (0, 1)$. Then*

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

PROOF. Follows immediately from the definition. □

LEMMA A.2. *If $X^e \in \mathcal{IR}$, then*

$$\limsup_{x \to \infty} \frac{x \mathbb{P}[X \geq x]}{\mathbb{P}[X^e \geq x]} < \infty.$$

PROOF. For any $\delta \in (0, 1)$ by definition of F^e,

$$\begin{aligned} \frac{x \mathbb{P}[X \geq x]}{\mathbb{P}[X^e \geq x]} &\leq \frac{\mathbb{P}[X^e \geq \delta x]}{\mathbb{P}[X^e \geq x]} \frac{x \mathbb{P}[X \geq x] \mathbb{E} X}{\int_{\delta x}^{x} \mathbb{P}[X \geq u]\, du} \\ &\leq \frac{\mathbb{P}[X^e \geq \delta x]}{\mathbb{P}[X^e \geq x]} \frac{\mathbb{E} X}{1 - \delta}. \end{aligned}$$

Hence, the result follows by Lemma A.1,

$$\limsup_{x \to \infty} \frac{x \mathbb{P}[X \geq x]}{\mathbb{P}[X^e \geq x]} \leq \frac{\mathbb{E} X}{1 - \delta} \limsup_{x \to \infty} \frac{\mathbb{P}[X^e \geq \delta x]}{\mathbb{P}[X^e \geq x]} < \infty. \qquad \square$$

For any bounded nondecreasing function F we say that $F \in \mathcal{IR}$ if it satisfies Definition A.3. Then the following lemma follows directly from the definition.

LEMMA A.3. *If $F_1, F_2 \in \mathcal{IR}$, then the following hold*:

(i) $F_1 F_2 \in \mathcal{IR}$;

(ii) $w_1 F_1 + w_2 F_2 \in \mathcal{IR}$, *for* $w_1, w_2 > 0$.

REFERENCES

[1] AGRAWAL, R., MAKOWSKI, A. and NAIN, PH. (1999). On a reduced load equivalence for fluid queues under subexponentiality. *Queueing Systems Theory Appl.* **33** 5–41.

[2] ANICK, D., MITRA, D. and SONDHI, M. (1982). Stochastic theory of a data handling system with multiple sources. *Bell Syst. Tech. J.* **61** 1871–1894.

[3] ATHREYA, K. B. and NEY, P. E. (1972). *Branching Processes*. Springer, Berlin.

[4] BOXMA, O. J. (1996). Fluid queues and regular variation. *Performance Evaluation* **27/28** 699–712.

[5] CHISTYAKOV, V. P. (1964). A theorem on sums of independent positive random variables and its application to branching random processes. *Theory Probab. Appl.* **9** 640–648.

[6] CHOUDHURY, G. L. and WHITT, W. (1997). Long-tail buffer-content distributions in broadband networks. *Performance Evaluation* **30** 177–190.

[7] COHEN, J. W. (1974). Superimposed renewal processes and storage with gradual input. *Stochastic Process. Appl.* **2** 31–58.

[8] COHEN, J. W. (1982). *The Single Server Queue*. North-Holland, Amsterdam.

[9] CROVELLA, M. and BESTAVROS, A. (1997). Self-similarity in World Wide Web traffic: Evidence and possible causes. *IEEE/ACM Trans. Networking* **5** 835–846.

[10] DUFFIELD, N. G. (1998). Queueing at large resources driven by long-tailed M/G/∞-modulated processes. *Queueing Systems Theory Appl.* **28** 245–266.

[11] DUMAS, V. and SIMONIAN, A. (2000). Asymptotic bounds for the fluid queue fed by subexponential on/off sources. *Adv. in Appl. Probab.* **32** 224–255.

[12] GLYNN, P. V. and WHITT, W. (1994). Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. In *Studies in Applied Probability* (J. Galambos and J. Gani, eds.) **31A** 131–156. Applied Probability Trust, Sheffield, UK. (Special issue of *J. Appl. Probab.*)

[13] GOLDIE, C. M. and KLÜPPELBERG, C. (1998). Subexponential distributions. In *A Practical Guide to Heavy Tails*: *Statistical Techniques for Analysing Heavy Tailed Distributions* (R. Adler, R. Feldman and M. S. Taqqu, eds.) 435–459. Birkhäuser, Boston.

[14] HEATH, D., RESNICK, S. and SAMORODNITSKY, G. (1998). Heavy tails and long range dependence in on/off processes and associated fluid models. *Math. Oper. Res.* **23** 145–165.

[15] HEATH, D., RESNICK, S. and SAMORODNITSKY, G. (1999). How system performance is affected by the interplay of averages in a fluid queue with long range dependence induced by heavy tails. *Ann. Appl. Probab.* **9** 352–375.

[16] HEYMAN, D. P. and LAKSHMAN, T. V. (1996). Source models for VBR broadcast-video traffic. *IEEE/ACM Trans. Networking* **4** 40–48.

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

[18] JELENKOVIĆ, P. (2000). On the asymptotic behavior of a fluid queue with a heavy-tailed M/G/∞ arrival process. *Oper. Res. Lett.* To appear.

[19] JELENKOVIĆ, P. and LAZAR, A. (1999). Asymptotic results for multiplexing subexponential on–off processes. *Adv. in Appl. Probab.* **31** 394–421.

[20] JELENKOVIĆ, P., LAZAR, A. and SEMRET, N. (1997). The effect of multiple time scales and subexponentiality of MPEG video streams on queueing behavior. *IEEE J. Select. Areas Comm.* **15** 1052–1071.

[21] JELENKOVIĆ, P. and MOMČILOVIĆ, P. (2001). Capacity regions for network multiplexers with heavy-tailed fluid on–off sources. In *Proc. IEEE Infocom*, *Anchorage*, *Alaska*.

[22] KRISHNAN, K. R. and MEEMPAT, G. (1997). Long-range dependence in VBR video streams and ATM traffic engineering. *Performance Evaluation* **30** 45–56.

[23] LELAND, W. E., TAQQU, M. S., WILLINGER, W. and WILSON, D. V. (1993). On the self-similar nature of Ethernet traffic. In *Proc. ACM SIGCOMM* 183–193. ACM, New York.

[24] LIKHANOV, N. and MAZUMDAR, R. (2000). Cell loss asymptotics in buffers fed by heterogeneous long-tailed sources. In *Proc. IEEE Infocom*, *Tel-Aviv*, *Israel*.

[25] LIU, Z., NAIN, PH., TOWSLEY, D. and ZHANG, Z. L. (1999). Asymptotic behavior of a multiplexer fed by a long-range dependent process. *J. Appl. Probab.* **36** 105–118.

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

[27] PAKES, A. G. (1975). On the tails of waiting-time distribution. *J. Appl. Probab.* **12** 555–564.

[28] PARULEKAR, M. and MAKOWSKI, A. M. (1997). Tail probabilities for M/G/∞ input processes (I): Preliminary asymptotics. *Queueing Systems Theory Appl.* **27** 271–296.

[29] RESNICK, S. and SAMORODNITSKY, G. (1999). Activity periods of an infinite server queue and performance of certain heavy tailed fluid queues. *Queueing Systems Theory Appl.* **33** 43–71.

[30] RESNICK, S. and SAMORODNITSKY, G. (2001). Steady state distribution of the buffer content for M/G/∞ input fluid queues. *Bernoulli* **7** 191–210.

[31] ROLSKI, T., SCHLEGEL, S. and SCHMIDT, V. (1999). Asymptotics of Palm-stationary buffer content distribution in fluid flow queues. *Adv. in Appl. Probab.* **31** 235–253.

[32] RUBINOVITCH, M. (1973). The output of a buffered data communication system. *Stochastic Process. Appl.* **1** 375–380.

[33] VAMVAKOS, S. and ANANTHARAM, V. (1998). On the departure process of a leaky bucket system with long-range dependent input traffic. *Queueing Systems Theory Appl.* **28** 191–214.

[34] WEISS, A. and SHWARTZ, A. (1995). *Large Deviations for Performance Analysis*: *Queues, Communications, and Computing*. Chapman and Hall, London.

[35] ZWART, B., BORST, S. and MANDJES, M. (2000). Exact asymptotics for fluid queues fed by multiple heavy-tailed on–off flows. *Ann. Appl. Probab.* To appear.

DEPARTMENT OF ELECTRICAL ENGINEERING
COLUMBIA UNIVERSITY
NEW YORK, NEW YORK 10027
E-MAIL: {predrag, petar}@ee.columbia.edu