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 independently and identically distributed, we derive explicit and asymptotically exact formulas for approximating the stationary loss 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 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 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 queuing system are the stationary overflow probability and loss rate. The analysis of a related infinite buffer queuing system dates back to [1], [2], [3] (see also [4] for additional references).

Most of the early work on multiplexing focused on On-Off processes with exponentially distributed On and Off periods (e.g., see [3]). However, repeated empirical measurements in modern multimedia 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 an 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 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 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.

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 mul-
Multiplexing a single heavy-tailed process with exponential streams was investigated in [11], [4], [12]. In [4] it was discovered that 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].

However, the understanding of multiplexing a finite number of heavy-tailed On-Off arrival processes is quite limited. General bounds can be found in [20], [21]. In this paper we derive explicit asymptotic 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 [22] (see also [23]). During the process of completing this paper we discovered that the complementary results for the infinite buffer model are derived in [24].

Informally, in the case of multiplexing $N$ homogeneous On-Off sources with peak rate $r$, average rate $\rho$ and probability of being on $p_{on}$ into a queue of capacity $c$ and buffer $B$ our result shows that the queue overflow probability $\mathbb{P}[Q^B = B]$ is asymptotically, as $B \to \infty$, equal to

$$
\left( \frac{N}{k_0} \right)^{k_0 \rho p_{on}} \left[ \tau^* r > \frac{B}{k_0 r + (N - k_0)\rho - c} \right],
$$

(1)

where $\tau^*$ 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 overflow probability decays polynomially in buffer size $B$ and exponentially in capacity $c$. This insight may prove to be important in designing network switching elements.

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 queuing 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 increasing jump times $\{T_0 = 0 < T_1 < T_2 < \cdots \}$. In this case, for any initial value $Q^B(0)$ and $t \in (T_n, T_{n+1}]$, $n \geq 0$, the evolution of $Q^B(t)$ is given by

$$
Q^B(t) = (Q^B(T_n) + (t - T_n)(A(T_n) - c)) + B,
$$

(2)

where $(x)^+ = \max(0, x)$ and $x \wedge y = \min(x, y)$. When necessary, we will use the notation $Q^B_A \equiv Q^B$ 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 [25], 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 computation of the overflow probability $\mathbb{P}[Q^B = B]$ and long time
average loss rate $\Lambda^B$ given by

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

where $L(0, t) = \{\text{amount of fluid lost in } (0, t)\}$. An equivalent representation of $\Lambda^B$, 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 $t$.

Next we prove a useful sample path bound. The bound formalizes an intuitively expected notion that multiplexing reduces the aggregate queuing workload.

**Proposition 1:** If $A(t) = \sum_{n=1}^{N} A_n(t)$ and $c = \sum_{n=1}^{N} c_n$, then

$$Q_A^B(t) \leq \sum_{n=1}^{N} Q_{A_n}^B(t), \quad \forall t \geq 0,$$

provided that $Q_{A_n}^B(0) = Q_{A_1}^B(0) = \ldots = Q_{A_N}^B(0) = 0$.

**Proof:** Given in Appendix II.

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: $\{\tau^{on}_n, \tau^{off}_n, n \geq 1\}, \{\tau^{off}_n, n \geq 1\}$. Define a point process $T^{off} = \sum_{i=1}^{n} (\tau^{on}_i + \tau^{off}_i)$, $n \geq 1$, $T^{off}_0 = 0$; 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

$$A^0(t) = r \quad \text{if } t \in \left[T^{off}_n - \tau^{on}_n, T^{off}_n\right], \quad n \geq 1,$$

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 $A^0(t) = 0$ can be extended to a stationary process on the whole real line [21]. We call that process $A(t)$. Note that the expected arrival rate $\rho$ 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 d.f. $F_r$ 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 $x \to \infty$.

**Proposition 2:** If $r > c > \rho$ and $\tau^{on}_r \in \mathcal{S}$, then as $B \to \infty$

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

**Proof:** In [22] it was shown that as $B \to \infty$

$$\Lambda^B \sim \frac{\mathbb{E}[\tau^{on}_r (r - c) - B]}{\mathbb{E}[\tau^{on}] + \mathbb{E}[\tau^{off}].} \quad (3)$$

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

$$G(0, t) = \frac{L(0, t)}{r - c},$$

and by ergodicity of $Q^B(t)$ and (3) as $B \to \infty$

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

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 $\tau^{on}_r \in \mathcal{S}$, then

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

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 3:** If $r > c_i > \rho$, $i = 1, 2$, $\tau^{on}_{r_i} \in \mathcal{T}_R$, and $\epsilon \in (0, 1)$, then for $m > l \geq 0$

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

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

$$0 \leq \frac{\mathbb{P}[Q^{B, c_1} \geq (1 - \epsilon)B]}{\mathbb{P}[Q^{B, c_2} \geq \epsilon B]} \leq \frac{\mathbb{P}[Q^{\infty, c_1} \geq (1 - \epsilon)B]}{\mathbb{P}[Q^{\infty, c_2} = B]}.$$
Next, Proposition 2 and Theorem 1 yield
\[
\lim_{B \to \infty} \mathbb{P}^m \left[ Q^{\infty,c_1} \geq (1 - \epsilon)B \right] = \mathbb{P}^m \left[ Q^{B,c_2} = B \right] = \lim_{B \to \infty} \mathbb{P}^{m,n} \left[ \tau^{on}_r > \frac{(1 - \epsilon)B}{R - c} \right] \leq \lim_{B \to \infty} \mathbb{P}^{m,n} \left[ \tau^{on}_r > \frac{B}{R - c} \right] \leq M \lim_{B \to \infty} \sup_{B} \mathbb{P}^{m-\delta} \left[ \tau^{on}_r > B \right] = 0,
\]
where \( M < \infty \); the second inequality is implied by Lemma 3. \( \blacksquare \)

The proposition below is the main technical result of this section. Due to space limitations the detailed proof is provided in [26].

**Proposition 4:** If \( r > c > \rho \) and \( \tau^{on}_r \in T_R \), then
\[
\lim_{c \uparrow 1} \lim_{B \to \infty} \mathbb{P} \left[ Q^B \geq \epsilon B \right] = \frac{1}{\mathbb{P} \left[ Q^B = B \right]}.
\]

### 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 \( A^{(i)}_j(t) \), \( 1 \leq i \leq M, 1 \leq j \leq n_i \) and the aggregate arrival process is denoted by \( A(t) = \sum_{i=1}^{M} \sum_{j=1}^{n_i} A^{(i)}_j(t) \). \( A^{(i)}_j(t) \) is the \( j \)th On-Off process of type \( i \) with On periods equal in distribution to \( \tau^{on,i}_j \); peak rate, average rate, and probability of the source being active are equal to \( (r_i, \rho_i, p_{on,i}) \), respectively. Random variables \( \{\tau^{on,i}_j, \tau^{on,i}_j\} \) 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], [27] and, most recently, in [24].

**Assumption 1:** The capacity of the queuing system satisfies the following
\[
c \notin \left\{ \sum_{i=1}^{M} [k_i(r_i - \rho_i) + n_i \rho_i] : k \in \mathbb{M} \right\},
\]
where \( k = (k_1, \ldots, k_M) \) 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 queuing performance, respectively.

Before starting and proving our main results we introduce two preparatory lemmas. The next lemma derives an asymptotic expression for the overflow probability in the special case when all sources need to be active in order to have a positive netput.

**Lemma 1:** Let \( R = \sum_{i=1}^{M} n_i r_i \). If \( 0 < R - c < r_i - \rho_i \), for \( 1 \leq i \leq M \), then for all \( B \geq 0 \) and \( 0 \leq \epsilon \leq 1 \)
\[
\lim_{B \to \infty} \prod_{i=1}^{M} p_{on,i}^{n_i} \left[ \frac{\tau^{on,i}_r > \epsilon B}{R-c} \right] \leq \mathbb{P} \left[ Q^{B,c}_{A^{(i)}} \geq \epsilon B \right] \leq \prod_{i=1}^{M} p_{on,i}^{n_i} \left[ \tau^{on,i}_r > \frac{B}{R-c} \right].
\]

In addition \( \tau^{on,i}_r \in S \) for \( 1 \leq i \leq M \), then as \( B \to \infty \)
\[
\mathbb{P} \left[ Q^{B,c} = B \right] \sim \prod_{i=1}^{M} p_{on,i}^{n_i} \left[ \tau^{on,i}_r > \frac{B}{R-c} \right].
\]

**Proof:** Let \( c_i = c - R + r_i \). Assume that at time \( t = 0 \) all considered queues are empty. For all \( 1 \leq i \leq M, 1 \leq j \leq n_i \), Proposition 1 yields
\[
Q^{B,c}_A(t) \leq Q^{B,c}_{A^{(i)}}(t) + Q^{B,c}_{A - A^{(i)}(j)}(t) = Q^{B,c}_{A^{(i)}_j}(t),
\]
which, by applying the operator \( \mathbb{P}[\cdot \geq \epsilon B] \), using independence of \( A^{(i)}_j \), and passing \( t \to \infty \), yields in stationarity
\[
\mathbb{P} \left[ Q^{B,c}_A \geq \epsilon B \right] \leq \prod_{i=1}^{M} p_{on,i}^{n_i} \left[ Q^{B,c}_{A^{(i)}} \geq \epsilon B \right].
\]

Obtaining the lower bound is straightforward from evaluating the system in stationarity at \( t = 0 \); for simplicity the time index is omitted. Let
\[
\alpha^{(i)}_j = \left\{ A^{(i)}_j = r_i, \tau^{on,i}_j > \frac{\epsilon B}{R-c} \right\},
\]
then,
\[
\mathbb{P} \left[ Q^{B,c}_A \geq \epsilon B \right] \geq \mathbb{P} \left[ Q^{B,c}_A \geq \epsilon B, \{\alpha^{(i)}_j, 1 \leq j \leq n_i\} \right] \geq \prod_{i=1}^{M} \left\{ \alpha^{(i)}_j, 1 \leq j \leq n_i \right\} \geq \prod_{i=1}^{M} p_{on,i}^{n_i} \left[ \tau^{on,i}_r > \frac{\epsilon B}{R-c} \right].
\]


Setting $\varepsilon = 1$ in the preceding upper and lower bounds and combining it with Proposition 2, we obtain the second statement of the proposition. \hfill \blacksquare

In order to state our second preliminary lemma and the 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 $\mathbf{e}_i = (e^{(i)}_1, \ldots, e^{(i)}_{n_i}) \in [0,1]^{n_i}$, for all $i$. In order to distinguish between scalar and vector quantities, vectors are denoted with bold letters. Let $|\mathbf{e}_i| = \sum_{j=1}^{n_i} e^{(i)}_j$.

Definition 1: Let $r_k = \sum_{i=1}^{M} k_i(r_i - \rho_i) + n_i \rho_i$. Here, we define the minimum overlap set

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

and the detailed minimum overlap set

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

Remark: The definition of $\mathcal{O}^*$ is similar to the definition of the minimal set in [21].

Next, let $\{X^{(i)}_j, 1 \leq j \leq n_i \}_{i=1}^{M}$ 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

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

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 $m_i \geq k_i$ for all $i$ and $m_j > k_j$ for at least one $j$.

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

$$\mathbb{P} \left[ \min_{\mathbf{e} \in \mathcal{O}^*} S_\mathbf{e} > y \right] \leq \prod_{m_i \in \mathcal{O}} \left( \frac{n_i}{m_i} \right) \mathbb{P}^{m_i} \left[ X^{(i)}_1 > \frac{y}{N} \right].$$

Proof: Here we prove the result for cases $M = 1$ and 2. An inductive proof for general $M$ is given in [26].

Define $D_i \equiv D_i(y)$ for $1 \leq i \leq M$ as the number of events $\{X^{(i)}_j > y/N\}$:

$$D_i \triangleq \sum_{j=1}^{n_i} \mathbf{1}[X^{(i)}_j > y/N].$$

For any $k \leq n_i$ the following simple union bound will be used repeatedly in this proof

$$\mathbb{P}[D_i \geq k] = \mathbb{P} \left[ \bigcup_{\mathbf{e} \in [0,1]^{n_i}} \bigcap_{j:e_j=1} \left\{ X^{(i)}_j > \frac{y}{N} \right\} \right] \leq \binom{n_i}{k} \mathbb{P}^k \left[ X^{(i)}_1 > \frac{y}{N} \right].$$

Case $M = 1$. 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 $\sum_{j=1}^{N} e_j = k_1$. Next, observe that

$$\{D_1 \leq k_1\} \cap \left\{ \min_{\mathbf{e} \in \mathcal{O}^*} S_\mathbf{e} > y \right\} = \emptyset$$

due to the fact that there is at least one $\mathbf{e} \in \mathcal{O}^*$ such that $S_\mathbf{e} \leq \frac{N-k_1}{N} y \leq y$ on $\{D_1 \leq k_1\}$. Thus, by inequality (5)

$$\mathbb{P} \left[ \min_{\mathbf{e} \in \mathcal{O}^*} S_\mathbf{e} > y \right] \leq \mathbb{P}[D_1 \geq k_1 + 1] \leq \binom{N}{k_1 + 1} \mathbb{P}^{k_1 + 1} \left[ X^{(1)}_1 > \frac{y}{N} \right],$$

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

Case $M = 2$. Note that for all $(m_1, m_2) \in \mathcal{O}$

$$\{D_1 \leq m_1, D_2 \leq m_2\} \cap \left\{ \min_{\mathbf{e} \in \mathcal{O}^*} S_\mathbf{e} > y \right\} = \emptyset,$$

for the same reason as in the previous case $M = 1$, and therefore,

$$\mathbb{P} \left[ \min_{\mathbf{e} \in \mathcal{O}^*} S_\mathbf{e} > y \right] \leq \mathbb{P} \left[ \left( \bigcup_{(m_1, m_2) \in \mathcal{O}} \{D_1 \leq m_1, D_2 \leq m_2\} \right) \right].$$

Now, let $k_1 = \max\{i : (i, j) \in \mathcal{O}\}$ and $l_1 = \min\{i : (i, j) \in \mathcal{O}\}$. Next, without loss of generality we can assume $r_1 - \rho_1 \geq r_2 - \rho_2$, which implies that $\mathcal{O}$ has a special form $\mathcal{O} = \{i, k_2\}, l_1 \leq i \leq k_1$, since it can be represented as

$$\mathcal{O} = \left\{ \mathbf{m} \in \mathcal{E} : \begin{array}{ll} c' < m_1 + \alpha m_2 < c + \alpha & \text{if } m_2 \neq 0 \\ c' < m_1 + \alpha m_2 < c + 1 & \text{if } m_2 = 0 \end{array} \right\},$$
where $\alpha \triangleq \frac{r_m - c}{r_m - \rho_1} \leq 1$ and $c' \equiv \frac{c n_i \rho_j - n_i \rho_1}{r_m - \rho_1}$. Then,

$$
\left( \bigcup_{(m_1, m_2) \in \mathcal{O}} \{D_1 \leq m_1, D_2 \leq m_2\} \right)^c = \{D_1 \geq k_1 + 1\} \cup \left( \bigcup_{i=k_1}^{k_2} \{D_1 = i, D_2 = k_2 + 1\} \right)
\subseteq \{D_1 \geq k_1 + 1\} \cup \left( \bigcup_{i=k_1}^{k_2} \{D_1 = i, D_2 = k_2 + 1\} \right).
$$

Next, we construct $\mathcal{O}$ as

$$
\mathcal{O} = \{(k_1 + 1, 0) \cap \mathcal{E}\} \cup \left( \bigcup_{i=k_1+1}^{k_2+1} (i, k_2 + 1) \right),
$$

where we use the notation $(i, k) \equiv (i, k)$. Clearly, $\mathcal{O}$ has the property that for each $m \in \mathcal{O}$ there is a $k \in \mathcal{O}$ such that $m > k$. Hence, equation (6), the preceding inclusion, the union bound, independence of $D_i$ and inequality (5) render

$$
\mathbb{P} \left[ \min_{e \in \mathcal{O}_*} S_e > y \right] 
\leq \mathbb{P}[D_1 \geq k_1 + 1] + \sum_{i=k_1}^{k_2+1} \mathbb{P}[D_1 \geq i, D_2 = k_2 + 1] 
\leq \left( \frac{n_1}{k_1 + 1} \right)^{k_1 + 1} \left( \frac{n_2}{k_2 + 1} \right)^{k_2 + 1} \mathbb{P} \left[ X_1^{(1)} > \frac{y}{N} \right] 
+ \sum_{i=k_1}^{k_2+1} \left( \frac{n_1}{i} \right) \mathbb{P} \left[ X_1^{(1)} > \frac{y}{N} \right] \mathbb{P} \left[ X_1^{(2)} > \frac{y}{N} \right] 
= \sum_{m \in \mathcal{O}} \prod_{i=1}^{\mathcal{O}} (n_i) \mathbb{P} \left[ X_1^{(1)} > \frac{y}{N} \right].
$$

Recall $(\frac{n_i}{k_1 + 1}) \equiv 0$ when $k_1 + 1 > n_i$. This finishes the proof of the case $M = 2$.

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

**Theorem 2:** If $\sum \rho_i n_i < c$ and $\tau_{on}^{n(i)} \in \mathcal{T}R$ for $1 \leq i \leq M$, then under Assumption 1 as $B \to \infty$

$$
\mathbb{P}[Q = B] \sim \hat{P}(B) \equiv \sum_{m \in \mathcal{O}} \prod_{i=1}^{\mathcal{O}} \left( \frac{n_i}{m_i} \right)^{m_i} \mathbb{P} \left[ \tau_{on}^{n(i)} > \frac{B}{r_m - c} \right],
$$

where $r_m = \sum_{j=1}^{M} (m_j (r_j - \rho_j) + n_j \rho_j)$.

**Remarks:** Complementary results for the infinite buffer model are recently obtained in [24]. A related result for a discrete time finite buffer queue with an $M/G/\infty$ arrival process can be found in [27].

**Proof:** Due to the space limitation we provide the proof only for the loss probability. The proof for the loss rate can be completed in the same spirit and is provided in [26]. The following proof consists of a lower and an upper bound.

**Upper bound.** Let $c_e \equiv c - \sum_{i=1}^{M} \sum_{j=1}^{n_i} (1 - e_j^{(i)}) \rho_i$ and $A_e \equiv \sum_{i=1}^{M} \sum_{j=1}^{n_i} e_j^{(i)} A_j^{(i)}$. For $\delta > 0$ consider the queues $Q_{A_e}^{B_e - \delta}$, $e \in \mathcal{O}_*$, $Q_{A_j^{(i)}}^{B_j + \delta / N}$. Assume that these queues are empty at time $t = 0$. For any $e \in \mathcal{O}_*$ and sufficiently small $\delta > 0$ such that all queues have their capacity greater than the average arrival rate Proposition 1 yields ($Q_{A_e}^{B_e} \equiv Q_B$)

$$
Q_{A_e}^{B_e, \delta}(t) \leq Q_{A_e}^{B_e - \delta}(t) + \sum_{i=1}^{M} \sum_{j=1}^{n_i} (1 - e_j^{(i)}) Q_{A_j^{(i)}}^{B_j + \delta / N}(t),
$$

and thus

$$
Q_{A_e}^{B_e, \delta}(t) \leq \min_{e \in \mathcal{O}_*} \left( Q_{A_e}^{B_e - \delta}(t) + \sum_{i=1}^{M} \sum_{j=1}^{n_i} (1 - e_j^{(i)}) Q_{A_j^{(i)}}^{B_j + \delta / N}(t) \right).
$$

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

$$
\mathbb{P}[Q = B] \leq \sum_{m \in \mathcal{O}} \prod_{i=1}^{\mathcal{O}} \left( \frac{n_i}{m_i} \right)^{m_i} \mathbb{P} \left[ \tau_{on}^{n(i)} > \frac{B}{r_m - c} \right] \mathbb{P} \left[ \min_{e \in \mathcal{O}_*} \left( Q_{A_e}^{B_e - \delta} + \sum_{i=1}^{M} \sum_{j=1}^{n_i} (1 - e_j^{(i)}) Q_{A_j^{(i)}}^{B_j + \delta / N} \right) \geq B \right].
$$

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

$$\mathbb{P}[Q_{A,B,C} = B] \leq \mathbb{P}\left[ \bigcup_{\alpha \in \mathcal{O}} \left\{ Q_{A_{\alpha},B,C} - \delta \geq \epsilon B \right\} \right]$$

$$+ \mathbb{P}\left[ \min_{\alpha \in \mathcal{O}} \sum_{i=1}^{M} \left( 1 - e_j^{(i)} \right) Q_{A_{\alpha},B,C} > (1 - \epsilon) B \right] \leq \sum_{\alpha \in \mathcal{O}} \mathbb{P}\left[ Q_{A_{\alpha},B,C} - \delta \geq \epsilon B \right]$$

$$+ \sum_{\alpha \in \mathcal{O}} \left( \frac{M}{n_i} \right) \prod m_i \left[ Q_{A_{\alpha},B,C} - \delta \geq \epsilon B \right]$$

$$\leq \prod_{\alpha \in \mathcal{O}} \left( \frac{M}{n_i} \right) \prod m_i \left[ Q_{A_{\alpha},B,C} - \delta \geq \epsilon B \right]$$

where $c_m \triangleq c_e - \sum_{j=1}^{M} m_j r_j + r_i$. Now, inequality (7), in conjunction with Proposition 3 and Lemma 2, results in

$$\limsup_{B \to \infty} \frac{\mathbb{P}[Q_{A,B,C} = B]}{P(B)} \leq \frac{1}{\prod_{\alpha \in \mathcal{O}} \left( \frac{M}{n_i} \right) \prod m_i \left[ Q_{A_{\alpha},B,C} - \delta \geq \epsilon B \right] \prod m_i \left[ Q_{A_{\alpha},B,C} - \delta \geq \epsilon B \right]}$$

Here, by recalling that Proposition 4 implies for all $m$ and $i$

$$\lim_{\delta, 0 \to 1} \limsup_{B \to \infty} \frac{\mathbb{P}\left[ Q_{A_{\alpha},B,C} - \delta \geq \epsilon B \right]}{\mathbb{P}\left[ Q_{A_{\alpha},B,C} - \delta \geq \epsilon B \right]} = 1,$$

which, by Proposition 2 and Lemma 5, yields

$$\lim_{\delta, 0 \to 1} \limsup_{B \to \infty} \frac{\mathbb{P}\left[ Q_{A_{\alpha},B,C} - \delta \geq \epsilon B \right]}{\mathbb{P}\left[ Q_{A_{\alpha},B,C} - \delta \geq \epsilon B \right]} = 1.$$
Corollary 1: Homogeneous sources ($M = 1$). If $\rho N < c < N r$, $\tau_m^c \equiv \tau_m \in T \mathcal{R}$, and there is an integer $k_0$ such that $0 < r_0 - c < r - \rho$ with $r_0 = k_0 r + (N - k_0) \rho$, then as $B \to \infty$

$$\mathbb{P}[Q^B_A = B] \sim \left(\begin{array}{c} N \\ k_0 \end{array}\right) p_{\rho m}^{k_0} p_{\rho 0} \left[\tau_m^c > \frac{B}{r_0 - c}\right],$$

and

$$\Lambda^B \sim (r_0 - c) \mathbb{P}[Q^B_A = B].$$

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 [27], [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 $\tilde{a}(t) \triangleq a([t] + 1)$, where $[t]$ denotes the integer part of $t$, and the corresponding fluid queue process $Q^B(t) \equiv Q^B_{\tilde{a}}(t)$. Then, simple sample path arguments yield for all $t \geq 0$

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

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 utmost importance both for network provisioning and admission control.

The first two examples demonstrate the accuracy of Theorem 2, or more precisely Corollary 1. Since the asymptotic results are insensitive to the distribution of Off periods we choose their distribution to be exponential $\mathbb{P}[\tau^{off} \leq x] = e^{-\lambda x}, x \geq 0$; On periods are selected from Pareto family $\mathbb{P}[\tau^{on} \geq x] = 1/x^\alpha, x \geq 1, \alpha > 1$. We select $\alpha$ in the range of recently measured file sizes ($\alpha = 1.44$), which we presented in Figure 1. Now, the asymptotic approximation from Corollary 1 computes explicitly to

$$\hat{P}(B) = \left(N \cdot \frac{p_{\rho m}}{\alpha B^{\alpha - 1}} (k_0 r + (N - k_0) \rho - c)^{\alpha - 1}\right)^{k_0},$$

(13)

where

$$p_{\rho m} = \frac{\lambda \alpha}{\lambda \alpha + \alpha - 1} \quad \text{and} \quad \rho = r p_{\rho m}.$$

To ensure an increased accuracy of our simulation experiments we select the length of the simulated sample path to be $t = 10^{10}$. Each experiment took approximately four hours on a Pentium III PC.

Example 1: Here, we select $N = 10$ i.i.d. On-Off sources with parameters $\lambda = 0.012, \alpha = 1.3$ and $r = 2$, which yield $p_{\rho m} = 0.05$ and $\rho = 0.1$. For the choice of capacity $c = 5$ we simulated the overflow probabilities for buffer sizes $B = 100 \times i, i = 1, \ldots, 10$. The results of the simulation are presented in Figure 2 with "+" symbols. The selected parameters render $k_0 = 3$ for the asymptotic approximation $\hat{P}(B)$ as defined in (13). The approximation is plotted on the same figure with dashed lines. Its striking accuracy is apparent. Furthermore, it is interesting to observe that, although the distribution of On periods of the arrival processes has a very heavy tail ($\alpha = 1.3$), it is possible to achieve high multiplexing gains and small overflow probabilities even with moderately large buffers.

---

![Fig. 2. Illustration for Example 1](image-url)
The results of the simulation and approximation are plotted with "+" symbols and dashed lines, respectively. Again, the approximation matches the simulated estimates accurately.

![Fig. 3. Illustration for Example 2](image)

Next, we describe how our result can be used for efficient computation of capacity regions for network multiplexers. Capacity region consists of all combinations of arrival streams that, when fed into a multiplexer of specified capacity, produce required QoS. Let $Q_n^{B,c}$, $n = (n_1, \ldots, n_M)$ be the workload in a multiplexer with capacity $c$, buffer $B$ and arrival sequence which consists of $n_j (1 \leq j \leq M)$ 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 $\delta$, then the capacity (i.e., admissible) region is defined as

$$C = \{n_1, \ldots, n_M : \mathbb{P}[Q_n^{B,c} = B] \leq \delta\}.$$  

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 $P(B) \equiv P$ as defined in Theorem 2 is easily computable. Our implementation in Matlab produced instant real-time answers. Here, we provide two simulation studies in which we check the correctness of the asymptotic method in computing $C$. These simulation studies were much lengthier than in the preceding examples and, therefore, we had 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 3:** We select 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 $\rho_1 = 1.04$, $\rho_2 = 1.5$. 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 $\delta$ is set to $10^{-5}$. The outcome of the experiment is presented on Figure 4 with "+" symbols. The experiment took seven hours on a Pentium III PC. On the same figure with symbols "o" we indicated the approximation of $C$ 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. 4. Illustration for Example 3](image)

**Example 4:** Consider the previous example with loss 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 $\rho_1 = 5$, $\rho_2 = 4$. The capacity region for this case is presented in Figure 5. 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 2 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.
where $F^{2^*}$ denotes the 2-nd convolution of $F$ with itself, i.e., $F^{2^*}(x) = \int_{[0,\infty]} F(x-y) F(y) dy$.

It is well known that $S \subset \mathcal{L}$ [29]. A recent survey on subexponential distributions can be found in [30]. The class of intermediately regularly varying distributions $\mathcal{I}R$ is a subclass of $S$. Pareto distributions are well known examples from $\mathcal{I}R$.

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

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

The following three basic lemmas on $\mathcal{I}R$ distributions are useful for analysis.

**Lemma 3**: Let $F \in \mathcal{I}R$, $\eta \in (0,1)$, then

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

**Lemma 4**: If $F_r \in \mathcal{I}R$, then

$$\lim_{x \to \infty} \frac{x \mathbb{P}[X \geq x]}{\mathbb{P}[X_r \geq x]} < \infty.$$ 

**Proof**: For any $\delta \in (0,1)$ by definition of $F_r$

$$x \mathbb{P}[X \geq x] \leq \mathbb{P}[X_r \geq \delta x] \mathbb{P}[X \geq x] \mathbb{E}X$$

$$\leq \mathbb{P}[X_r \geq \delta x] \mathbb{E}X \leq \mathbb{P}[X_r \geq \delta x] \mathbb{E}X \leq \frac{x \mathbb{P}[X \geq x]}{1 - \delta}.$$ 

By Lemma 3 the result follows

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

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

**Lemma 5**: If $F_1, F_2 \in \mathcal{I}R$, then

(i) $F_1 F_2 \in \mathcal{I}R$, 
(ii) $w_1 F_1 + w_2 F_2 \in \mathcal{I}R$, for $w_1, w_2 > 0$.

**II. Proof of Proposition 1**

Let

$$0 = T_0 < T_1 < T_2 \cdots < T_m < T_{m+1} \cdots \text{ a.s.}$$

be the jump points in $\Lambda(t)$. Then, by assumption and (2), the statement holds for any $t \in [0, T_1]$

$$Q^{Bc}_A(t) = \left( t \sum_{n=1}^{N} (A_n(0) - c_n) \right)^+ \wedge B$$

$$\leq \sum_{n=1}^{N} (t(A_n(0) - c_n))^+ \wedge B = \sum_{n=1}^{N} Q^{Bc}_A(t),$$

The following three basic lemmas on $\mathcal{I}R$ distributions are useful for analysis.
where the last inequality follows from

\[
\left( \sum_{n=1}^{N} x_n \right) ^+ \land B \leq \left( \sum_{n=1}^{N} x_n^+ \right) ^+ \land B \leq \sum_{n=1}^{N} x_n^+ \land B.
\]

(14)

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

\[
Q_{A}^{B,c}(t) \leq \left( \sum_{n=1}^{N} \left( Q_{A, n}^{B,c_n}(T_m) + (t-T_m)(A_n(T_m)-c_n) \right) \right) ^+ \land B \leq \sum_{n=1}^{N} Q_{A, n}^{B,c_n}(t)
\]

and, therefore, it holds for all \( t \). This concludes the proof.

References


