Performance Evaluation of a Queue Fed by a Poisson Lomax Burst Process

Chen, Jiongze; Addie, Ronald G.; Zukerman, Moshe; Neame, Timothy D.

Published in:
IEEE Communications Letters

Published: 01/03/2015

Document Version:
Post-print, also known as Accepted Author Manuscript, Peer-reviewed or Author Final version

License:
Unspecified

Publication record in CityU Scholars:
Go to record

Published version (DOI):
10.1109/LCOMM.2014.2385083

Publication details:
Chen, J., Addie, R. G., Zukerman, M., & Neame, T. D. (2015). Performance Evaluation of a Queue Fed by a Poisson Lomax Burst Process. *IEEE Communications Letters*, *19*(3), 367-370. Advance online publication. https://doi.org/10.1109/LCOMM.2014.2385083

Performance Evaluation of a Queue Fed by a Poisson Lomax Burst Process

Jiongze Chen, Ronald G. Addie, Moshe Zukerman and Timothy D. Neame

***Abstract*—We consider a special case of the M/G/∞ traffic process, named as Poisson Lomax Burst Process (PLBP), where the burst lengths follow the Lomax distribution. We illustrate its advantage in modelling Internet traffic flow sizes, particularly, its ability to capture a large number of small flows. We provide two approximations based on analytical and fast simulation methods for the overflow probability of a single server queue fed by PLBP, and illustrate their accuracy by discrete-event simulations.**

***Index Terms*—Long Range Dependence Process, Pareto, Lomax, Queueing analysis, Traffic modelling**

I. Introduction

THERE has been extensive research on Internet traffic modelling following the discovery of its Long Range Dependence (LRD) characteristics [1]. The Poisson Pareto Burst Process (PPBP) [2]–[4] has been considered a suitable model because of its inherent structure and statistical characteristics. PPBP consists of bursts with Pareto distributed length that arrive according to a Poisson process. Several authors have referred to this as an M/G/∞ input process [5], [6] with Pareto distributed service times. There has been a significant research effort on the performance analysis of queues fed by PPBP [3]–[8]. Of particular interest was a discrete G/D/1 system (considered also here) fed by PPBP. Here, we replace the Pareto burst distribution of PPBP with a Lomax distribution [9]–[12] so that small traffic flows are taken into account. The resulting input process is named the *Poisson Lomax Burst Process* (PLBP).

Notice that the Pareto distribution is characterized by two parameters – the shape parameter, $\gamma > 0$, and the scale parameter, $\delta > 0$. Meanwhile, δ also serves as the lower bound of the Pareto random variable, which implies that the modeled flow sizes have a positive minimum value. The fact that δ plays the roles of both the shape parameter and the minimum flow size introduces a certain difficulty. As a scale parameter, δ is one of the key parameters of the Pareto distribution that together with the shape parameter and possibly other traffic parameters are often selected to fit certain statistical characteristics of a process or distribution which are of interest in traffic modelling applications. Such fitting normally results in δ taking a value that is too large to be appropriate for the minimum flow size which can be only few bytes. To release the scale parameter δ from its second role as the minimum flow size, we adopt the *Lomax* distribution, which retains the Pareto heavy tail characteristics but also enables the capture of a large number of small traffic flows that are excluded using the traditional Pareto distribution.

The remainder of this paper is organized as follows. In section II, we clarify the definition of PLBP and provide its key statistical parameters. In Section III, we describe approximation and simulation methodologies for evaluation of the overflow probability of a PLBP queue. In Section IV, we validate the approximation by simulations and compare between PLBP and PPBP.

II. The Poisson Lomax Burst Process

Like PPBP, PLBP is based on a stream of bursts with burst arrivals following a Poisson process with rate λ [bursts/s]. In PLBP, burst durations are assumed to be independent Lomax distributed random variables with parameters γ and δ. Let d [s] be a random variable representing a burst duration. Assume that each burst generates work at a constant rate, r [B/s]. Because arrivals of new bursts may occur before the completion of existing bursts, multiple overlapping bursts may be transmitted simultaneously, and since the number of overlapping bursts is unlimited, this arrival process can be viewed as a special case of an M/G/∞ queueing model where the burst length (or the transmission time) are Lomax distributed. Figure 1 presents how PLBP is formed by the Exponential and Lomax distributions. In the following, the key parameters of the Lomax distribution and PLBP are provided.

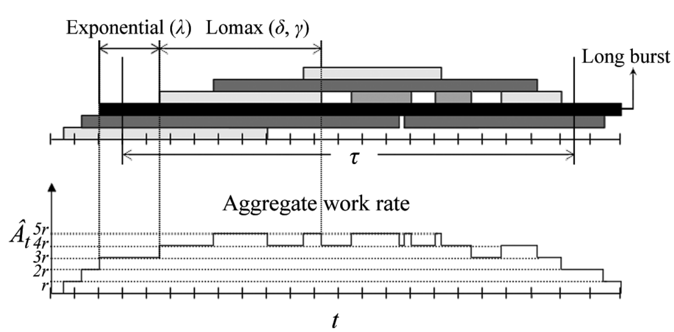

Fig. 1. Formation of the Poisson Lomax Burst Process.

A. The Lomax distribution

As a variant of Pareto, the Lomax distribution permits values arbitrarily close to 0 with non-zero probability. The complementary distribution function (CDF) of a Lomax random variable d is

$$P(d > x) = \left(1 + \frac{x}{\delta}\right)^{-\gamma}, \quad x \geq 0 \tag{1}$$

where δ is the *scale parameter* which differs from the original Pareto by no longer determining the minimum value of d, and

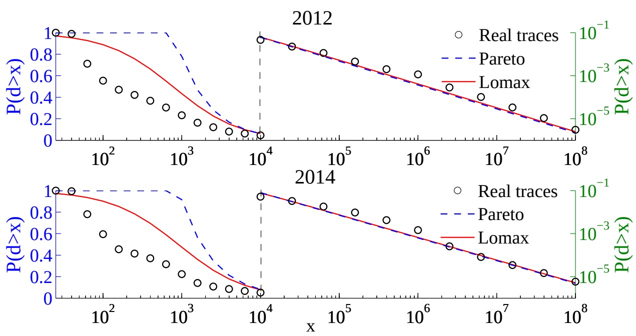

Fig. 2. The complementary distribution, $P(d > x)$, of Pareto (with $\gamma = 1.108$) and Lomax (with $\gamma = 1.054$) vs. real traces.

γ is the *shape parameter* that controls the tail behavior of the distribution. Also the mean of d is

$$E(d) = \begin{cases} \frac{\delta}{\gamma-1}, & \gamma > 1, \\ \infty, & \text{otherwise.} \end{cases} \quad (2)$$

Let ω be the Forward Recurrence Time (FRT) of a Lomax random variable with parameters, γ and δ. Its CDF is

$$P(\omega > x) = \int_x^\infty dP(\omega > t) = \left(1 + \frac{x}{\delta}\right)^{1-\gamma}. \quad (3)$$

It is understandable that δ is larger than the minimum flow size of some real networks where Lomax seems to be a good alternative for Pareto. Figure 2 shows one example when Lomax is a better choice than Pareto. The real traces are flows captured by CAIDA's equinix-sanjose monitors on high-speed Internet backbone links in 2012 and 2014 [13]. For simplicity, we assume $r = 1$. The relationship between γ and δ for Lomax is obtained by (2) to fit the mean flow size (8192.7 and 11015.8 [Byte]). Then we fit the curve of the real traces as shown in the plot, by the curve fitting tool of Matlab applying the nonlinear least squares method [14]. We obtain $\gamma = 1.108$ and $\delta = 884.8$ for 2012 data, and $\gamma = 1.054$, $\delta = 969.5$ for 2014 data. To ensure similar tail behavior, we retain the same value for γ of Pareto and the value of δ is estimated to fit the mean of the real traces. We observe in Figure 2 that Lomax fits better than Pareto for small flows ($< 10^4$ Bytes) while for larger flows, the tails of Pareto and Lomax are close to each other and reasonably close to the real traces (with Lomax slightly closer).

B. Statistics of PLBP

Let B_t be the number of active bursts contributing work at time t. By [15, p. 188], as the burst arrivals follow a Poisson process, B_t are Poisson-distributed, with mean $E(B_t) = \lambda E(d)$. As illustrated in Figure 1, the work contributed by all bursts during the interval $(0, t]$ can be expressed as $\hat{A}_t = r\int_0^t B_s ds$. Since we have assumed that all bursts contribute work at a constant rate, the mean of $\hat{A}_t$ is

$$\mu(t) \triangleq E(\hat{A}_t) = rtE(B_t) = rt\lambda E(d) = \frac{\lambda t r \delta}{\gamma - 1}. \quad (4)$$

In addition, the variance of $\hat{A}_t$ can be obtained by repeatedly integrating the CDF of the burst duration [16]:

$$v(t) \triangleq var(\hat{A}_t) = 2\lambda r^2 \int_0^t du \int_0^u dv \int_v^\infty P(d > x)dx. \quad (5)$$

Following the same reasoning as in [7]:

$$v(t) = \frac{2\lambda r^2 \delta^3}{(\gamma-1)(\gamma-2)}\left(\frac{t}{\delta} - \frac{(1+\frac{t}{\delta})^{3-\gamma}}{3-\gamma} + \frac{1}{3-\gamma}\right). \quad (6)$$

From (6), we see that for large t, the dominant term is

$$\frac{2\lambda r^2 \delta^3 (1+\frac{t}{\delta})^{3-\gamma}}{(\gamma-1)(\gamma-2)(\gamma-3)}.$$

Define

$$H = \frac{3-\gamma}{2},$$

and observe that the growth of $v(t)$ is proportional to t^{2H}. This implies that for $1 < \gamma < 2$, like PPBP, PLBP is asymptotically self-similar with Hurst parameter H.

III. Performance evaluation of the PLBP SSQ

Consider a single server queue (SSQ) with constant service rate, C [B/s], and an infinite buffer, fed by a PLBP input with parameters, λ, γ, δ and r, henceforth called PLBP SSQ. An SSQ with constant service rate has been often used as a model of statistical multiplexer used to multiplex variable bit rate (VBR) traffic from multiple sources into a single output (e.g. a wavelength). A related SSQ with constant service rate was used to study the performance of an ATM buffer in [8]. Also, simulations of an SSQ with constant service rate fed by a VBR video trace were used to explore the issue of resource allocation for statistically multiplexed video sources in [17] that revealed the LRD phenomenon of VBR video traffic. In this section we provide an algorithm for computation of the buffer overflow probability of a PLBP SSQ by the Quasi-Stationary (QS) approximation. A satisfactory approximation was obtained for the PPBP SSQ in [3], [4]. Concise details of applying the QS idea are given here while more detailed justifications or concepts can be found in [3], [4]. In addition, the simulation method is introduced in Section III-B.

A. The QS approximation

The QS algorithm makes use of an idea which was used in [18] to find the rate function for a large deviations characterization of multi-source heavy tailed on-off traffic as the number of sources increases. The idea is to separate the bursts into long and short. Considering a time period, τ, those bursts existing for the entire period are identified as long bursts and the rest are short bursts, as shown in Figure 1. We separate the process into long and short burst processes, denoted: L_τ and S_τ, respectively. Since the traffic contributed by each long burst is constant during the period τ, the performance of the PLBP SSQ is equivalent to that of a queue fed by S_τ with service rate reduced by L_τ. Thus for a given τ, the overflow probability of a PLBP SSQ ($P_\tau(Q > x)$) is

$$P_\tau(Q > x) = \sum_{n=0}^{\infty} P(\eta = n)P(Q_S > x), \quad (7)$$

where η is the number of long bursts within the period, Q and Q_S are the steady-state queue size of the queues fed by the PLBP input and S_τ, respectively.

Because the start-times for bursts form a Poisson process, and it is assumed that the duration of each burst is independent from all other aspects of the process, L_τ is statistically independent from S_τ. The number of long bursts is Poisson distributed with mean β, equal to the burst density (λ) times the probability that the forward recurrence time of the burst length distribution (Lomax) is longer than τ, which is

$$\beta = \lambda E(d)P(\omega > \tau) = \frac{\lambda\delta}{\gamma - 1}\left(1 + \frac{\tau}{\delta}\right)^{1-\gamma},$$

by (3). Then $P(\eta = n)$ can be expressed as

$$P(\eta = n) = e^{-\beta}\frac{\beta^n}{n!}. \tag{8}$$

So, the mean and variance of the long bursts process (L_τ) are

$$\mu_L(t) \triangleq E(L_\tau) = rt\beta, \quad t < \tau, \tag{9}$$

and

$$v_L(t) \triangleq var(L_\tau) = r^2t^2\beta, \quad t < \tau. \tag{10}$$

Now we consider the queue fed by the short burst process (S_τ). The service rate of the queue, c, is the original service rate deducted by the constant rate contributed by L_τ, $c = C - r\eta$. Since S_τ and L_τ are independent, the mean/variance of S_τ is equal to the mean/variance of PLBP reduced by the mean/variance of L_τ, giving

$$\mu_S(t) \triangleq E(S_\tau) = \mu(t) - \mu_L(t) = \mu(t) - rt\beta, \tag{11}$$

and

$$v_S(t) \triangleq var(S_\tau) = v(t) - v_L(t) = v(t) - r^2t^2\beta, \tag{12}$$

for $t < \tau$, where $\mu(t)$ and $v(t)$ are given by (4) and (6). As in [2], [3], assuming that S_τ is Gaussian, and using the approximation of [2], we obtain

$$P(Q_S > x) \approx K\exp\left(-\frac{(x + (c - \mu_S(1))t^*)^2}{2v(t^*)}\right), \tag{13}$$

where K is chosen so that for $x = 0$, the right hand side in (13) is equal to a Gaussian estimate of the probability that ηr exceeds the available capacity of the server, and t^* is a value of t (≥ 0) which minimizes the expression, $(x + (c - \mu_S(1))t^*)^2/2v(t^*)$.

Finally, we have the QS approximation for the overflow probability of the PLBP SSQ by exhausting all possible τ for (7) to find a global maximum of $P_\tau(Q > x)$, that is

$$P(Q > x) \approx \sup_{\tau\geq 0} P_\tau(Q > x). \tag{14}$$

B. A fast simulation method

Instead of a straightforward slow simulation which is well-known to be time-consuming when it involves LRD processes, we adopt the fast simulation method provided by [3] for time efficiency. We now consider a discrete-time PLBP SSQ. The details of the queue has been provided in the previous section. By dividing the whole simulation time T [s] into M sampling intervals, we can obtain the amount of work buffered in the queue at the end of the nth interval by Lindley's equation:

$$Q_m = \text{Max}\left\{Q_{m-1} + A_n - C, 0\right\}, m = 1, \dots, M, \tag{15}$$

where $Q_0 = 0$ and A_m is the work contributed by PLBP within the mth interval. In this section, we will demonstrate the procedure of obtaining Q_m by the fast simulation method.

The only difference between the fast simulation and the slow simulation is that we assume there already exist η long bursts when the simulation is initialized. By simulation, we can obtain the estimation of the overflow probability of the queue started with η long bursts, denoted: $\hat{P}(Q_{(\eta,T)} > x)$. Then the overflow probability of the PLBP SSQ is

$$P(Q > x) = \sum_{n=0}^{\infty} P(\eta = n)\hat{P}(Q_{(\eta,T)} > x), \tag{16}$$

where $P(\eta = n)$ is shown in (8). For practice, we set n going to N instead of ∞; N is the smallest number that makes $P(\eta = N) \leq \varepsilon$. Here, the tolerable error, ε, is set to be 10^{-9}.

In sum, for a selected set of x values, X, the fast simulation is accomplished by the following steps, started with $\eta = 0$:

1) Start a thread of simulation with initial η long bursts, and total period, T. Assign the short bursts to the queue. The number of the short bursts follows a Poisson process with mean as $\lambda E(d) - \beta$, and their durations are bounded Pareto FRT number with maximal value as T. Finally, feed the queue with normal PLBP input, monitor the queue size by (15), and obtain $\hat{P}(Q_{(\eta,T)} > x)$, $x \in X$.
2) Go to 3) if $\eta = N$, else set $\eta = \eta + 1$ and go to 1).
3) Obtain $P(Q > x)$, $x \in X$, by (16).

The time to reach a consistent state for traditional simulation for a LRD queue is very long. The merit of the fast simulation method is that it reduces this time by initializing the system in a range of states.

IV. Numerical Results

In this section, we validate the QS approximation with simulation results. Also, comparisons between the PLBP and the PPBP SSQs are presented. All simulations results are provided with 95% confidence intervals based on Student's t-distribution, which are presented as the coloured area in the figures. In Figure 3, we compare the overflow probabilities of a PLBP SSQ obtained by the QS approximation with the simulation results from both the slow and the fast simulation methods, for various cases involving a range of values for H, r and C. Table I shows the total times required to obtain the overflow probabilities for the approximation and the simulations, for all selected x values, in each of the six cases presented in Figure 3. Notice the advantage of the fast simulation in speed over the slow simulation, especially for $H = 0.85$. As expected, the approximation takes far less time, which does not vary much with H. The consistency of the approximation and the two simulation methods can be found in the plots. However, there exist discrepancies between the QS approximation and the simulations for some x values. The approximation relies on modelling the short burst process as Gaussian. In cases where the short-burst process plays a major role in determining performance, this may lead to inaccuracy. This is probably the main cause of discrepancy between the QS approximation and simulation results. Observe that the results of the fast simulation, that do not rely on this assumption, are consistent with those of the slow simulation. However, for large H, both slow and fast simulations of LRD queues may require even longer (impractically long) run-times to achieve the required level of accuracy.

TABLE I
TOTAL TIMES FOR OBTAINING $P(Q > x)$ FOR ALL SELECTED x VALUES.

	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
Approx.	0.2 s	0.3 s	0.4 s	0.6 s	0.7 s	1.2 s
Fast sim.	53 min	14 min	4 h	21 min	48 min	2 h
Slow sim.	93 min	46 min	6 days	33 min	2 h	4 days

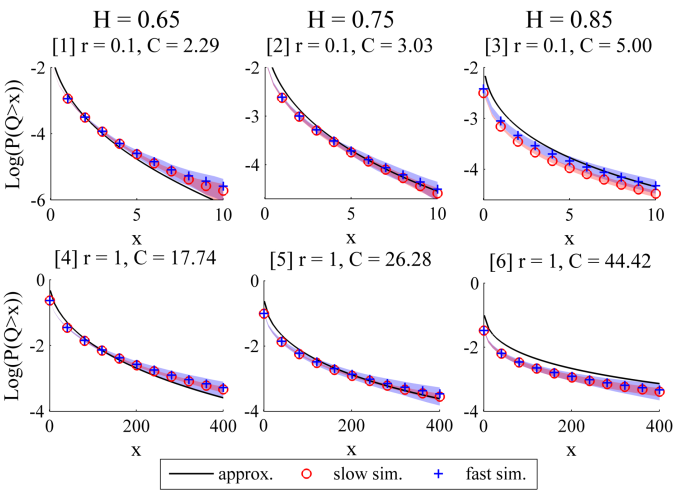

Fig. 3. Overflow probabilities based on QS approximation vs. simulation results for Case 1 to 6, with $\lambda = 10$ and $\delta = 1$ for all cases.

Figure 4 illustrates the differences in overflow probabilities obtained by PPBP and PLBP for the same fitted key parameters values. We set $\gamma = 1.5$ for both processes so that the Hurst parameter, $H = 0.75$. The plot shows two cases with $\lambda = 1$, 10. For each case, the parameters of both PPBP and PLBP are chosen to fit the same mean and variance as shown in Table II. The capacities (C) of both queues are set as Mean + $3\sqrt{\text{Variance}}$. The overflow probabilities for both processes are obtained by simulations. We see that the overflow probabilities of the PLBP SSQ are higher than those of the PPBP SSQ for both cases. The reason is that while there are more small bursts for PLBP, these extra small bursts must be balanced by more large bursts to fit the mean, compared to PPBP, and this leads to more overflows. We observe that the performance prediction results of PLBP can be significantly more conservative than PPBP. Recalling the results in Figure 2 indicating that PLBP accurately represents measured traces, we propose PLBP as a useful tool in network performance modelling.

TABLE II
PARAMETERS OF PPBP AND PLBP.

λ	Mean	Variance	PPBP		PLBP	
			δ	r	δ	r
1	1.836	1	1	0.612	1.537	0.597
10	5.82	1	1	0.194	1.545	0.188

V. CONCLUSION

We have motivated the use of PLBP by using real traces to demonstrate that Lomax overcomes a weakness of Pareto in better representing short bursts while retaining similar heavy tail behavior. Then we have provided an analytical approximation and a fast simulation method for evaluating the overflow probability of a PLBP SSQ. In all the cases considered, the accuracy of the fast simulation method was demonstrated by showing that its results are very close to those obtained by slow simulations. We have observed some inaccuracies of the approximation which were explained by its Gaussian assumption of the short bursts.

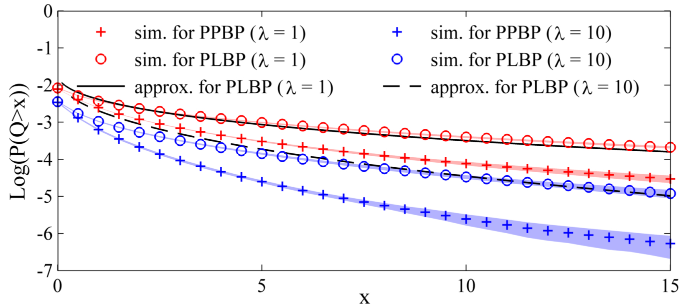

Fig. 4. Overflow probabilities for PPBP and PLBP SSQs.

REFERENCES

[1] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of Ethernet traffic (extended version)," *IEEE/ACM Trans. Networking*, vol. 2, no. 1, pp. 1–15, Feb. 1994.
[2] R. G. Addie, P. Mannersalo, and I. Norros, *Performance formulae for queues with Gaussian input*, ser. Teletraffic Engineering in a Competitive World. Elsevier Science, Jun. 1999, pp. 1169–1178.
[3] R. G. Addie, T. D. Neame, and M. Zukerman, "Performance evaluation of a queue fed by a Poisson Pareto burst process," *Computer Networks*, vol. 40, no. 3, pp. 377–397, Oct. 2002.
[4] ——, "Performance analysis of a Poisson-Pareto queue over the full range of system parameters," *Computer Networks*, vol. 53, no. 7, pp. 1099–1113, May 2009.
[5] N. G. Duffield, "Queueing at large resources driven by long-tailed M/G/∞-modulated processes," *Queueing Syst.*, vol. 28, no. 1-3, pp. 245–266, May 1998.
[6] M. M. Krunz and A. M. Makowski, "Modeling video traffic using M/G/∞ input processes: a compromise between Markovian and LRD models," *IEEE J. Sel. Areas Commun.*, vol. 16, pp. 733–748, Jun. 1998.
[7] T. D. Neame, M. Zukerman, and R. G. Addie, "Modelling broadband traffic streams," in *Proc. GLOBECOM '99*, Nov. 1999.
[8] N. Likhanov, B. Tsybakov, and N. D. Georganas, "Analysis of an ATM buffer with self-similar ("fractal") input traffic," in *Proc. IEEE INFOCOM '95*, vol. 3, Apr. 1995, pp. 985–992.
[9] K. Lomax, "Business failures: Another example of the analysis of failure data," *J. Amer. Statist. Assoc.*, vol. 49, no. 268, pp. 847–852, 1954.
[10] O. Holland, A. Golaup, and A. Aghvami, "Traffic characteristics of aggregated module downloads for mobile terminal reconfiguration," *IEE Proc.-Commun.*, vol. 153, no. 5, pp. 683–690, 2006.
[11] M. Weigle, "Improving confidence in network simulations," in *Proc. WSC '06*, 2006, pp. 2188–2194.
[12] C. Mattos, E. Ribeiro, E. Fernandez, and C. Pedroso, "An unified VoIP model for workload generation," *Multimedia Tools and Applications*, vol. 70, no. 3, pp. 2309–2329, 2014.
[13] "The CAIDA UCSD Anonymized Internet Traces 2012 (2014) - [20120216 (20140320)]." [Online]. Available: http://www.caida.org/data/passive/passive_2012(2014)_dataset.xml
[14] T. F. Coleman and Y. Li, "An interior trust region approach for nonlinear minimization subject to bounds," *SIAM Journal on optimization*, vol. 6, no. 2, pp. 418–445, 1996.
[15] D. R. Cox and V. Isham, *Point processes*. Chapman and Hall, 1980.
[16] J. Roberts, U. Mocci, and J. Virtamo, *Broadband network teletraffic: performance evaluation and design of broadband multiservice networks*, ser. Lecture Notes in Computer Science. Springer Verlag, Nov. 1996.
[17] M. W. Garrett and W. Willinger, "Analysis, modeling and generation of self-similar VBR video traffic," *ACM Sigcomm Comput. Commun. Rev.*, vol. 24, no. 4, pp. 269–280, Oct. 1994.
[18] M. Mandjes, "Overflow behavior in queues with many long-tailed inputs," *Adv. in Appl. Probab.*, vol. 32, no. 4, pp. 1150–1167, 2000.