An insensitivity property for light-tail shot noise traffic overflow asymptotics

Pierre Brémaud[1]

ABSTRACT

In this short note, we derive the large deviations estimate of the tail of the buffer occupancy distribution in a communications link with a very general integrated Poisson shot noise model for the total input. The result is obtained by a straightforward application of the general theory developped in [KWC93], [GlWh94] and [DuO'C95]. The interesting outcome of these computations is that, in the light-tail case, it is largely model independent, and this makes the statistical analysis of traffic in view of link dimensioning, with a procedure implementable by a simple network-based software.

1 Introduction

In the fluid models of traffic, the total number of packets presented to a communications link during the time interval $(a, b]$ is $\int_a^b X(s)ds$, where $X(t)$ is the traffic intensity at time t, whereas in the hard model, it is $\sum_n \sigma_n \ 1_{(a,b]}(T_n)$ where σ_n is the work brought at time T_n. The fluid model is justified as long as the packet size is very small compared to the average traffic intensity, as is indeed the case in broadband communications networks.

It remains to choose a realistic probabilistic model for the traffic intensity, and one which is also analytically tractable. Several have been proposed in the litterature, for instance on-off and Markov fluids [KWC93], Ornstein-Uhlenbeck models [KuRo94], and fractal brownian

[1]Laboratoire des Signaux et Systèmes, CNRS, France, and Département Systèmes de Communications, EPFL, Switzerland

motion models [No94]. In this article, we take for the integrated traffic an integrated Poisson shot noise, which encompasses fluid as well as hard models, in particular, the usual hard M/GI input and the fluid $M/GI/\infty$ input. However we consider the very general model, leaving the identification problem aside for the time being.

We are interested in computing the rate of exponential decay (with respect to B) of the probability of overshoot of level B in a link of given capacity c, which in turn gives the effective bandwidth ([Ke91]). Talking about exponential rates implies some kind of light-tail and/or short dependence assumption, which will be explicited in section 3. The large deviations asymptotics of the overshoot probability are obtained as a direct applications of the results in [KWC93], [GlWh94] and [DuO'C95] and of well known formulas concerning Poisson shot noise.

The result is that the rate of decay does not depend on the fine details of the utterly general model we consider. It depends only on the distribution of the total number of packets in a typical shot of the Poisson shot noise, it does not depend on the rate of delivery of the shot. We show how this feature can be exploited in model independent statistical analysis leading directly to an estimate of the effective bandwidth using simple software tools.

2 Poisson shot noise traffic

A Poisson shot noise (PSN) is a process of the form

$$X(t) = \sum_n h(t - T_n, X_n)\ 1_{(0,t]}(T_n),$$

where $\{T_n\}_{n\in\mathbb{Z}}$ is the sequence of times of a homogeneous Poisson process of rate λ, and $\{X_n\}_{n\in\mathbb{Z}}$is an i.i.d sequence of random elements of a measurable space $(E, \mathcal{E})$, independent of the Poisson process, and $h(t, x) = 0$ for negative times, and is otherwise non-negative. From

$$E[X(t)] = \lambda \int_0^\infty E[h(s, X_1)]ds,$$

we deduce that the PSN, which is always well defined since the function h is non-negative, is finite if

$$\rho \stackrel{\text{def}}{=} \lambda \int_0^\infty E[h(t, X_1)]dt < \infty.$$

In a communications context ρ is called the traffic rate.

A typical example is the fluid $M/GI/\infty$ input, where $h(t, x) = 1_{\{t\le x\}}$ in which case $X(t)$ represents the number of virtual customers present at time t in a virtual pure delay system, a typical virtual customer arriving at time T_n and departing at time $T_n + X_n$.

A PSN is also called a (randomly) filtered Poisson process, with the interpretation that $h(t-T_n, X_n)$ is the random impulse response associated with spike, or impulse, at T_n.

The total (or integrated) input in the time interval $(0,t]$ is

$$A((0,t]) \;=\; \int_0^t X(s)ds,$$

and it takes the form

$$A((0,t]) \;=\; \sum_n H(t-T_n, X_n)\; 1_{(0,t]}(T_n) + \sum_n \left(H(t-T_n, X_n) - H(-T_n, X_n)\right) 1_{(-\infty,0]}(T_n), \tag{2.1}$$

where

$$H(t,x) \;=\; \int_0^t h(s,x)ds.$$

To account for traffic of a general type, not necessarily fluid, we adopt the integrated Poisson shot noise (IPSN) model (2.1) where

$$H(t,x) \;=\; \mu(x,[0,t]) \tag{2.2}$$

and $\mu(x,\cdot)$ is for each x a measure on the non-negative half-line.

If $\mu(x,\cdot) = x\delta(\cdot)$ where $\delta(\cdot)$ is the Dirac unit mass at 0, we have the input of the classical $M/GI/1/\infty$ queue. In this case we denote X_n by σ_n, and therefore

$$A((0,t]) \;=\; \sum_n \sigma_n 1_{(0,t]}(T_n).$$

In the general model

$$E[A([0,t))] \;=\; \lambda \int_0^t E[H(s,X_1)]ds \;+\; \lambda \int_0^\infty E[H(t+s,X_1) - H(s,X_1)]ds.$$

In particular if this quantity is finite for some t, then it is finite for all t, and proportional to t. The traffic intensity is then

$$\rho \;\stackrel{\text{def}}{=}\; \frac{1}{t}\left\{\lambda \int_0^t E[H(s,X_1)]ds\lambda + \int_0^\infty E[H(t+s,X_1) - H(s,X_1)]ds\right\}. \tag{2.3}$$

This model is very rich, because the marks X_n are of an arbitrary nature. The measure A is also called a Poisson cluster measure, where $\mu(X_n,\cdot)$ is the cluster measure at T_n. An important random variable is the total cluster size, or total input per shot,

$$H(\infty, X_1) \;=\; \mu(X_1, \mathbb{R}_+).$$

The total (integrated) input in the interval $(0,t]$ can be written as a sum

$$A((0,t]) \;=\; A^0((0,t]) + D((0,t]), \tag{2.4}$$

where

$$A^0((0,t]) \;=\; \sum_n H(t-T_n, X_n)\; 1_{(0,t]}(T_n), \tag{2.5}$$

is the transient total input in interval $(0,t]$, corresponding to an initially empty system, and

$$D((0,t]) \;=\; \sum_n \left(H(t-T_n, X_n) - H(-T_n, X_n)\right) 1_{(-\infty,0]}(T_n) \tag{2.6}$$

is the part of the total input in interval $(0,t]$ due to clusters initiated before time 0. Note that D and A^0 are independent. The transient total input measure process $A^0(t+\cdot)$ converges in distribution to the stationary measure A as $t \to \infty$ if and only if for all intervals $(a,b] \in \mathbb{R}_+$ the random variable

$$D((t+a,t+b]) \;=\; \sum_n \left(H(t+b-T_n, X_n) - H(t+a-T_n, X_n)\right) 1_{(-\infty,0]}(T_n)$$

accounting for that part of the traffic in the interval $(a,b]$ initiated before time 0, converges in distribution to 0 as $t \to \infty$. The characteristic function of this variable is

$$\exp\left(\lambda \int_0^\infty E[e^{iu(H(s+t+b,X_1)-H(s+t+a,X_1)} - 1]ds\right)$$

and therefore a necessary and sufficient condition for convergence in distribution is that for all intervals $(a,b] \in \mathbb{R}_+$,

$$\int_0^\infty E[e^{iu(H(s+t+b,X_1)-H(s+t+a,X_1)} - 1]ds$$

tends to zero as t tends to infinity.

Note that in some occasions, $D((\tau,\infty) = 0$ for some almost surely finite random time τ. In this case the transient and the stationary total input measures couple, and in particular the convergence is in variation. This is the case for the M/GI/∞ process. A necessary and sufficient condition for this is that the support $[0, Z_1]$ of the function $h(t, X_1)$ verifies $E[Z_1] < \infty$. This follows from the consideration of the associated M/GI/∞ process with typical delay Z_1.

3 Effective bandwidth

We consider a communications link of capacity c. We assume the stability condition

$$\rho \ < \ c \tag{3.1}$$

which guarantees existence of a stationary buffer content W when the buffer is infinite. For finite buffer capacity B the overflow probability is overestimated by $P(W > B)$. The stationary buffer content is

$$W \ = \ \sup_{t\geq 0} \{A((-t,0]) - ct\} \, . \tag{3.2}$$

Let $\{W(t)\}_{t\geq 0}$ be the stationary buffer content process, i.e $W(0) = W$, and let $\{W_0(t)\}_{t\geq 0}$ be the transient buffer content starting empty, i.e $W_0(0) = 0$. The former dominates the latter, and both couple as soon as the stationary process becomes null. Therefore in order to show that coupling occurs in finite time, and therefore the transient process converges in variation to the stationary one, it suffices to show that $P(W = 0) > 0$. But we have the inequality

$$W(t) \leq W(0) + A((0,t]) - c \int_0^t 1_{W(s)>0},$$

from which it follows that

$$P(W = 0) \ \geq \ 1 - \frac{\rho}{c} > 0. \tag{3.3}$$

We now assume that the traffic is light-tailed, that is

$$E\left[e^{\theta H(\infty, X_1)}\right] \ < \ \infty, \tag{3.4}$$

for all θ in a neighborhood of 0. Also, we assume that in the same neighborhood of 0,

$$\lim_{t\uparrow\infty} \frac{1}{t} \int_0^\infty E[e^{\theta(H(t+s,X_1)-H(s,X_1))} - 1]ds \ = \ 0 \tag{3.5}$$

It then holds that

$$\lim_{B\uparrow\infty} \frac{1}{B} \ln(P(W > B)) \ = \ -R \tag{3.6}$$

where $R > 0$ is the solution of

$$\lambda E\left[e^{\theta H(\infty, X_1)} - 1\right] \ = \ c\theta. \tag{3.7}$$

Proof. We use the general results of [GlWh94] and [DuO'C95], which give

$$R = \sup\{\theta \ ; \ \lambda(\theta) \leq 0\}, \tag{3.8}$$

where

$$\lambda(\theta) \;=\; \lim_{t\uparrow\infty} \frac{1}{t} \ln E\left[e^{\theta(A([0,t))-ct)}\right]. \tag{3.9}$$

Straightforward computations (see section 2.1 of [KlMi95]) show that

$$\frac{1}{t}\ln E[e^{\theta A([0,t))}] = \lambda\frac{1}{t}\int_0^t E[e^{\theta H(s,X_1)}-1]ds \;+\; \lambda\frac{1}{t}\int_0^\infty E[e^{\theta(H(t+s,X_1)-H(s,X_1))}-1]ds. \tag{3.10}$$

Under the light-tailed assumption (3.4), the first term in the left-hand side converges to

$$\lambda E[e^{\theta H(\infty,X_1)}-1],$$

whereas the second term tends to 0 under the condition (3.5). □

Condition (3.5) is implied by

$$\lambda E\left[\int_0^\infty \left(e^{\theta(H(\infty,X_1)-H(s,X_1))}-1\right)ds\right] \;<\; \infty, \tag{3.11}$$

This condition implies the convergence of the transient traffic process to the stationary traffic process.

One should start by checking the simpler condition (3.11). Note however that it may happen that (3.11) is verified but not (3.5). Here is a simple example: Take

$$H(t,X_1) = H(t,Y_1,Z_1) = Y_1 \inf(t,Z_1),$$

where $Z_1 = 1/Y_1$. This corresponds to a total input per shot $H(\infty,X_1)=1$ delivered at rate Y_1. If $E[1/Y_1]=\infty$ (very slow delivery), the left hand side of (3.11) is

$$(e^\theta - 1 - \theta)E[1/Y_1] = \infty,$$

whereas the limit to be taken in (3.5) is that of

$$1_{t>Y_1^{-1}}\frac{1}{t}\frac{1}{\theta Y_1}(e^\theta-1-\theta) + 1_{t\le Y_1^{-1}}(e^{\theta Y_1 t}-1)\left(\frac{1}{Y_1 t}\frac{1+\theta}{\theta}-1\right).$$

The second term is easily bounded by a deterministic constant times $1_{t\le Y_1^{-1}}$ and therefore its expectation tends to zero since Y_1^{-1} is a.-s. finite. As for the first term we write

$$E\left[1_{t>Y_1^{-1}}\frac{1}{t}\frac{1}{\theta Y_1}\right] = \frac{1}{\theta t}\int_t^\infty z f(z)dz,$$

where f is the probability density of $1/Y_1$. If for large z this density behaves like $\frac{1}{z^{1+c}}$, where $c \in (0,1)$, the last displayed quantity converges to zero as t tends to ∞.

Suppose we can caracterize K independent inputs to the link as Poisson shot noises with rates λ_i and total input per shot $\sigma^i = H(\infty, X^i)$, $1 \leq i \leq K$. If we impose on the link a large buffer size B and a given probability of overflow corresponding to a given decay rate R, then the effective bandwidth ([Ke91]) of flow i is

$$\beta_i = \frac{\lambda_i}{R} E[e^{R\sigma^i} - 1]. \tag{3.12}$$

Indeed the aggregated flow requires, in order to guarantee the above bound of the overflow probability, a capacity c equal to

$$\frac{\lambda}{R} E[e^{R\sigma} - 1] \; = \; \sum_i \frac{\lambda_i}{R} E[e^{R\sigma^i} - 1]. \tag{3.13}$$

A difficulty with traffic models of the shot noise type is the choice of a realistic "shot shape" $h(t, X_1)$ and of the "shot rate" λ. Efficient estimation procedures, say for a parametric model of the shot shape, are not generally available. However, as far as effective bandwidth is concerned, the modelling problem only concerns the rate and the total "one-shot" input $\sigma_1 = H(\infty, X_1)$. The modeling and statistical problems however do not completely disappear.

We now take a global point of view, by considering sufficiently aggregated traffic, in a given link, say from a local network to the entry point of a broadband network. We then identify T_n as the start of a session and σ_n as the total number of packets transferred in the session. It is reasonable to consider that the Poisson model is adequate, since aggregation of small streams leads to a Poisson limit, as follows from the various versions of the law of rare events (see [DaV-J88]). The statistics concerning the session rates and the session sizes require standard statistical software tools, and the modeling problem is by-passed. Of course such statistical analysis is best performed off-line and is potentially useful for network architecture design, in view of predicting resource needs.

4 Conclusion

The result of this note depends in a crucial way on the light tail hypothesis. This diminishes the impact of it, since it has been observed that the traffic often exhibits heavy tail or long range dependence phenomena. However, since heavy tailed traffic is in a certain sense, catastrophic, and in any case detrimental to the other type of traffic, it could be a good idea

to segregate heavy tail traffic from the rest, via a simple threshold rule on the size of the file (which could, for instance discriminate between emails with or without attachment), and to allocate to each of them its own bandwidth.

Another direction of research concerns heavy-tail Poisson shot noise traffic, using the corresponding theory in [DuO'C95] and the same formulas for the moment generating functions of shot noise. A recent article in this direction is [PaMa98], where the discrete-time $M/G/\infty$ process, a special case of discrete-time shot noise is considered. In the latter article, Theorem 3 and Corollary 4 coincide (modulo the fact that time is discrete, which is not a fundamental issue) with our results for the special case considered. However, the main import of our work consists in the observation that, for the general shot noise and in the light-tail case, the result depends only on the statistics of the total number of packets in a shot, and not on the rate of delivery of a shot.

Finally, let us observe that the result of this note can be adapted to the insurance context [Bre99].

References

[Bre99] Brémaud, P. (1999) An insensitivity property of Lundberg's estimate for delayed claims, to appear, J. Applied Probability

[DaV-J88] Daley, D., and Vere-Jones (1988) An Introduction to Point Processes, Springer, NY

[DuO'C95] Duffield, N.G., and N. O'Connell (1995) Large Deviations and Overflow Probabilities for the general Single-Server Queue, with Applications, Math. Proc. Camb. Phil. Soc., 118, 363-374

[GlWh94] Glynn, P.W., and W. Whitt (1994) Logarithmic Asymptotics for Steady State Tail Probabilities in a Single Server Queue, J. Appl. Proba., 31 A, 131-159

[Ke91] Kelly, F. (1991) Effective Bandwith at Multiclass Queues, Queuing Systems, 9, 5-16

[KWC93] Kesidis, G.; J. Walrand, and C.S. Chang (1993) Effective Bandwidths for Multiclass Markov fluids and other ATM sources, IEEE/ACH Trans. Networking, 1, 424-428

[KlMi95] Kluppelberg, C., and T. Mikosch (1995) Explosive Poisson Shot Noise Processes with Applications to Risk Reserves, Bernoulli, 1, 1/2, 125-147

[KuRo94] Kulkarni, V., and T. Rolski (1994) Fluid Models Driven by an Ornstein-Uhlenbeck process,

[No94] Norros, I. (1994) A storage model with self-similar input, Queuing Systems, 16, 387-396

[PaMa98] Parulekar, M., and A. Makowski (1998) Tail Probabilities for $M/G/\infty$ input processes (I): Preliminary Asymptotics, Queuing Systems, 27, 271-296