DIAS-APG-94-08

Large deviations for arrivals, departures, and overflow in some queues of interacting traffic.

N.G. Duffield [1] and Neil O'Connell [2]

Introduction. The theory of Large Deviations provides a mechanism with which to characterize the statistical properties of traffic arriving at a buffer. Indeed, the *thermodynamic entropy* of a traffic stream (the Large Deviation rate function) is a robust physical quantity which can be determined empirically. (This is described in the talk presented by John Lewis [6]). The entropy can be used to estimate the probabilities of rare events (such as buffer overflow) which determine the Quality of Service experienced by traffic. The theoretical relationship between the entropy of a source and the overflow probabilities in a single server queue is by now well understood in a great deal of generality, as we shall review shortly. The purpose of this paper is to apply such a "calculus of entropy" to treat classes of queueing problems involving traffic of different priorities. The main new ingredient here is that traffic with a lower service priority experiences a varying service rate: it takes the service unused by traffic of a higher service. The entropic techniques developed here enable us to characterize the *unused bandwidth* available to it. As an application we show how such low priority traffic can achieve very low loss ratios if buffered separately from a stream of comparatively high intensity. Such an arrangement could be used to transmit low volume traffic which is extremely sensitive to loss. We also derive the entropy of the output of such a stream from the buffer.

Consider a single server queue. Denote by $A_{s,t}$ the work arriving in the interval $[s,t)$, and by $S_{s,t}$ the amount of work which can be processed in the same interval. (Set $A_{t,t} = S_{t,t} = 0$). $S_{s,t} = c(t-s)$ for deterministic service at rate c. Define the workload for the interval $W_{s,t} = A_{s,t} - S_{s,t}$. Then the queue length at time t is

$$Q_t := \sup_{t' \leq t} W_{t',t} \tag{1}$$

[1]School of Mathematical Sciences, Dublin City University, Dublin 9, Ireland; Dublin Institute for Advanced Studies, 10 Burlington Road, Dublin 4, Ireland. E-mail: duffieldn@dcu.ie. Author presenting paper at 11th UK Teletraffic Symposium, Cambridge, 23-25 March 1994.

[2]Dublin Institute for Advanced Studies, 10 Burlington Road, Dublin 4, Ireland. E-mail: oconnell@stp.dias.ie

a relation which can be iterated to give

$$Q_t = \max\left\{ \sup_{t':t'\le t''<t} W_{t'',t} \ , \quad Q_{t'} + W_{t',t} \right\} \tag{2}$$

for any $t' < t$.

A and S are random variables on some underlying probability space, and we shall assume that the increments of A and S are stationary in the sense that the distributions of $(A_{s,t})_{s\le t\in\mathbb{R}}$ is identical to that of $(A_{s+t',t+t'})_{s\le t\in\mathbb{R}}$ for any $t' \in \mathbb{R}$, and similarly for S. For simplicity we shall take the time variable to be integer-valued in what follows.

The Theory of Large Deviations deals with the probabilities of rare events. As detailed below, if the excursions of the workload have probabilities which are exponentially small

$$P[W_{-t,0} > xt] \approx e^{-tI(x)} \qquad \text{for some function } I \tag{3}$$

then so do the excursions of the queue length

$$P[Q_0 > b] \approx e^{-\delta b}. \tag{4}$$

One can view I as the *thermodynamic entropy function* for the workload, and δ can be found in terms of it as follows.

For any such process X defined through the stationary increments $(X_{s,t})_{s<t}$ define the cumulant generating function (or simply the cumulant) by

$$\lambda_X(\theta) := \lim_{t\to\infty} t^{-1} \log E[e^{\theta X_{0,t}}] \tag{5}$$

for $\theta \in \mathbb{R}$ such that this limit exists. Clearly, if A and S are independent

$$\lambda_W = \lambda_A + \lambda_{-S} \tag{6}$$

when both terms on the right hand side exist. Entropies are related to cumulants through the Legendre-Fenchel transform: the transform of a function λ is denoted by λ^* and is defined by

$$\lambda^*(x) = \sup_\theta \left(\theta x - \lambda(\theta)\right). \tag{7}$$

The relationship is that when λ_W satisfies certain conditions (see Theorem 1 below), λ_W^* is the entropy (or *rate function*) I in (3). More precisely, the distribution of $W_{-t,0}/t$ degenerates as $t \to \infty$ to its mean $\lambda_W'(0)$ and for $x \ge \lambda_W'(0)$

$$\lim_{t\to\infty} t^{-1} \log P[W_{-t,0} > xt] = -\lambda_W^*(x). \tag{8}$$

An account of these matters (and the Theory of Large Deviations in general) can be found in the book of Dembo and Zeitouni [4].

Under very general circumstances the large deviation properties of Q (namely, the exponential decay rate of the queue length distribution) can be expressed in terms of those of W as follows. (This is the precise statement of the relationship between (3) and (4)).

Theorem 1 *Suppose*

(i) For each $\theta \in \mathbb{R}$, $\lambda_W(\theta)$ *exists as an extended real number.*

(ii) (Note that λ *is automatically convex.)* $\lambda_W(\cdot)$ *is essentially smooth, lower semi-continuous and there exists* $\theta > 0$ *for which* $\lambda_W(\theta) < 0$.

Then the distribution of Q_0 *has the following large deviation property.*

$$-\lim_{b\to\infty} b^{-1} \log P[Q_0 > b] = \delta := \inf_{d>0} d^{-1}\lambda_W^*(d) = \sup\{\theta : \ \lambda_W(\theta) \leq 0\}. \tag{9}$$

Note that since λ_W is convex, $\delta > 0$ if and only if $\lambda_W'(0) < 0$ (i.e. the mean workload is negative), in which case δ is the unique solution of $\lambda_W(\delta) = 0$. In the case of constant service rate c this can be rewritten as the *bandwidth equation* for the bandwidth $\sigma(\delta)$:

$$\sigma(\delta) := \frac{\lambda_A(\delta)}{\delta} = c. \tag{10}$$

Remarks: This result was proposed on heuristic grounds by Kesidis *et al* [8] and later made rigorous by Glynn and Whitt [7]. (See also [12] for further bibliographical details). It has recently been extended in two directions by Duffield and O'Connell [5]. First, the result can be proved for continuous time under an additional hypothesis on the local growth of W. Second, an analogous result holds with large deviation scalings more general than t^{-1}, b^{-1} in (5) and (9). For example, Fractional Brownian Motion (FBM) has been proposed as a model for the workload by Leland *et al* [9], based on observations of Ethernet traffic. Taking $W_{-t,0} = Z_t - ct$ where Z is FBM with Hurst parameter $H \in (0,1)$ (see Mandelbrot and Van Ness [10])

$$\lim_{b\to\infty} b^{-2(1-H)} \log P(Q_0 > b) = -\inf_{d>0} d^{-2(1-H)}(d+c)^2/2, \tag{11}$$

agreeing with the lower bound of Norros [11].

Define the *departures* $D_{s,t}$ in the interval $[s,t)$ by

$$D_{s,t} := Q_s + A_{s,t} - Q_t \tag{12}$$

and the *unused service* $U_{s,t}$ in $[s,t)$ by

$$U_{s,t} := S_{s,t} - D_{s,t}. \tag{13}$$

By means of (2) this can we rewritten in the form

$$U_{s,t} = -\min\left\{0\ ,\ Q_s + \inf_{t':s\leq t'<t} W_{s,t'}\right\}. \tag{14}$$

The unused service $U_{s,t}$ is the amount of work arriving in a secondary stream of arrivals $A^{(2)}$ which could be processed in the interval $[s,t)$, after arrivals A have been served. If $A^{(2)}$ is independent of A and S, then from (6) and (9) we see that the queueing problem for $A^{(2)}$ reduces to finding solutions of $\lambda_{A^{(2)}}(\theta) + \lambda_{-U}(\theta) = 0$. Thus we are motivated to examine the cumulant λ_{-U}, or what amounts to the same, the large deviation properties of $-U$.

Large Deviations for Unused Service. We begin with Large Deviation heuristics. The key here is the principle that *rare events occur in the most likely way*. This is based upon the observation that in a union of rare events, each of which has an exponentially small probability, the probability of the union is dominated by that of the most likely event. We now apply this idea. When the primary queue is stable the queue length distribution decays exponentially, whereas $W_{-t,0}/t$ degenerates onto $\lambda_W'(0) < 0$ as $t \to \infty$. Thus for large t we expect $-U_{0,t}/t \approx \inf_{t':0\le t'<t} W_{0,t'}/t$, in the sense that their large deviations have the same distribution as $t \to \infty$. Moreover, this infimum is overwhelmingly likely to occur at $t' = t$. Too see this, condition on $W_{0,t/t} = x < 0$, and observe that if it is not the infimum then the path $t' \to W_{0,t'}/t$ must fall below x at some $t' < t$. This would require the increment $W_{t',t}/t$ to be positive and bounded away from 0, and increasingly rare event as $t \to \infty$. Thus, $-U_{0,t}/t \approx W_{0,t}/t$. More discussion of such path-wise heuristics can be found in [1, 2, 3]. For the moment we substantiate them though a large deviation upper bound for $-U$.

Define

$$\hat{\lambda}_{-U}(\theta) := \begin{cases} \lambda_W(\theta) & \text{if } \theta \le \delta_* := (\lambda_W^*)'(0) \\ -\lambda_W^*(0) & \text{if } \theta > \delta_* \end{cases} \tag{15}$$

That is $\hat{\lambda}_{-U}$ is equal to λ_W to the left of the point δ_* such that where $\lambda_W'(\delta_*) = 0$ and takes the value $\lambda_W(\delta_*) = \lambda_W\left((\lambda_W')^{-1}(0)\right) = -\lambda_W^*(0)$ to the right of δ_*. From this description one sees that $\hat{\lambda}_{-U}$ has Legendre transform

$$\hat{\lambda}_{-U}^* = \begin{cases} \lambda_W^*(x) & \text{if } x \le 0 \\ \infty & \text{if } x > 0. \end{cases} \tag{16}$$

Theorem 2 *Under the hypotheses of Theorem 1, $-U_{0,t}$ satisfies a large deviation upper bound with rate function λ_W^*. In particular, for $x \ge \lambda_W'(0)$*

$$\limsup_{t\to\infty} t^{-1} \log P[-U_{0,t} \ge xt] \le -\hat{\lambda}_{-U}^*(x). \tag{17}$$

The proof of the theorem is given in the Appendix. One checks that $\hat{\lambda}_{-U}^{**} = \hat{\lambda}_{-U}$. Consequently one can use Varadhan's Theorem (see e.g. [4]) to conclude that $\hat{\lambda}_{-U}$ is an upper bound for the cumulant of $-U$. In the following we take $\hat{\lambda}_{-U}$ to be the cumulant: but this means that all the statements made below concerning queue lengths are, at worst, conservative.

We call

$$\sigma_U(\theta) := \hat{\lambda}_{-U}(\theta)/\theta \tag{18}$$

the *unused bandwidth* available to low traffic of lower priority. For consider the following arrangement. Two streams of traffic with arrival processes $A^{(1)}$ and $A^{(2)}$, with respective arrival cumulants $\lambda_{A^{(1)}}$ and $\lambda_{A^{(2)}}$, are buffered separately at a switch with constant service rate c. Arrivals $A^{(i)}$ have bandwidths

$$\sigma^{(i)}(\theta) := \frac{\lambda_{A^{(i)}}(\theta)}{\theta}. \tag{19}$$

Stream 1 has service priority: no arrival from $A^{(2)}$ are processed while the queue from $A^{(1)}$ is non-empty. Thus stream 1 is 'blind' to stream 2: provided its mean activity $\bar{a}^{(1)} := \lambda'_{A^{(1)}}(0)$ is less than c, it has a stable queue length distribution with asymptotic exponential decay constant $\delta^{(1)}$ which from (6) and (19) satisfies the bandwidth equation

$$\sigma^{(1)}(\delta^{(1)}) = c, \tag{20}$$

or, what is the same, $\lambda_W(\delta^{(1)}) = 0$.

Stream 2 takes the unused service U of $A^{(1)}$, characterized by the cumulant λ_{-U} corresponding through Theorem 2 to the workload process $W^{(1)}_{s,t} = A^{(1)}_{s,t} - (t-s)c$. The workload process for stream 2 is $W^{(2)} = A^{(2)} - U$. Provided its mean activity

$$\bar{a}^{(2)} := \lambda'_{A^{(2)}}(0) < -\lambda'_{-U}(0) = c - \bar{a}^{(1)}, \tag{21}$$

stream 2 has a stable queue distribution with asymptotic exponential decay constant $\delta^{(2)} > 0$ which from (6) and (19) satisfies the bandwidth condition

$$\sigma^{(2)}(\delta^{(2)}) = \sigma_U(\delta^{(2)}), \tag{22}$$

or equivalently $\lambda_{A^{(2)}}(\delta^{(2)}) + \lambda_{-U}(\delta^{(2)}) = 0$. This is illustrated in Figure 1.

The foregoing analysis has interesting ramifications for the design of multiplexers to service traffic with differing Quality of Service requirements. We consider the case that stream 2 has lower service priority than stream 1, but is more sensitive to loss: a higher value of the decay rate δ is required for stream 2. Simply aggregating the traffic without priority in a common buffer leads to a common decay rate $\tilde{\delta} < \delta^{(1)}$, the solution of $\lambda_{A^{(2)}}(\tilde{\delta}) = -\lambda_{W^{(1)}}(\tilde{\delta})$. (See Figure 1).

For separate buffering the decay constant of stream 2 is $\delta^{(2)} > \delta^{(1)}$. When stream 2 has low intensity then $\delta^{(2)}$ can be substantially larger than $\delta^{(1)}$. From Figure 2 we see that

$$\delta^{(2)} > \tilde{\delta} \quad \text{if and only if} \quad \lambda_{A^{(2)}}(\delta_*) < -\lambda_{-U}(\delta_*) = \lambda^*_W(0) = \lambda^*_{A^{(1)}}(c). \tag{23}$$

Otherwise, $\delta^{(2)} = \tilde{\delta}$, and there is no advantage to stream 2 (in terms of loss probabilities) in being buffered separately: it has the same asymptotic queue length distribution as it would have if buffered commonly with stream 1.

Large Deviations for Departure Processes. The large deviation properties of the departure process have been investigated for constant service rate by de Veciania *et al* [12]. We are able to extend the large deviation upper bound to the case of general service, but a heuristic justification of the accuracy of this estimate seems intricate.

Theorem 3 *Under the hypotheses of Theorem 1. with $Q_0 = 0$, then $D_{0,t}$ satisfies the following large deviation upper bound: for $x \geq \bar{a} := \lambda'_A(0)$*

$$\limsup_{t\to\infty} t^{-1} \log P[D_{0,t} \geq xt] \leq -\sup_{\theta\geq 0}(\theta x - \min[\lambda_A, \lambda_S](\theta)). \tag{24}$$

where min *denotes the pointwise minimum.*

Proof: Since from (12) (13) and (14), $D_{0,t} = \inf_{t':0\le t'\le t}(A_{0,t'} + S_{t',t})$,

$$\begin{aligned}
&\limsup_{t\to\infty} t^{-1}\log P[D_{0,t}\ge xt] \\
&\le \limsup_{t\to\infty}\inf_{\nu\in[0,1]} t^{-1}\log P[A_{0,\lceil\nu t\rceil} + S_{\lceil\nu t\rceil,t}\ge xt] && (25)\\
&\le \inf_{\theta\ge 0}\inf_{\nu\in[0,1]}\limsup_{t\to\infty} t^{-1}\log(E[e^{\theta A_{0,\lceil\nu t\rceil}}]E[e^{\theta S_{\lceil\nu t\rceil,t}}]) - \theta x && (26)\\
&\le \inf_{\theta\ge 0}\inf_{\nu\in[0,1]} \nu\lambda_{_A}(\theta) + (1-\nu)\lambda_{_S}(\theta) - \theta x && (27)\\
&= -\sup_{\theta\ge 0}(\theta x - \min[\lambda_{_A},\lambda_{_S}](\theta)) && (28)
\end{aligned}$$

(Here $\lceil y\rceil$ is the least integer $\ge y$). □

Note that $\min[\lambda_{_A},\lambda_{_S}]$ is not convex, but invoking Varadhan's Theorem we see that the convex envelope of its restriction to $[0,\infty]$ is an upper bound the departure cumulant (in $[0,\infty]$).

Appendix. Proof of Theorem 2: If $x>0$ then $P[-U_{0,t}\ge xt]=0$ and hence LHS of (17) is $+\infty$. For $x<0$,

$$\begin{aligned}
\limsup_{t\to\infty} t^{-1}\log P[-U_{0,t}\ge xt] &\le \limsup_{t\to\infty} t^{-1}\log P[Q_0 + \inf_{0<t''\le t} W_{0,t''}\ge xt] && (29)\\
&\le \limsup_{t\to\infty} t^{-1}\log \inf_{\nu\in(0,1]} P[Q_0 + W_{0,\lceil\nu t\rceil}\ge xt] && (30)\\
&= \limsup_{t\to\infty} t^{-1}\log \inf_{\nu\in(0,1]} P[\sup_{t'\ge 0} W_{-t',\lceil\nu t\rceil}\ge xt] && (31)\\
&\le \inf_{\nu\in(0,1]}\limsup_{t\to\infty} t^{-1}\log P[\sup_{t'\ge 0} W_{-t',\lceil\nu t\rceil}\ge xt] && (32)
\end{aligned}$$

For all $\nu\in(0,1]$, for all $\theta\in(0,\delta)$ (so that $\lambda_{_W}(\theta)<0$), then for any $\varepsilon\in(0,-\lambda_{_W}(\theta))$,

$$\begin{aligned}
P[\sup_{t'\ge 0} W_{-t',\lceil\nu t\rceil}\ge xt] &\le \sum_{t'\ge 0} P[W_{-t',\lceil\nu t\rceil}\ge xt] && (33)\\
&\le \sum_{t'\ge 0} e^{-\theta xt} E[e^{\theta W_{-t',\lceil\nu t\rceil}}] && (34)\\
&\le \sum_{t'\ge 0} e^{-\theta xt + (t'+\nu t)(\lambda_{_W}(\theta)+\varepsilon)} && (35)
\end{aligned}$$

for t sufficiently large. The sum is finite since $\lambda_{_W}(\theta)+\varepsilon<0$, and so

$$\begin{aligned}
\limsup_{t\to\infty} t^{-1}\log P[-U_{0,t}\ge xt] &\le \inf_{\theta\in(0,\delta)}\inf_{\nu\in(0,1]} -x\theta + \nu\lambda_{_W}(\theta) && (36)\\
&= -\sup_{\theta\in(0,\delta)}(x\theta + \lambda_{_W}(\theta)) = -\lambda^*_{_W}(x), && (37)
\end{aligned}$$

provided $x\in[\lambda'_{_W}(0),0]$. In a similar manner it is shown that

$$\limsup_{t\to\infty} t^{-1}\log P[-U_{0,t}\le xt] \le -\sup_{\theta<0}(x\theta - \lambda_{_W}(\theta)) = -\lambda^*_{_W}(x), \qquad (38)$$

for $x < \lambda'_w(0)$. The proof is omitted. □

Acknowledgements. One of us (NO'C) was supported by grants from EOLAS and Mentec Computer Systems Ltd, under the Higher Education-Industry Cooperation Scheme.

References

[1] V. Anantharam (1988) How large delays build up in a GI/G/1 queue. *Queueing Systems*, 5:345-368

[2] S. Asmussen (1982) Conditional limit theorem relating the random walk to its associate, with applications to risk processes and the GI/G/1 queue. *Avd. Appl. Prob.*, 14:143-170

[3] A. Dembo and T. Zajic (1992). Large deviations for sample paths of partial sums. Preprint.

[4] A. Dembo and O. Zeitouni (1993). *Large Deviation Techniques and Applications.* Jones and Bartlett, Boston-London.

[5] N.G. Duffield and Neil O'Connell (1993). Large deviations and overflow probabilities for the general single-server queue, with applications. Preprint DIAS-APG-93-30

[6] N.G. Duffield, J.T. Lewis, N. O'Connell, R. Russell and F. Toomey (1994) The entropy of an arrival process: a tool for estimating QoS parameters of ATM traffic. *Proceedings of 11th IEE Teletraffic Symposium* Cambridge, March 1994.

[7] Peter W. Glynn and Ward Whitt (1993). Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *J. Appl. Prob.*, to appear.

[8] G. Kesidis, J. Walrand and C.S. Chang (1993). Effective bandwidths for multiclass Markov fluids and other ATM Sources. Preprint.

[9] Will E. Leland, Murad S. Taqqu, Walter Willinger and Daniel V. Wilson (1993). Ethernet traffic is self-similar: stochastic modeling of packet traffic data. Preprint.

[10] Benoit B. Mandelbrot and John W. Van Ness (1968). Fractional Brownian motions, fractional noises and applications. *SIAM Review*, 10:422–437.

[11] Ilkka Norros (1994). A storage model with self-similar input. *Queueing Systems* , to appear.

[12] G. de Veciana, C. Courcoubetis and J. Walrand (1993) Decoupling bandwidths for networks: a decomposition approach to resource management. Preprint.