Large deviations, the shape of the loss curve, and economies of scale in large multiplexers

D.D. Botvich
School of Electronic Engineering,
Dublin City University, Dublin 9, Ireland;

N.G. Duffield
School of Mathematical Sciences,
Dublin City University, Dublin 9, Ireland;
and
Dublin Institute for Advanced Studies,
10 Burlington Road, Dublin 4, Ireland.
E-mail: `duffieldn@dcu.ie`

Revision, March 13, 1995

Abstract

We analyse the queue Q^L at a multiplexer with L inputs. We obtain a large deviation result, namely that under very general conditions

$$\lim_{L\to\infty} L^{-1}\log \mathbf{P}[Q^L > Lb] \quad = \quad -I(b)$$

provided the offered load is held constant, where the shape function I is expressed in terms of the cumulant generating functions of the input traffic. This provides an improvement on the usual effective bandwidth approximation $\mathbf{P}[Q^L > b] \approx e^{-\delta b}$, replacing it with $\mathbf{P}[Q^L > b] \approx e^{-LI(b/L)}$. The difference $I(b) - \delta b$ determines the economies of scale which are to be obtained in large multiplexers. If the limit $\nu = -\lim_{t\to\infty} t\lambda_t(\delta)$ exists (here λ_t is the finite time cumulant of the workload process) then $\lim_{b\to\infty}(I(b) - \delta b) = \nu$. We apply this idea to a number of examples of arrivals processes: heterogeneous superpositions, Gaussian processes, Markovian additive processes and Poisson processes. We obtain expressions for ν in these cases. ν is zero for independent arrivals, but positive for arrivals with positive correlations. Thus economies of scale are obtainable for highly bursty traffic expected in ATM multiplexing.

Keywords: Large deviations, scaling limits, ATM multiplexers, heterogeneous superpositions.

1 Introduction.

The problem of determining loss probabilities in queueing systems is crucial in the development of emergent technology of telecommunications networks using the Asynchronous Transfer Mode (ATM). Much recent work has focused on the analysis of the single server queue with general arrivals. This enables one to analyse queues with correlated arrivals, such as those which occur in the buffer of an ATM multiplexer whose input is a superposition of highly bursty sources.

Consider a general single server queue. For $t \in T$ (here $T = \mathbf{R}_+$ or $\mathbf{Z}_+$) denote by A_t the amount of work which arrives to be processed in the interval $[-t, 0)$ and by S_t the amount which can be processed in the same interval. If more work arrived than can be processed, the surplus waits in the queue. The workload process W is defined by $W_0 = 0$ and

$$W_t = A_t - S_t, \tag{1.1}$$

and the queue of unprocessed work at time zero is

$$Q = \sup_{t \geq 0} W_t. \tag{1.2}$$

The relation between the tail of the queue length distribution and the large deviation properties of the workload processes has been established in progressive degrees of generality. Following a heuristic proposal by Kesidis *et al* [22] (see also [32, 5] for further bibliographical details), Glynn and Whitt [16] showed for $T = \mathbf{Z}_+$ that if the pair $(W_t/t, t)$ satisfy a *large deviation principle* then

$$\lim_{b \to \infty} b^{-1} \log \mathbf{P}[Q > b] = -\delta, \tag{1.3}$$

where

$$\delta = \sup\{\theta \mid \lambda(\theta) \leq 0\}, \tag{1.4}$$

and λ is the cumulant generating function of the workload process defined by

$$\lambda(\theta) = \lim_{t \to \infty} t^{-1} \log \mathbf{E}[e^{\theta W_t}]. \tag{1.5}$$

Alternatively, δ can be expressed through

$$\delta = \inf_{t > 0} t \lambda^*(t^{-1}) \tag{1.6}$$

where λ^*, the Legendre-Fenchel transform of λ, is defined through

$$\lambda^*(x) := \sup_{\theta} \left(x\theta - \lambda(\theta) \right). \tag{1.7}$$

(We refer the reader to the book Dembo and Zeitouni [9] as a comprehensive reference for large deviations, that of Bucklew [3] for a more heuristic approach, and the article of Lewis and Pfister [26] for a general introduction).

Recently, Duffield and O'Connell have extended this result in two directions [12]. Firstly, the same result is shown to hold when $T = \mathbf{R}_+$, subject to a local growth condition on W. Secondly an analogous result holds with large deviation scalings more general than the linear scalings b and t in (1.3) and (1.5). These are appropriate for treating, for example, the case where the workload is fractional Brownian motion: this has been proposed as a model for the workload by Leland *et al* [25], based on observations of Ethernet traffic.

The relation (1.3) is the basis of the *effective bandwidth approximation* to the queue length distribution:

$$\mathbf{P}[Q > b] \approx e^{-\delta b}. \tag{1.8}$$

(See for example, [5, 18, 19, 15, 21, 34] for development, applications and further references). The motivation here is that for ATM multiplexers one wants to estimate exponentially small loss probabilities, which in practice are to be as small as 10^{-9}. However, there is already theoretical and numerical work indicating that (1.8) is insufficient for this purpose. For a queue where the input is an L-fold superposition of Markovian sources served at constant rate s, Duffield [10] proves the upper bound

$$\mathbf{P}[Q > b] \leq e^{-\mu L} e^{-\delta b} \tag{1.9}$$

δ is as before and μ and δ depend only the the traffic due to a single source, and on the offered load through the ratio s/L. In the example of on-off Markov sources with positive autocorrelation, μ is strictly positive (see Buffet and Duffield [4]). Thus in a large superposition, the loss probabilities may be exponentially small even for small b: the effective bandwidth approximation (1.8) can be extremely conservative through over-estimating the loss probabilities. On the other hand, if μ were negative, then (1.9) suggests that (1.8) will under-estimate the loss probabilities at large L, even for large b. Moreover, both types of error become more severe as L increases at constant load. Both these types of behaviour have been observed though numerical studies of queueing models by Choudhury *et al* [6], in which the asymptotic approximation

$$\mathbf{P}[Q > b] \approx \beta e^{-\eta L} e^{-\delta b} \tag{1.10}$$

is proposed. The effective bandwidth approximation, when compared with these results, is shown to over-estimate the loss probabilities in examples of bursty sources, and under-estimate them in examples of sub-bursty sources. Finally, we note the work of Weiss [33] for L-fold superpositions of On-Off Markov fluids. There a sample-path large deviation argument was used to identify the most likely path to a rare event (such as overflow from a large buffer) and its probability, with the asymptotics $\mathbf{P}[Q > Lb] \approx e^{-LJ(b)}$ where for large b, $J(b) \approx \eta + \delta b$.

Thus we are led to investigate the large deviation properties of the queue length distribution *in the size L*. Apart from the above considerations, we are motivated by the observation in examples that the broad features of the queue length distribution remain roughly invariant when both the size L and the queue length b are jointly scaled. (See simulation results in the thesis of Corcoran [7], heuristic arguments by Rasmussen *et al* [28] and by Courcoubetis *et al* [8]). For example, let Q^L be the queue due to a superposition of L identical sources *not necessarily Markovian*, served at constant rate sL (s fixed), and denote by λ_t the finite time cumulant due to a single source with arrival process A served at rate s:

$$\lambda_t(\theta) = t^{-1} \log \mathbf{E}[e^{\theta(A_t - st)}]. \tag{1.11}$$

The workload W_t^L of the superposition is the sum of L independent copies of $W_t = A_t - st$ and $Q_t^L = \sup_{t \geq 0} W_t^L$. As a consequence of Theorem 1 of the next section it follows (under convergence of λ_t as $t \to \infty$ and suitable local regularity conditions on the workload) that for $b \geq 0$

$$\lim_{L \to \infty} L^{-1} \log \mathbf{P}[Q^L > Lb] = -I(b) \tag{1.12}$$

where I is the *shape function* defined by

$$I(b) = \inf_{t>0} t\lambda_t^*(b/t). \tag{1.13}$$

The heuristics behind (1.12) are as follows. By (1.2), $\mathbf{P}[Q^L > Lb] = \mathbf{P}[\cup_{t\geq 0}\{W_t^L > Lb\}]$. The probability of each event in the union is exponentially small for large L, and so the probability is dominated by that of the most likely event $\sup_{t\geq 0}\mathbf{P}[W_t^L > Lb]$. If, for each fixed t, the family of superposed workloads satisfies a large deviation principle with some rate function I_t, i.e. $\mathbf{P}[W_t^L > Lb] \approx e^{-LI_t(b)}$, then altogether we get $\mathbf{P}[Q_t^L > Lb] \approx e^{-\inf_t I_t(b)}$. The required large deviation principle for independent superpositions follows from Cramér's Theorem, and $I_t = (t\lambda_t)^*$. (However, independence is not a requirement for our proof). We note that this large deviation principle was obtained for alternating renewal processes by Simonian and Guibert [30].

The shape function I can be seen to give the large scale corrections to the effective bandwidth approximation: (1.8) is replaced by

$$\mathbf{P}[Q^L > b] \approx e^{-LI(b/L)}. \tag{1.14}$$

(Observe that if one replaces λ_t^* by λ^* in (1.13), where $\lambda = \lim_{t\to\infty}\lambda_t$, then by (1.6) $I(b)$ reverts to $b\delta$). $I(b) - \delta b$ determines the error incurred by using the effective bandwidth at large L, or to make a more positive statement, it determines the *economies of scale* to be obtained in multiplexers of a large number of sources.

We examine the initial value $I(0)$ and asymptotics as $b \to \infty$ of the shape function I in Theorems 2 and 3. For $T = \mathbf{Z}_+$, a workload with stationary increments, $I(0) = \lambda_1^*(0)$. This is just the large deviation result for the loss probability in a bufferless resource as found by Hui [18, 19]. The asymptotics of I are

$$\lim_{b\to\infty}(I(b) - \delta b) = \nu \tag{1.15}$$

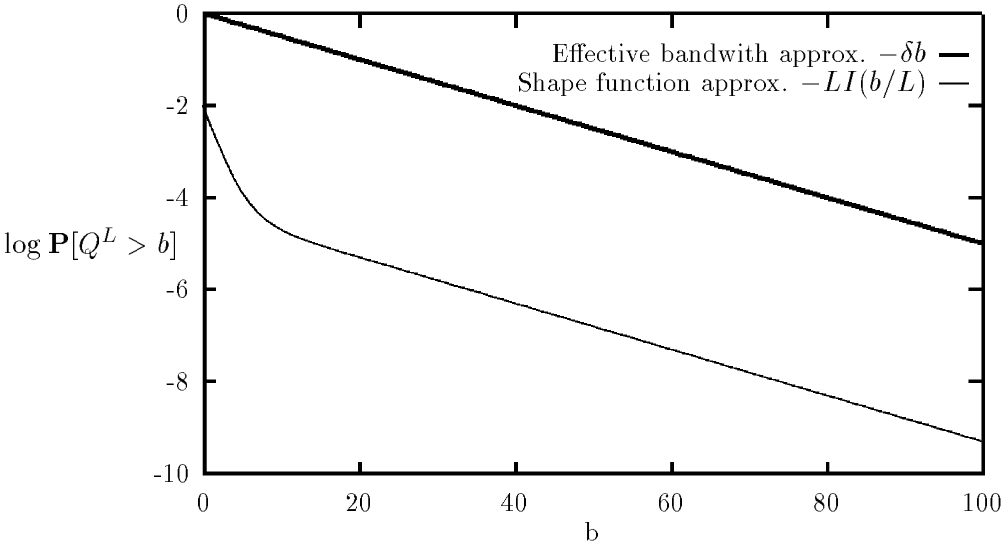
where

$$\nu = -\lim_{t\to\infty} t\lambda_t(\delta) \tag{1.16}$$

provided this limit exists, and subject to some regularity conditions in the case $T = \mathbf{Z}_+$. For large b (at least of order L) this means we can make the approximation $I(b) \approx \nu + \delta b$. This establishes that the asymptotics found by Weiss for Markov fluid sources hold far more generally, and provides a theoretical justification for (1.10).

One sees from (1.4) that $\nu = 0$ for uncorrelated arrivals, since then $\lambda_t(\delta) = \lambda(\delta) = 0$. Thus there are no economies of scale to be gained at large (rescaled) buffer sizes for uncorrelated arrivals, since then $-\delta b$ is asymptotic to $I(b)$ for large b. On the other hand, a sufficient condition for ν to be positive is that the workloads on disjoint time intervals are positively correlated (Theorem 4). This is typically the case for highly bursty sources. A generic configuration with $\nu > I(0) > 0$ is illustrated in Figure 1.

Figure 1: Economies of Scale with $\nu > I(0) > 0$



æ

It is interesting to note that ν depends on finer details of the workload process than those which determine the asymptotic slope δ: it depends not only on the limiting cumulant λ but rather on the manner in which the λ_t approach λ as $t \to \infty$. To borrow from the terminology of physics, ν is not a thermodynamic quantity.

The paper is organised in the following way. The basic large deviation result is stated and proved in section 2. The analysis of the shape function I is done in section 3. In section 4 we apply them to a number of examples. The case of heterogeneous superpositions is worked out in 4.1. Gaussian workload process are covered in 4 and the specific example of Ornstein-Uhlenbeck processes in 4.3, including a calculation of the shape function for a heterogeneous superposition. Markov Additive Processes are treated in 4.4. In this case we can express ν in terms of the the maximal eigenfunction of the (Laplace transform) of the Markov transition kernel (Corollary 5). Comparisons of the approximation (1.14) with simulation are made for superpositions of Markovian on-off sources. Finally, in 4.5 we apply the results to a very simple class of examples: independent Poissonian arrivals with general service distribution. In light of the explicit distribution for $\sum^L M(\kappa)/M(L\mu)/1$ it is not surprising that in this case $I(0) = \nu = 0$: there are no economies of scale to be obtained for Poissonian arrivals at *any* buffer size.

2 Large deviations.

We begin by stating our hypotheses concerning the workload processes, then give some examples which satisfy the hypotheses. For each $L \in \mathbf{N}$, $(W_t^L)_{t \in T}$ (where $T = \mathbf{Z}_+$ or $\mathbf{R}_+$) is a stochastic process, and $W_0^L = 0$. The queue length at time zero is

$$Q^L = \sup_{t \in T} W_t^L. \tag{2.1}$$

(Note that if the increments of W^L are stationary, then the distribution of Q^L is also stationary). For $\theta \in \mathbf{R}$ define the cumulant generating function

$$\lambda_t^L(\theta) = (Lt)^{-1} \log \mathbf{E}[e^{\theta W_t^L}]. \tag{2.2}$$

Hypothesis 1

(i) For each $\theta \in \mathbf{R}$, the limits

$$\lambda_t(\theta) = \lim_{L\to\infty} \lambda_t^L(\theta) \qquad \text{and} \qquad \lambda(\theta) = \lim_{t\to\infty} \lambda_t(\theta) \tag{2.3}$$

exist as extended real numbers. Moreover, the first limit exists uniformly for all t sufficiently large.

(ii) λ (and λ_t) are essentially smooth: i.e. λ is differentiable on the interior of its effective domain (the region where it is finite), and $\lim_{n\to\infty} |\lambda'(\theta_n)| = +\infty$ *for any sequence (θ_n) in the effective domain which converges to a point on its boundary.*

(iii) There exists $\theta > 0$ for which $\lambda_t(\theta) < 0$ for all $t \in T$.

(iv) ($T = \mathbf{R}_+$) For all $t \geq r \geq 0$ define $\tilde{W}_{t,r}^L = \sup_{0<r'<r} W_{t-r'}^L - W_t^L$. Then for all $\theta \in \mathbf{R}$

$$\limsup_{r\to 0} \limsup_{L\to\infty} L^{-1} \sup_{t\geq 0} \log \mathbf{E}[e^{\theta \tilde{W}_{t,r}^L}] \leq 0. \tag{2.4}$$

Remark: if Hypotheses 1(i),(ii) are satisfied, then by the Gärtner-Ellis theorem, for each t the pair $(W_t^L/L, L)$ satisfies a large deviation principle with good rate function given by the Legendre-Fenchel transform of $t\lambda_t$. In other words, for any Borel set Γ,

$$\limsup_{L\to\infty} L^{-1} \log \mathbf{P}(W_t^L/L \in \Gamma) \leq - \inf_{x\in\overline{\Gamma}} (t\lambda_t)^*(x), \tag{2.5}$$

and

$$\liminf_{L\to\infty} L^{-1} \log \mathbf{P}(W_t^L/L \in \Gamma) \geq - \inf_{x\in\Gamma^\circ} (t\lambda_t)^*(x), \tag{2.6}$$

where the Legendre-Fenchel transform of a function f is

$$f^*(x) := \sup_{\theta\in\mathbf{R}} \{\theta x - f(\theta)\}, \tag{2.7}$$

from which it follows that $(t\lambda_t)^*(x) = t\lambda_t^*(x/t)$. Let x_t^* be the solution of $\lambda_t^*(x_t^*) = 0$: by Hypothesis 1(iii) it is negative. Then for $x \geq x_t^*$

$$\limsup_{L\to\infty} L^{-1} \log \mathbf{P}(W_t^L/L > x) \leq -t\lambda_t^*(x/t), \tag{2.8}$$

and

$$\liminf_{L\to\infty} L^{-1} \log \mathbf{P}(W_t^L/L > x) \geq -t\lambda_t^*(x^+/t), \tag{2.9}$$

where $^+$ indicates limit from above. Hypothesis 1(iii) is a stability condition: then there exists a strictly positive solution δ of the equation $\lambda(\delta) = 0$, which is the asymptotic decay rate of the queue length distribution. Hypothesis 1(iv) is a local regularity condition on the sample paths of the workload.

Examples

Homogeneous superpositions. There are L identical sources whose backward arrival processes are independent copies of $(A_t)_{t\in T}$. The superposition is served at a constant rate sL: the offered load is independent of L. Then

$$\lambda_t(\theta) = \lambda_t^L(\theta) = t^{-1}\log \mathbf{E}[e^{\theta A_t}] - s\theta, \tag{2.10}$$

independent of L. Thus λ_t is the cumulant corresponding to the workload $W_t = A_t - st$ of a single source served at rate s.

Heterogeneous superpositions. Sources are classified by type j in some finite index set J. There are L_j sources of type j, with $L = \sum_j L_j$ sources in total. The backward arrival process for a source of type j is $(A_{j,t})_{t\in T}$. All sources are independent. Then

$$\lambda_t^L(\theta) = \sum_{j\in J} p_j^L c_{j,t}(\theta) - s\theta, \tag{2.11}$$

where

$$c_{j,t}(\theta) = t^{-1}\log \mathbf{E}[e^{\theta A_{j,t}}] \quad \text{and} \quad p_j^L = L_j/L. \tag{2.12}$$

The limits $c_j(\theta) = \lim_{t\to\infty} c_{j,t}(\theta)$ are assumed to exist with $c_j(\cdot)$ essentially smooth. Then for any $\theta \in \mathbf{R}$, $L \to \lambda_t^L(\theta)$ is convergent provided the limits $p_j = \lim_{L\to\infty} p_j^L$ exist, and the convergence is uniform in t since J is finite. Heterogeneous superpositions have been previously analysed through the effective bandwidth approximation (see references above) and through eigenfunction expansions for classes of Markovian fluid models by Kosten [23] (building on the early work of Anick *et al*[1] on homogeneous superpositions) and finite state models by Elwalid *et al* [14].

For $T = \mathbf{R}_+$ and $A_{j,t}$ having stationary increments, Hypothesis 1(iv) is satisfied if, for each $j \in J$,

$$\limsup_{r\to 0} \mathbf{E}[\exp(\theta \sup_{0<t<r} |A_{j,t}|)] = 1. \tag{2.13}$$

Time rescalings. The single source arrival process is $A_{L,t}$ where the process $(A_{L,Lt})$ is convergent in distribution to some process (A_t) as $L \to \infty$. The superposition is served at a *fixed* rate s. Thus with $\sum^L$ denoting an L-fold superposition:

$$\begin{aligned} Q_t^L &= \sup_{t\geq 0}\left((\textstyle\sum^L A_{L,t}) - st\right) &(2.14)\\ &= \sup_{t\geq 0}\left((\textstyle\sum^L A_{L,Lt}) - Lst\right). &(2.15)\end{aligned}$$

We shall take the limit $L \to \infty$ and so assume $\lambda_t^L(\theta) = (Lt)^{-1}\log \mathbf{E}[\exp\theta \sum^L (A_{L,Lt} - st)]$ satisfies Hypothesis 1. This class of models is motivated by examples of rescaled Markovian sources simulated by Corcoran [7], and of rescaled renewal processes examined by Sriram and Whitt [31] and Rasmussen *et al* [28].

Theorem 1 *Under Hypothesis 1, for each $b > 0$,*
Upper bound:

$$\begin{aligned} \limsup_{L\to\infty} L^{-1}\log \mathbf{P}[\sup_{t>0} W_t^L > Lb] &\leq -I(b) &(2.16)\\ \text{where } I(b) &:= \inf_{t>0} t\lambda_t^*(b/t) &(2.17)\end{aligned}$$

Lower bound:

$$\liminf_{L\to\infty} L^{-1}\log \mathbf{P}[\sup_{t>0} W_t^L > Lb] \geq -I(b^+). \tag{2.18}$$

Proof of Theorem 1 : *Lower Bound*

$$\begin{aligned}
\liminf_{L\to\infty} L^{-1}\log \mathbf{P}[\sup_{t>0} W_t^L > Lb] &\geq \liminf_{L\to\infty} L^{-1}\sup_{t>0}\log \mathbf{P}[W_t^L > Lb] && (2.19)\\
&\geq \sup_{t>0}\liminf_{L\to\infty} L^{-1}\log \mathbf{P}[W_t^L > Lb] && (2.20)\\
&= \sup_{t>0} -t\lambda_t^*(b^+/t) = -I(b^+), && (2.21)
\end{aligned}$$

by (2.9), where the last equality follows since, by Hypothesis 1(iii), λ_t^* is increasing on $[0,\infty) \subset [x_t^*,\infty)$.

Upper Bound: $T = \mathbf{Z}_+$. For any $t,\theta > 0$ and $\theta_{t'} > 0, 0 < t' < t$,

$$\begin{aligned}
\mathbf{P}[\sup_{t'>0} W_{t'}^L > Lb] &\leq t\max_{0<t'<t}\mathbf{P}[W_{t'}^L > Lb] + \sum_{t'\geq t}\mathbf{P}[W_{t'}^L > Lb] && (2.22)\\
&\leq t\max_{0<t'<t} e^{-Lb\theta_{t'}+t'L\lambda_{t'}^L(\theta_{t'})} + e^{-\theta Lb}\sum_{t''\geq t} e^{Lt''\lambda_{t''}^L(\theta)} && (2.23)
\end{aligned}$$

by Chebychev's inequality. Since $\lambda_t^L(\theta) \to \lambda_t(\theta)$ uniformly in t, $\lambda_t(\theta) \to \lambda(\theta)$ and $\lambda(\theta) < 0$ on $(0,\delta)$, we can find $\theta > 0$ and $\varepsilon < 0$ such that $\lambda_t^L(\theta) < \varepsilon$ for all L, t sufficiently large. This means that for such L and t the geometric series in (2.23) is summable, yielding

$$\mathbf{P}[\sup_{t'>0} W_{t'}^L > Lb] \leq t\max_{0<t'<t} e^{-Lb\theta_{t'}+t'L\lambda_{t'}^L(\theta_{t'})} + e^{-\theta Lb+Lt\varepsilon}/(1-e^{L\varepsilon}) \tag{2.24}$$

Taking logarithms, dividing by L, taking the lim sup as $L \to \infty$ and finally taking the infimum over the $\theta_{t'}$ we obtain

$$\limsup_{L\to\infty} L^{-1}\log \mathbf{P}[\sup_{t'>0} W_{t'}^L > Lb] \leq \max\left(\max_{0<t'<t}\left(-t'\lambda_{t'}^*(b/t')\right),\ -\theta b + t\varepsilon\right) \tag{2.25}$$

Recall $\varepsilon < 0$ so that taking the limit $t \to \infty$ we obtain (in conjunction with the lower bound) the stated result.

Upper Bound: $T = \mathbf{R}_+$. For any $\epsilon > 0$ and $n \in \mathbf{N}$ define

$$\hat{W}_n^L = \sup_{(n-1)\epsilon<t\leq n\epsilon} W_t^L \quad \text{and} \quad \hat{\lambda}_n^L = (nL)^{-1}\log \mathbf{E}[e^{\theta \hat{W}_n^L}]. \tag{2.26}$$

By Hölder's inequality then for any p in $(0,1)$:

$$n\hat{\lambda}_n^L(\theta) \leq n\epsilon\, p\lambda_{n\epsilon}^L(\theta/p) + (1-p)L^{-1}\log \mathbf{E}[e^{\theta\tilde{W}_{n\epsilon,\epsilon}^L/(1-p)}], \tag{2.27}$$

with $\tilde{W}^L$ as in Hypothesis 1(iv). According to Hypothesis 1, for any $p \in (0,1)$ we can make the second term of the right hand side of (2.27) as small as we like by choosing ϵ sufficiently small then L sufficiently large. Thus we can repeat the steps (2.22), (2.23) and (2.24) with ϵ and p fixed, take the limits $t \to \infty$ then $\epsilon \to 0$ to obtain

$$\begin{aligned}
\limsup_{L\to\infty} L^{-1}\log \mathbf{P}[\sup_{t>0} W_t^L > Lb] &\leq \limsup_{L\to\infty} L^{-1}\log \mathbf{P}[\sup_{n>0}\hat{W}_n^L > Lb] && (2.28)\\
&\leq p\sup_{t>0} -t\lambda_t^*(b/t), && (2.29)
\end{aligned}$$

since $(p\lambda_t(\cdot/p))^* = p\lambda_t^*(\cdot)$, and finally let $p \nearrow 1$ to get the stated result. ■

3 Analysis of the shape function.

In this section we analyse the initial value $I(0)$ of the shape function, and the asymptotics of $I(b)$ as $b \to \infty$.

For $t > 0, \theta \in \mathbf{R}$ define $\Lambda_t(\theta) = t\lambda_t(\theta/t)$.

Hypothesis 2

(i) W_t^L *has stationary increments.*

(ii) $(T = \mathbf{R}_+)$ *The limit* $\Lambda(\theta) = \lim_{t\to 0} \Lambda_t(\theta)$ *exists as an extended real number for all* $\theta \in \mathbf{R}$.

(iii) 0 *lies in the interior of the effective domain of* Λ^*.

Theorem 2 *Under Hypotheses 1 and 2*

$$(T = \mathbf{Z}_+) \qquad I(0) = \lambda_1^*(0).$$

$$(T = \mathbf{R}_+) \qquad I(0) = \Lambda^*(0).$$

Remark: Theorem 2 says that a stable workload with stationary increments is (asymptotically) most likely to exceed 0 at the smallest times. But this need not be the case for non-stationary workloads. For the proof of Theorem 2 we need the following Lemma:

Lemma 1 *Under Hypothesis 2(i), for* $n \in \mathbf{N}$, $r \in T$

$$\Lambda_r(\theta) \geq \Lambda_{nr}(\theta). \tag{3.1}$$

Proof: This follows from Hölder's inequality and the stationarity of the increments of W. For $(\mu_i)_{1\leq i\leq n}$, with $\mu_i > 0$ and $\sum_{i=1}^n \mu_i = 1$

$$\mathbf{E}[e^{\theta W_t/t}] \leq \prod_{i=1}^n \mathbf{E}[e^{\theta W_{\mu_i t}/(\mu_i t)}]^{\mu_i} \tag{3.2}$$

and hence

$$\Lambda_t(\theta) \leq \sum_{i=1}^n \mu_i \Lambda_{\mu_i t}(\theta), \tag{3.3}$$

from which (3.1) follows by taking $\mu_i = n^{-1}$ and $t = nr$. ■

Proof of Theorem 2:

$$t\lambda_t^*(0) = \sup_\theta -t\lambda_t(\theta) = \sup_\theta -t\lambda_t(\theta/t) = \Lambda_t^*(0), \tag{3.4}$$

so that $I(0) = \inf_t \Lambda_t^*(0)$.

For $T = \mathbf{Z}_+$ observe from Lemma 1 that $\Lambda_1(\theta) \geq \Lambda_t(\theta)$ and hence $\Lambda_1^*(0) \leq \Lambda_t^*(0)$. Hence $I(0) = \Lambda_1^*(0) = \lambda_1^*(0)$.

For $T = \mathbf{R}_+$ since, by Hypothesis 2(ii) Λ_t converges to Λ, then by Lemma 1 of [11], Λ_t^* converges to Λ^* on the interior of the effective domain of Λ^*. By Lemma 1, $n \mapsto \Lambda_{r/2^n}$ is increasing for any $r > 0$, and so $n \mapsto \Lambda^*_{r/2^n}$ is decreasing. Hence $\Lambda^*(b) = \inf_{t>0} \Lambda_t^*(b)$ for any b in the interior of the effective domain of Λ: for any $\varepsilon > 0$ we can choose r such that

$$\inf_t \Lambda_t^*(b) + \varepsilon \geq \Lambda_r^*(b) \geq \Lambda^*_{r/2^n}(b) \geq \inf_t \Lambda_t^*(b) \tag{3.5}$$

and $\lim_{n\to\infty} \Lambda^*_{r/2^n}(b) = \Lambda^*(b)$. The result now follows by Hypothesis 2(iii). ■

The identification of the asymptotics of I requires some technical conditions, as follows.

Hypothesis 3

(i) ($T = \mathbf{Z}_+$ or $T = \mathbf{R}_+$) The following limit exists:

$$\nu := -\lim_{t\to\infty} t\lambda_t(\delta). \tag{3.6}$$

(ii) ($T = \mathbf{Z}_+$) λ_t and λ are strictly convex and $t \mapsto (t+1)\lambda'_{t+1}(\delta) - t\lambda'_t(\delta)$ is bounded above; or

(ii′) ($T = \mathbf{Z}_+$) λ_t and λ are strictly convex; $(\lambda_t^*)'$ and $(\lambda^*)'$ are uniformly Lipschitz continuous on some neighbourhood of $\lambda'(\delta)$; and

$$(t+1)\lambda'_{t+1}(\delta) - t\lambda'_t(\delta) = \mathbf{o}(\sqrt{t}). \tag{3.7}$$

Remark: Hypothesis 3(i) can be understood as follows. Let $\lambda_t(\delta_t) = 0$. Then $\lambda_t(\delta) \approx \lambda'_t(\delta)(\delta - \delta_t)$. So the existence of a finite limit ν means that $\delta_t - \delta \sim t^{-1}$ for large t.

Theorem 3 *Under Hypothesis 3(i), and with the addition of Hypotheses 3(ii) or 3(ii′) for $T = \mathbf{Z}_+$, then*

$$\lim_{b\to\infty} (I(b) - \delta b) = \nu. \tag{3.8}$$

According to Hypothesis 1(ii), λ_t and λ are differentiable, so the convergence of λ_t to λ implies the pointwise convergence of λ'_t to λ'. (See, for example, Lemma IV.6.3 of [13]).

Proof of Theorem 3: Define

$$\beta(t) := t\lambda'_t(\delta) \tag{3.9}$$

Since $\lambda'_t(\delta) \to \lambda'(\delta) > 0$ as $t \to \infty$, $t \mapsto \beta(t)$ is increasing for t sufficiently large and $\lim_{t\to\infty} \beta(t) = +\infty$. Set

$$\tau(b) := \sup\{t \in T \mid \beta(t) \leq b\}. \tag{3.10}$$

$\mathrm{Ran}(\beta) \ni b \mapsto \tau(b)$ is increasing and $\lim_{b\to\infty} \tau(b) = +\infty$. By definition of the Legendre-Fenchel transform of λ_t^* and (3.9)

$$t\lambda_t^*(b/t) - \delta b \geq t\lambda_t^*(\beta(t)/t) - \delta\beta(t) = -t\lambda_t(\delta). \tag{3.11}$$

We obtain upper bounds for $\limsup_{b\to\infty}(I(b) - \delta b)$, first for $T = \mathbf{R}_+$, then for $T = \mathbf{Z}_+$. We then show these are equal to a lower bound for $\liminf_{b\to\infty}(I(b) - \delta b)$.

Upper Bound: ($T = \mathbf{R}_+$) $\mathrm{Ran}(\tau) = \mathbf{R}_+$, and so for any $b \in \mathbf{R}_+$,

$$\begin{aligned} I(b) - \delta b &= \inf_t t\lambda_t^*(b/t) - \delta b && (3.12)\\ &\le \tau(b)\lambda_{\tau(b)}^*(b/\tau(b)) - \delta b && (3.13)\\ &= -\tau(b)\lambda_{\tau(b)}(\delta), && (3.14)\end{aligned}$$

the last equality following from (3.11) because $\beta(\tau(b)) = b$ for $T = \mathbf{R}_+$. Since $\tau(b) \to \infty$ as $b \to \infty$, then by Hypothesis 3,

$$\limsup_{b\to\infty}(I(b) - \delta b) \le \nu. \qquad (3.15)$$

($T = \mathbf{Z}_+$) In this case $\mathrm{Ran}(\beta)$ is a discrete set, but the conclusion (3.15) holds provided we take the limit along $\mathrm{Ran}(\beta)$, since $\beta \circ \tau$ acts as the identity there. But for any $b \in \mathbf{R}_+$ we have

$$\tau(b)\lambda_{\tau(b)}^*(b/\tau(b)) - \delta b = \tau(b)\lambda_{\tau(b)}^*(\beta(\tau(b))/\tau(b)) - \delta\beta(\tau(b)) + E_b \qquad (3.16)$$

where

$$\begin{aligned} E_b &= \tau(b)\left(\lambda_{\tau(b)}^*(b/\tau(b)) - \lambda_{\tau(b)}^*(\beta(\tau(b))/\tau(b))\right) - \delta\,(b - \beta(\tau(b))) && (3.17)\\ &\le (b - \beta(\tau(b)))\,(\delta_b - \delta) && (3.18)\end{aligned}$$

where

$$\delta_b = (\lambda_{\tau(b)}^*)'(b/\tau(b)). \qquad (3.19)$$

Here we have used the fact that since λ_t is strictly convex, λ_t^* is differentiable (see Theorem 26.3 in [29]).

The proof is complete if we can show that $\limsup_{b\to\infty} E_b \le 0$, since then $\limsup_{b\to\infty}(I(b) - \delta b) \le \limsup_{b\to\infty} -\tau(b)\lambda_{\tau(b)}(\delta) = \nu$. Note

$$\beta(\tau(b)) \le b \le \beta(\tau(b) + 1) \qquad (3.20)$$

(for sufficiently large b) and the relations

$$\tau(b)\lambda'_{\tau(b)}(\delta_b) = b;\ \tau(b)\lambda'_{\tau(b)}(\delta) = \beta(\tau(b));\ (\tau(b)+1)\lambda'_{\tau(b)+1}(\delta) = \beta(\tau(b)+1) \qquad (3.21)$$

from which it follows that $\delta_b \ge \delta$ since λ'_t is increasing. (This means E_b is non-negative). Combining these gives

$$\frac{\beta(\tau(b)+1) - \beta(\tau(b))}{\tau(b)} \ge \frac{b - \beta(\tau(b))}{\tau(b)} = \lambda'_{\tau(b)}(\delta_b) - \lambda'_{\tau(b)}(\delta) \ge 0. \qquad (3.22)$$

We now proceed under Hypothesis 3(ii) Since, from (3.21),

$$\lim_{b\to\infty} \beta(\tau(b))/\tau(\beta) = \lim_{b\to\infty} \lambda'_{\tau(b)}(\delta) = \lambda'(\delta), \qquad (3.23)$$

then by (3.22)

$$\lim_{b\to\infty}\left(\lambda'_{\tau(b)}(\delta_b) - \lambda'_{\tau(b)}(\delta)\right) = 0. \qquad (3.24)$$

Hence

$$\lim_{b\to\infty} \delta_b = \delta; \qquad (3.25)$$

for if not then $\delta_{b_i} > \delta+\varepsilon$ for some $\varepsilon > 0$ and subsequence $b_i \to \infty$. so $\limsup_{i\to\infty} \lambda'_{\tau(b_i)}(\delta_{b_i}) - \lambda'_{\tau(b_i)}(\delta) \geq \limsup_{i\to\infty} \lambda'_{\tau(b_i)}(\delta+\varepsilon) - \lambda'_{\tau(b_i)}(\delta) = \lambda'(\delta+\varepsilon) - \lambda'(\delta) > 0$, since λ is strictly convex, in contradiction with (3.24). Finally,

$$0 \leq b - \beta(\tau(b)) \leq \beta(\tau(b)+1) - \beta(\tau(b)) = (\tau(b)+1)\lambda'_{\tau(b)+1}(\delta) - \tau(b)\lambda'_{\tau(b)}(\delta) \tag{3.26}$$

which is bounded according to Hypothesis 3(ii). Combining with (3.25), then $\lim_{b\to\infty} E_b = 0$ as required.

Alternatively, under Hypothesis 3(ii′), $0 \leq \delta_b - \delta \leq k(b - \beta(\tau(b))$ for some $k > 0$ independent of b and so

$$\begin{aligned} 0 \leq E_b &\leq k\,(b - \beta(\tau(b)))^2/\tau(b) && (3.27)\\ &\leq (\beta(\tau(b)+1) - \beta(\tau(b)))^2/\tau(b) && (3.28)\\ &= \left((\tau(b)+1)\lambda'_{\tau(b)+1}(\delta) - \tau(b)\lambda'_{\tau(b)}(\delta)\right)^2/\tau(b) && (3.29)\end{aligned}$$

which goes to 0 as $b \to \infty$ by (3.7).

Lower Bound: ($T = \mathbf{R}_+$ or $\mathbf{Z}_+$) Suppose first that $\inf_t t\lambda^*_t(b/t)$ is attained at $\hat{\tau}(b)$.

$$\begin{aligned} \inf_t t\lambda^*_t(b/t) - \delta b &= \hat{\tau}(b)\lambda^*_{\hat{\tau}(b)}(b/\hat{\tau}(b)) - \delta b && (3.30)\\ &\geq -\hat{\tau}(b)\lambda_{\hat{\tau}(b)}(\delta) && (3.31)\end{aligned}$$

by (3.11) and so

$$\liminf_{b\to\infty}(I(b) - \delta b) \geq \nu \tag{3.32}$$

provided $\hat{\tau}(b) \to \infty$ as $t \to \infty$. But if this is not the case, $\hat{\tau}(b)$ is bounded. Thus we obtain a contradiction with the upper bound (3.15) if we can show that

$$\lim_{b\to\infty}(t\lambda^*_t(b/t) - \delta b) = +\infty \tag{3.33}$$

for any fixed t. But this is true since $b \mapsto t\lambda^*_t(b/t) - \delta b$ is strictly convex and by (3.9) achieves it infimum at $\beta(t) < \infty$.

If $\inf_t t\lambda_t(b/t)$ is not attained, then we can repeat the above arguments replacing $\hat{\tau}(b)$ with $\hat{\tau}_\varepsilon(b)$ for which the infimum is approximated to within ε, uniformly in b, then take $\varepsilon \to 0$ at the end. ■

We shall say more concerning the existence of ν in the context of Markov Additive Processes in section 4.4. However, we can make a general statement concerning the sign of ν.

Theorem 4 *Let W^L_t have stationary increments, and suppose for each L and for each $0 \leq t_1 \leq t_2 \leq t_3 \leq t_4$ that $W_{t_4} - W_{t_3}$ and $W_{t_2} - W_{t_1}$ are non-negatively correlated. Then if ν exists, it is non-negative.*

Proof: Since W_t and $W_{t+t'} - W_t$ are non-negatively correlated and $w \mapsto e^{\theta w}$ is non-decreasing for $\theta \geq 0$,

$$\mathbf{E}[e^{\theta W_{t+t'}}] = \mathbf{E}[e^{\theta W_t} e^{\theta(W_{t+t'} - W_t)}] \geq \mathbf{E}[e^{\theta W_t}]\mathbf{E}[e^{\theta(W_{t+t'} - W_t)}] = \mathbf{E}[e^{\theta W_t}]\mathbf{E}[e^{\theta W_{t'}}], \tag{3.34}$$

the last equality being due to the stationarity of the increments of W. Thus $t \mapsto t\lambda_t(\theta)$ is superadditive. Applying the sub-additivity theorem (see Lemma 6.1.11 of [9]) to $-t\lambda_t(\delta)$ we obtain:

$$\lim_{t\to\infty}\lambda_t(\delta) = \sup_{t>0}\lambda_t(\delta) = \lambda(\delta) = 0. \tag{3.35}$$

Thus $\lambda_t(\delta)$ is non-positive for all t and so $\lim_{t\to\infty} t\lambda_t(\delta)$, if it exists, is also non-positive. ■

4 Applications and Examples.

4.1 Heterogeneous Superpositions

We examine $I(0)$ and ν for the class of heterogeneous superpositions described in the section 2. Recall there are $L = \sum_j L_j$ sources in total in the superposition, L_j of type j, each having backward arrival process $A_{j,t}$. We set $c_{j,t}(\theta) = t^{-1} \log \mathbf{E}[e^{\theta A_{j,t}}]$ and assume the existence of the limits $c_j(\theta) = \lim_{t\to\infty} c_{j,t}(\theta)$ and $p_j = \lim_{L\to\infty} L_j/L$. The service rate is sL so $\lambda_t(\theta) = \sum_j p_j c_{j,t}(\theta) - s\theta$.

First, we examine $I(0)$. Consider the case $T = \mathbf{R}_+$. We assume the existence of the limits $C_i(\theta) = \lim_{t\to 0} tc_{i,t}(\theta/t)$. Then

$$\Lambda(\theta) = \sum_j p_j C_j(\theta) - s\theta. \tag{4.1}$$

Thus

$$I(0) = \Lambda^*(0) = \sum_j p_j C_j^*(s_j^{(0)}) \tag{4.2}$$

where $s_j^{(0)} = C_j'(\hat\theta)$, $\hat\theta$ being the unique solution of the equation $\sum_j p_j C_j'(\hat\theta) = s$. For $T = \mathbf{Z}_+$, the same formulae hold, but with the C_j replaced by $c_{j,1}$.

Second, we examine ν. δ is the unique solution of the equation $\lambda(\delta) = 0$, i.e.

$$\sum_j p_j c_j(\delta) - s\delta = 0 \tag{4.3}$$

and so

$$\nu = \sum_j p_j \nu_i \tag{4.4}$$

where

$$\nu_i = -\lim_{t\to\infty} t\left(c_{j,t}(\delta) - s_j^{(\infty)}\delta\right), \tag{4.5}$$

and $s_j^{(\infty)} = c_j(\delta)/\delta$. Note $s_j^{(\infty)} > s_j^{(0)}$ due to the convexity of c_j, since $c_j(0) = 0$. $s_j^{(0)}$ and $s_j^{(\infty)}$ are the usual *effective bandwidths* of sources of type j for zero and infinite buffers respectively. (See, for example, [21] for further details).

4.2 Superpositions of Gaussian Arrival Processes.

In this section we take $W_t^L = A_t^L - sLt$ where for each L, A_t^L is an L-fold superposition of independent copies of A_t: a zero-mean Gaussian process with stationary increments and variance σ_t^2. We make the following Hypotheses concerning A

Hypothesis 4

(i) For some function k_t increasing to $+\infty$ as $t \searrow 0$,

$$\limsup_{t\to 0} k_t m_t < \infty \quad \text{where} \quad m_t = \mathbf{E}[\sup_{0<r<t} |A_r|], \quad \text{and} \quad \limsup_{t\to 0} k_t^2 \sigma_t^2 < \infty. \tag{4.6}$$

The following limits exist as extended real numbers:

(ii) $\sigma^2 := \lim_{t\to\infty} \sigma_t^2/t$.

(iii) $\hat\sigma_0^2 := \lim_{t\to 0} \sigma_t^2/t^2$.

(iv) $\hat\sigma_\infty^2 := \lim_{t\to\infty}(\sigma_t^2 - t\sigma^2)$.

Proposition 1 *Under Hypothesis 4, Theorems 1, 2 and 3 hold with*

$$\delta = 2s/\sigma^2; \qquad I(0) = \frac{s^2}{2\hat\sigma_0^2} \quad \textit{and} \quad \nu = -\frac{2s^2\hat\sigma_\infty^2}{\sigma^4}. \tag{4.7}$$

Proof: Hypothesis 4(ii) means that λ exists, and

$$\lambda_t(\theta) = \theta^2\sigma_t^2/(2t) - s\theta \quad \text{and} \quad \lambda(\theta) = \theta^2\sigma^2/2 - s\theta \tag{4.8}$$

are clearly convex and differentiable.

To check that Hypothesis 1(iv) holds, we show that (i) implies (2.13). By Borell's inequality (see [2], Theorem 5.2),

$$\mathbf{P}[\sup_{0<r<t} |A_r| \geq x] \leq 1 - \Phi\left((x - 2m_t)/\bar\sigma_t\right) \tag{4.9}$$

for any $x \geq 2m_t$, where $\bar\sigma_t = \sup_{0<r<t}\sigma_t$ and Φ is the canonical Gaussian distribution function. From this is follows that for any θ,

$$\mathbf{E}[\exp(\theta k_t \sup_{0<r<t} |A_r|)] \leq e^{2\theta k_t m_t}\left(1 + e^{\theta^2 k_t^2 \bar\sigma_t^2/2}\left(1 - \Phi(-\theta k_t \bar\sigma_t)\right)\right). \tag{4.10}$$

With (i) we get that $k_t m_t$, $k_t\sigma_t$ and hence also $k_t\bar\sigma_t$ remain bounded as $t \to 0$, so that $\mathbf{E}[\exp(\theta k_t \sup_{0<r<t}|A_r|)]$ is also bounded as $t \to 0$. Hence when $k_t \geq 1$,

$$\mathbf{E}[\exp(\theta \sup_{0<r<t} |A_r|)] \leq \mathbf{E}[\exp(\theta k_t \sup_{0<r<t} |A_r|)]^{1/k_t} \to 1 \tag{4.11}$$

as $t \to 0$.

From (4.8) the decay rate is $\delta = 2s/\sigma^2$. Under (iii) the limit

$$\Lambda(\theta) = \lim_{t\to 0} t\lambda_t(\theta/t) = \lim_{t\to 0}\frac{\theta^2\sigma_t^2}{2t^2} - \theta s = \theta^2\hat\sigma_0^2/2 - \theta s \tag{4.12}$$

exists and $\Lambda^*(b) = (b-s)^2/(2\hat\sigma_0^2)$. Thus Hypothesis 2 is satisfied and $I(0) = \Lambda^*(0) = s^2/(2\hat\sigma_0^2)$.

Finally, under (iv) the limit

$$-\nu = \lim_{t\to\infty} t\lambda_t(\delta) = \lim_{t\to\infty} 2(\sigma_t^2 - \sigma^2 t)s^2/\sigma^4 = 2s^2\hat\sigma_\infty^2/\sigma^4 \tag{4.13}$$

exists, so Hypothesis 3(i) is satisfied. ■

The proof goes through for heterogeneous superpositions as described in sections 2 and 4.1, where A_t^L is a superposition of sums of L_j copies of independent Gaussian processes $A_{j,t}$ with mean zero and variance $\sigma_{j,t}^2$ satisfying Hypothesis 4, provided the limits $p_j = \lim_{L\to\infty} L_j/(\sum_j L_j)$ exist. In this case $\sigma_t^2 = \sum_j p_j \sigma_{j,t}^2$.

We note that when A is Brownian motion (take $k_t = t^{-1/2}$), $\hat\sigma_0 = +\infty$ and $\hat\sigma_\infty = 0$, so $I(0) = \nu = 0$. (Compare with the discussion in section 4.5).

4.3 Ornstein-Uhlenbeck Arrival Processes

An example where the workload is modelled by a Gaussian process with stationary increments is the following. Consider a queue with constant service rate, for which the workload W_t is the position component of a stationary Ornstein-Uhlenbeck process with added negative drift. Such an arrival process has been proposed by Norros *et al* [27] as a model of continuous correlated arrivals. It arises as the heavy traffic limit of superposed 2-state Markov fluid sources under suitable rescaling of time and mean activity (see [24]).

We consider the stationary Ornstein-Uhlenbeck velocity process $(V_t,\ t \in \mathbf{R}_+)$, defined to be the solution of the stochastic differential equation

$$dV_t = -V_t dt + \sqrt{2}(s/\kappa)\, dB(t) \tag{4.14}$$

where V_0 is normally distributed with zero mean and variance $(s/\kappa)^2$. Here B is standard Brownian motion, $\kappa > 0$ is a load parameter (the case $\kappa = 0$ corresponding to unit load), and $s > 0$ can be viewed as a service rate. The corresponding position process (with zero initial condition) is

$$A_t = \int_0^t V_s ds, \tag{4.15}$$

and the workload is

$$W_t = A_t - st. \tag{4.16}$$

W_t is Gaussian with mean $-st$ and variance

$$\sigma_t^2 = 2(s/\kappa)^2(t + e^{-t} - 1). \tag{4.17}$$

Hence

$$\lambda_t(\theta) = \frac{\theta^2\sigma_t^2}{2t} - s\theta\ , \qquad \text{and} \qquad \lambda(\theta) = \frac{\theta^2 s^2}{\kappa^2} - s\theta. \tag{4.18}$$

This gives $\sigma^2 = 2s^2/\kappa^2$, $\delta = \kappa^2/s$, $\hat\sigma_0^2 = s^2/\kappa^2$ and $\hat\sigma_\infty^2 = -2s^2/\kappa^2$. Thus items (ii), (iii) and (iv) of Hypothesis 4 is satisfied and

$$I(0) = \kappa^2/2 \qquad \text{and} \qquad \nu = \kappa^2. \tag{4.19}$$

In item (i) we can take $k_t = t^{-1}$ since, as $t \to 0$, $\sigma_t^2 \sim (st/\kappa)^2$ by (4.17), and $m_t \sim \text{const.}\ t^{3/2}$ as we shall now show. We use the integral representation of (4.14),

$$V_t = e^{-t}V_0 + e^{-t}\sqrt{2}(s/\kappa)\int_0^t e^s\ dB(s). \tag{4.20}$$

Then for $t \leq 1$

$$\begin{aligned}
\mathbf{E}[\sup_{0<r<t} |Z_r|] &\leq 2\mathbf{E}[\sup_{0<r<t} Z_r] \qquad \text{by symmetry} && (4.21)\\
&\leq 2t\mathbf{E}[\sup_{0<r<t} V_r] && (4.22)\\
&\leq (2\sqrt{2}s/\kappa)t\mathbf{E}[\sup_{0<r<t}\int_0^r dB(r')] && (4.23)\\
&\leq (4\sqrt{2}s/\kappa)t\mathbf{E}[\sup_{0<r<t} B(r)] && (4.24)\\
&= \text{const.}\ t^{3/2} && (4.25)
\end{aligned}$$

In fact we can easily calculate $I(b)$ numerically . Normalizing b by s, a routine calculation yields

$$t\lambda_t^*(sb/t) = \frac{s^2(b+t)^2}{2\sigma_t^2} = \kappa^2 \frac{(t+b)^2}{4(t-1+e^{-t})}. \tag{4.26}$$

We can also perform the same calculations for heterogeneous superpositions. Arrivals of type j are Ornstein-Uhlenbeck position processes with mean 0 and variance

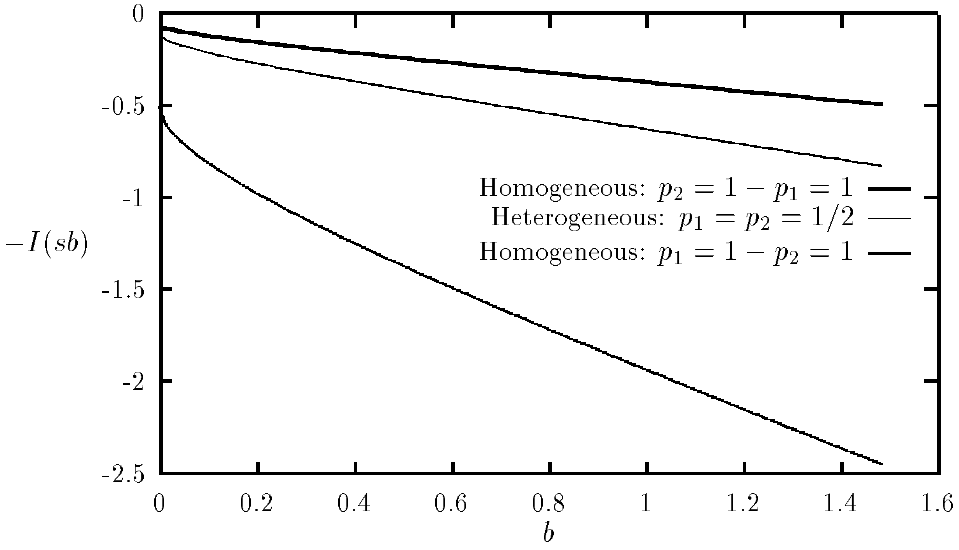
$$\sigma_{j,t}^2 = 2(s/\kappa_j)^2(r_j t + e^{-r_j t} - 1), \tag{4.27}$$

and occur with limiting proportion p_j in the superposition. Here we have included possible time rescalings r_j on each process. Thus the superposition has variance $\sigma_t^2 = \sum_j p_j \sigma_{j,t}^2$, and the analysis of the previous section gives:

$$\begin{aligned}
\sigma^2 &= 2s^2 \sum_j p_j r_j \kappa_j^{-2} && (4.28)\\
\delta^{-1} &= s^{-1} \sum_j p_j r_j \kappa_j^{-2} && (4.29)\\
I(0) &= \left(2 \sum_j p_j r_j^2 \kappa_j^{-2}\right)^{-1} && (4.30)\\
\nu &= \frac{\sum_j p_j r_j^2 \kappa_j^{-2}}{\left(\sum_j p_j r_j \kappa_j^{-2}\right)^2} && (4.31)\\
t\lambda_t^*(sb/t) &= \frac{(b+t)^2}{4\sum_j p_j \kappa_j^{-2}\left(r_j t + e^{-r_j t} - 1\right)}. && (4.32)
\end{aligned}$$

In Figure 2 the curve of $b \mapsto -I(sb)$ is plotted for two types with $r_1 = \kappa_1 = 1$, $r_2 = \kappa_2^2 = 2$ and $p_1 = p_2 = 1/2$. The curve lies between those obtained for homogeneous arrivals of each type separately: these are also plotted.

Figure 2: Ornstein-Uhlenbeck superpositions



æ

4.4 Markov Additive Arrival Processes.

In this section we obtain an expression for ν in the case that the increments of the workload W occur at integer times and are distributed according to the state of an underlying Markov process X describing the configuration of the source of the arrivals. (Specifically, one could consider W to be the single source workload in a homogeneous superposition described in section 2; the corresponding results for heterogeneous superpositions follow from section 4.1). A convenient description for this is that of a Markov Additive Process.

To be precise, let $X = (X_t)_{t\in \mathbf{Z}_+}$ be a irreducible aperiodic Markov process on a state space E (with σ-field $\mathcal{E}$), and adjoin to it an additive component $W = (W_t)_{t\in\mathbf{Z}_+}$ with $W_0 = 0$ such that (X, W) is a Markov process on the state space $E \times \mathbf{R}$. Furthermore, for each $t \in \mathbf{N}$ the joint distribution of the increment $Z_{t+1} := W_{t+1} - W_t$ and X_{t+1}, conditioned on $(X_{t'}, W_{t'})_{0\leq t'\leq t}$ depends only on X_t. This dependence can be expressed through the kernel

$$P(x, G \times B) := \mathbf{P}[X_{t+1} \in G, Z_{t+1} \in B \mid X_t = x], \tag{4.33}$$

for $G \in \mathcal{E}$ and B a Borel set of $\mathbf{R}$.

For $\theta \in \mathbf{R}$ define the transformed kernel $\hat{P}(\theta)$ by

$$\hat{P}(x, G; \theta) := \int P(x, G \times dz)e^{\theta z}, \tag{4.34}$$

and denote by $\hat{P}^t$ its t-fold convolution. A technical recurrence condition on the kernel P (see [20]) is required for what follows:

Hypothesis 5 *(i) There exists a probability measure μ on $E\times\mathbf{R}$, an integer m_0 and real numbers $0 < a \leq b < \infty$ such that*

$$a\mu(G \times B) \leq P^{m_0}(x, G \times B) \leq b\mu(G \times B) \tag{4.35}$$

for all $x \in E$, $G \in \mathcal{E}$ and Borel sets B of $\mathbf{R}$.

(ii) The convex hull of the support of $\mu(E \times \cdot)$ has non-empty interior.

(iii) The set $\{\theta \in \mathbf{R} \mid \hat{\mu}(E, \theta) < \infty\}$ is open.

Note that Hypothesis 5 is automatically satisfied in the case that E is finite and $P(x, E \times dz)$ has compact support for all $x \in E$.

The main technical result we require concerning the kernel $\hat{P}$ is an extension of the standard Perron-Frobenious to non-discrete state spaces. Lemma 3.1. and Lemma 3.4 of [20] and Theorem III.10.1 of [17]). Let q denote the stationary distribution of X.

Proposition 2 *λ is strictly convex and essentially smooth. For all θ in the effective domain of λ, $e^{\lambda(\theta)}$ is the simple maximal eigenvalue of $\hat{P}(\theta)$. The corresponding (right) eigenfunction $r(\cdot;\theta)$ and Radon-Nikodym derivative $d\ell(\cdot;\theta)/dq$ of the (left) eigenmeasure $\ell(\cdot;\theta)$ are uniformly bounded and positive. With the normalization $\int \ell(dx;\theta)r(x;\theta) = 1$*

$$\hat{P}^t(x, G; \theta) = r(x; \theta)\ell(G; \theta)e^{t\lambda(\theta)} \left(1 + \mathbf{O}\left(\varepsilon(\theta)^t\right)\right), \tag{4.36}$$

where $0 < \varepsilon(\theta) < 1$.

Corollary 5

$$\nu = -\lim_{t\to\infty} t\lambda_t(\delta) = -\log\left(\ell(E;\delta)\int q(dx)r(x;\delta)\right). \tag{4.37}$$

Proof: This follows since $t\lambda_t(\delta) = \log\int q(dx)\hat{P}^t(x,E;\delta)$ and $\lambda(\delta)=0$. ■

To calculate $I(0)$ we note that

$$\lambda_1(\theta) = \log\int q(dx)P(x,dy\times dz)e^{\theta z}. \tag{4.38}$$

In the special case, frequent in modelling, that the increment Z_t is a non-random function ζ of X_t when both are conditioned on X_{t-1} we have

$$P(x,dy\times dz) = R(x,dy)\delta_{\zeta(y)}(dz) \tag{4.39}$$

where R is the transition kernel for X, and so (4.38) reduces to

$$\lambda_1(\theta) = \log\int q(dy)e^{\theta\zeta(y)}. \tag{4.40}$$

Finally, note that the complexity of the calculation of ν and $I(0)$ (and indeed the whole curve of $I(b)$) is independent of the number of sources L in the superposition.

Application: two-state Markov chain. We consider a L-fold homogeneous superposition of arrivals streams, each of which is generated by a stationary discrete time Markov chain on two states: on and off. In the on state an arrival of unit length is generated; in the off state no arrival is generated. Transitions from off to on occur with probability a; the reverse transition with probability d. The superposition is serviced are serviced at constant rate sL with $s<1$.

Within the general framework above we have $E=\{0,1\}$, $Z_t=\zeta(X_t)$ with $\zeta(0)=-s, \zeta(1)=1-s$, so that $P(x,dy\times dz)=R(x,dy)\delta_{\zeta(y)}(dz)$ where R is the transition matrix of X:

$$R = \begin{pmatrix} 1-a & a \\ d & 1-d \end{pmatrix} \quad . \tag{4.41}$$

The stationary distribution of R is $q=(\frac{d}{a+d},\frac{a}{a+d})$: the stability condition is $a/(a+d)<s$.

The eigenvector/eigenvalue analysis of $\hat{P}$ yields the following. $e^{\lambda(\theta)}$ is the maximal eigenvalue of the matrix

$$\hat{P}(\theta) = R\ e^{\theta\zeta(\cdot)} = \begin{pmatrix} 1-a & ae^{\theta} \\ d & (1-d)e^{\theta} \end{pmatrix} e^{-s\theta}. \tag{4.42}$$

Let $y=e^\theta$ and set

$$x_\pm = (2a)^{-1}\left(y(1-d)-(1-a)\pm\sqrt{((1-a)-y(1-d))^2+4ady}\right). \tag{4.43}$$

The eigenvalues of $\hat{P}(\theta)$ are $v_\pm = y^{-s}(ax_\pm+1-a)$. Hence $\delta=\log y$ for y such that $1=y^{-s}(ax_+ + 1-a)$. The (unnormalised) eigenmeasures and eigenfunctions are

$$\ell_\pm(\cdot;\theta)=\ell_\pm := (d, ax_\pm) \quad\text{and}\quad r_\pm(\cdot;\theta)=r_\pm := (y,x_\pm) \tag{4.44}$$

respectively. Thus

$$e^{t\lambda_t(\theta)} = \int q(du)\hat{P}^t(u,E;\theta) = \frac{q\cdot r_+ \;\; \ell_+\cdot(1,1)v_+^t}{\ell_+\cdot r_+} + \frac{q\cdot r_- \;\; \ell_-\cdot(1,1)v_-^t}{\ell_-\cdot r_-}, \tag{4.45}$$

and

$$\nu = -\log\left(\frac{q\cdot r_+ \;\; \ell_+\cdot(1,1)}{\ell_+\cdot r_+}\right) = -\log\frac{(ax_+ + dy)(ax_+ + d)}{(ax_+^2 + dy)(a+d)}, \tag{4.46}$$

using the values for x_+ obtained from (4.43) with $y = e^\delta$.

We check that all the required hypothesis are satisfied. Proposition 2 gives Hypotheses 1(i,ii), coupled with the fact that for $\lambda_t^L = \lambda_t$ for homogeneous superpositions. One verifies by explicit differentiation that $\lambda_t'(0)$ is bounded (negatively) away from zero for all t if the stability condition $s > a/(a+d)$ is satisfied: this gives 1(iii). Hypothesis 2(i) follows from the assumed stationary of the arrival streams. As seen previously, Hypothesis 3(i) follows since ν exists and is finite. One sees that from the decomposition (4.45) that $t \mapsto t\lambda_t'(\delta) = A_t + tB_t$ where A_t is bounded and $t(B_{t+1} - B_t) \to 0$ as $t \to \infty$. Hence Hypothesis 3(ii) is satisfied in this model (and indeed in any Markovian model for which such a differentiable decomposition exists). Hypothesis 5 is automatically satisfied for a model with deterministic arrivals and E finite.

At $b = 0$ we find

$$I(0) = \lambda_1^*(0) = -\inf_\theta \lambda_1(\theta) = -\log\inf_\theta \frac{ae^\theta + d}{(a+d)e^{\theta s}} = -\log\frac{a\hat{x} + d}{(a+d)\hat{x}^s} \tag{4.47}$$

where $\hat{x} = sd/(a(1-s))$. This agrees with the large deviation (upper) bound according to Hui [18] for the probability of overflow at a bufferless resource (i.e. with $b = 0$).

The sign of ν can be related to the sign of the correlations of the arrivals process. One sees from (4.46) that $\text{sgn}(\nu) = \text{sgn}\left((x_+ - 1)(x_+ - y)\right)$. But it is shown from Proposition 3 of [4] that $1 - a - d > 0 \Longrightarrow x_+ > y > 1$ while $1 - a - d < 0 \Longrightarrow y > x_+ > 1$ and $1 - a - d = 0 \Longrightarrow x_+ = y > 1$. Furthermore the covariance of successive arrivals $\text{Cov}(Z_t, Z_{t+1}) = ad(1-a-d)/(a+d)^2$. Summarizing:

$$\text{sgn}(1-a-d) = \text{sgn}(\nu) = \text{sgn}\left(\text{Cov}(Z_t, Z_{t+1})\right). \tag{4.48}$$

Bursty sources will modelled with $a + d < 1$: successive arrivals are positively correlated. A sub-bursty Markov model (i.e. with negatively correlated arrivals) has been studied numerically by Choudhury *et al* [6]. It is found that the log-loss curves are concave, and asymptotic to a straight line with *positive* intercept at $b = 0$: correspondingly, our value for ν will be negative.

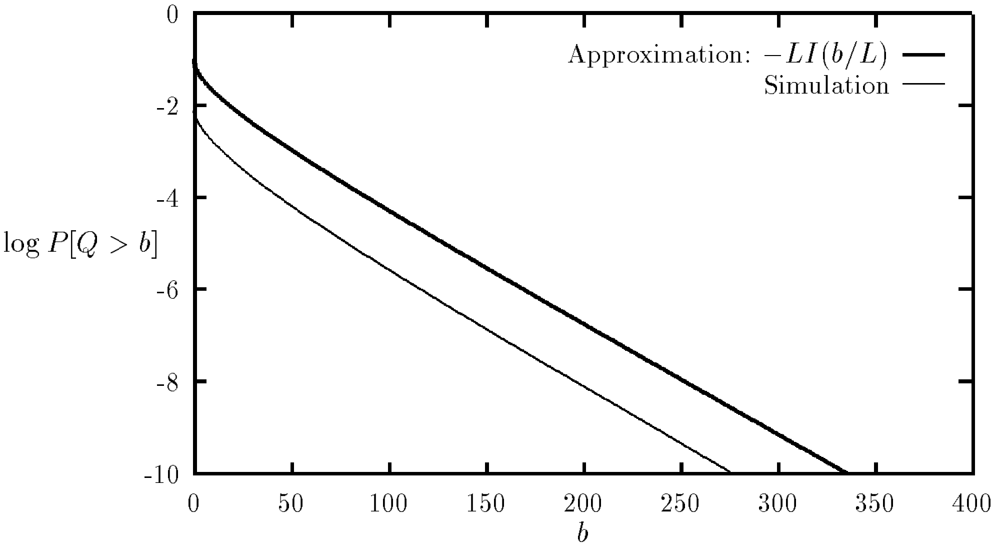
Comparisons and estimates. Theorem 1 can be used as a basis for approximation of superpositions of finitely many lines: we take

$$\mathbf{P}[Q^L > b] \approx e^{-LI(b/L)}. \tag{4.49}$$

The λ_t are using (4.45) with $y = e^\theta$. The resulting approximation is compared with simulations in three cases. Figure 3 takes $a = 0.03$, $d = 0.045$, $L = 84$ and $s = 40/84$, a superposition of highly bursty sources. In Figure 4 the parameters $a = 0.3$ $d = 0.5344$, $L = 100$ and $s = 40/100$ are chosen to make $\nu = I(0)$: the curve is very close to linear. In Figure 5 a sub-bursty case $a = 0.55$, $d = 0.825$, $L = 84$ and $s = 40/84$ is shown. In these examples, the shape of the log-loss curve is closely reproduced by the approximation, but with a shift which makes the approximation conservative (in these cases). This may well be a limitation of the first-order large deviation method.

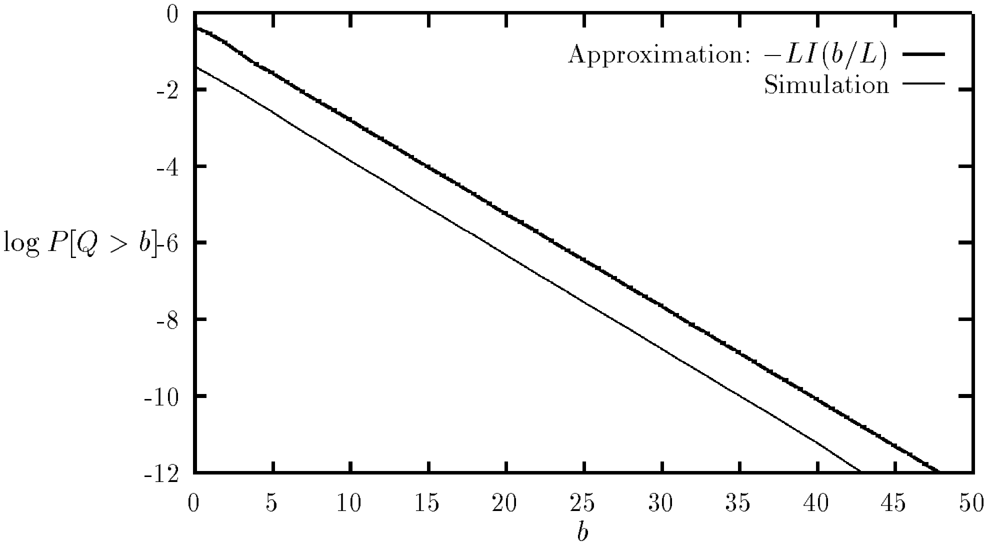
In fact the discrepancy is well within the $\mathbf{O}(\log L)$ refinements to the large deviation estimate of $\mathbf{P}[W_1^L > 0]$ in (4.47), so some improvement may be possible with further work involving these refinements to the first-order large deviation result. As a final numerical example we take a large superposition of extremely bursty sources: $a = .0003$, $d = .0007$, $s = 400$, $L = 1000$. For these parameters $-LI(0)/\log 10 = 9.8$ and $-L\nu/\log 10 = 20.2$: the desired loss probabilities of 10^{-9} are already obtained at $b = 0$.

Figure 3: $a + d < 1$ with $a = 0.03$, $d = 0.045$, $s = 40$ and $L = 84$.



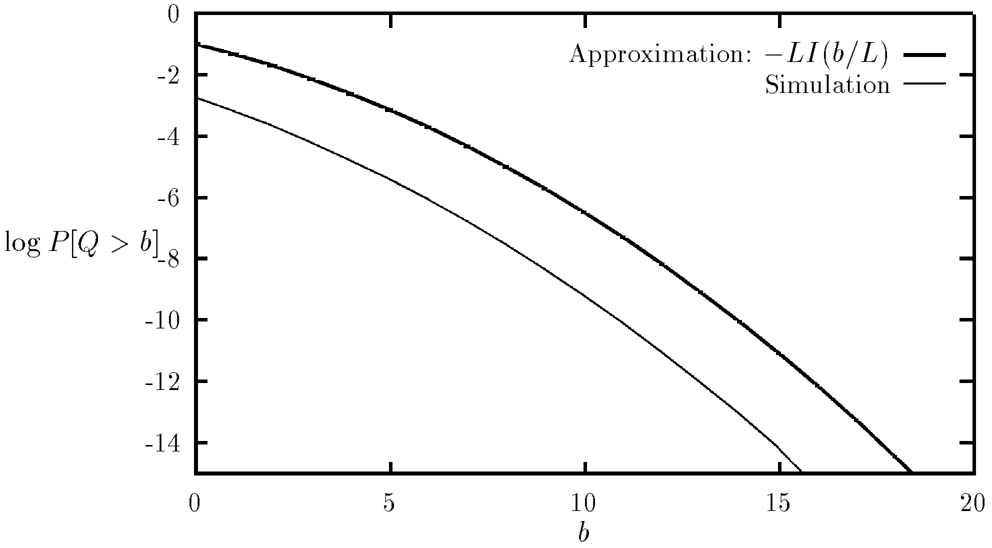
æ

Figure 4: $I(0) = \nu$ with $a = 0.3$, $d = .5344$, $s = 40$ and $L = 100$



æ

Figure 5: $a + d > 1$ with $a = 0.55$, $d = 0.825$, $s = 40$ and $L = 84$.



æ

4.5 Poissonian Arrivals.

We conclude with a brief discussion of Poissonian arrivals. In this case one sees easily that $\lambda_t = \lambda$ independent of t, and hence from (1.6)

$$I(b) = \inf_t t\lambda^*(b/t) = b\delta. \tag{4.50}$$

$I(0) = \nu = 0$: we draw the conclusion that there are no economies of scale to be obtained from a superposition of Poissonian arrivals at any buffer size b. In contrast, Bernoulli arrivals will generally give $I(0) > \nu = 0$: take $a = 1 - d$ in the on-off model as an example. The difference in $I(0)$ between the two cases can be shown to go to zero if one constructs the Poissonian arrivals process as a continuum limit of Bernoulli arrivals.

Acknowledgements. The authors thank John Lewis, Neil O'Connell, Ad Ridder, Fergal Toomey and particularly Raymond Russell for their comments. Raymond Russell wrote the software used in the simulations of the on-off Markov model.

References

[1] D. Anick, D. Mitra and M.M. Sondhi (1982) Stochastic theory of a data-handling system with multiple sources. *Bell Sys. Tech. J.*, 61:1872–1894

[2] C. Borell (1975). The Brunn-Minkowski inequality in Gauss space. *Invent. Math.*, 30:205–216.

[3] J.A. Bucklew (1990). *Large deviation techniques in decision, simulation and estimation.* Wiley, New York.

[4] E. Buffet and N.G. Duffield (1994) Exponential upper bounds via martingales for multiplexers with Markovian arrivals. *J. Appl. Prob.* 31:1049–1061.

[5] C.S. Chang (1994). Stability, queue length and delay of deterministic and stochastic queueing networks. *IEEE Trans. on Automatic Control.* 39:913-931.

[6] G.L. Choudhury, D.M. Lucantoni and W. Whitt (1993). Squeezing the most out of ATM. *IEEE Transactions on Communications,* to appear.

[7] T.J. Corcoran (1994) Prediction of ATM multiplexer performance by simulation and analysis of a model of packetized voice traffic. M.Sc. Thesis, Dublin City University.

[8] C. Courcoubetis, G. Kesidis, A. Ridder, J. Walrand and R. Weber (1994). Admission control and routing in ATM networks using inferences from measured buffered occupancy. *IEEE Transactions on Communications.* to appear.

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

[10] N.G. Duffield (1994). Exponential bounds for queues with Markovian arrivals. *Queueing Systems,* 17:413–430

[11] N.G. Duffield (1994). Economies of scale in queues with sources having power-law large deviation scalings. Submitted to *J. Appl. Prob.*

[12] N.G. Duffield and N. O'Connell (1993). Large deviations and overflow probabilities for the general single-server queue, with applications. *Math. Proc. Camb. Phil. Soc.* to appear.

[13] R.S. Ellis (1985). *Entropy, Large Deviations, and Statistical Mechanics*, Springer, New York.

[14] A.I. Elwalid, D. Mitra and T.E. Stern (1991). Statistical multiplexing of Markov modulated sources: theory and computational algorithms. In: *Teletraffic and Datatraffic in a period of change, ITC-13*, A. Jensen & V.B. Iversen (Eds.) Elsevier Science Publishers B.V. (North-Holland).

[15] R.J. Gibbens and P.J. Hunt (1991). Effective Bandwidths for the multi-type UAS channel *Queueing Systems*, 9:17–28

[16] P.W. Glynn and W. Whitt (1994). Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *J. Appl. Prob.* 31A:131–159

[17] T.E. Harris (1963), *The theory of branching processes.* Springer, Berlin.

[18] J.Y. Hui (1988). Resource allocation for broadband networks. *IEEE J. Selected Areas in Commun.* 6:1598–1608

[19] J.Y. Hui (1990). *Switching and traffic theory for integrated broadband networks.* Kluwer. Boston.

[20] I. Iscoe, P. Ney and E. Nummelin (1985). Large deviations of uniformly recurrent Markov additive processes. *Adv. in Appl. Math.* 6:373–412

[21] F.P. Kelly (1991). Effective bandwidths at multi-type queues. *Queueing Systems* 9:5–16

[22] G. Kesidis, J. Walrand and C.S. Chang (1993). Effective bandwidths for multiclass Markov fluids and other ATM Sources. *IEEE/ACM Trans. Networking* 1:424-428.

[23] L. Kosten (1988) Stochastic Theory of data handling systems with groups of multiple sources. *Proc. 2nd Int. Smyp. on the Performance of Computer Communication Systems,* eds. H. Rudin & W. Bux, North-Holland.

[24] V. Kulkarni and T. Rolski (1993). Fluid Model Driven by an Ornstein-Uhlenbeck Process. Preprint.

[25] W.E. Leland, M.S. Taqqu, W. Willinger and D.V. Wilson (1993). On the self-similar nature of Ethernet traffic. *ACM SIGCOMM Computer Communications Review* 23:183-193.

[26] J.T. Lewis and C.-E. Pfister (1994). Thermodynamic probability theory: some aspects of large deviations. *Theor. Prob. Appl.*, to appear.

[27] I. Norros, J.W. Roberts, A. Simonian and J. Virtamo (1991). The superposition of variable bitrate sources in ATM multiplexers. *IEEE J. Selected Areas in Commun.* 9:378–387.

[28] C. Rasmussen, J.H. Sorensen, K.S. Kvols and S.B. Jacobsen (1991). Source-independent call acceptance procedures in ATM networks. *IEEE J. Selected Areas in Commun.* 9:351–358.

[29] R.T. Rockafellar (1970) *Convex Analysis.* Princeton University Press, Princeton.

[30] A. Simonian and J. Guibert (1994). Large deviations approximation for fluid queues fed by a large number of on-off sources. *Proceedings of ITC 14, Antibes, 1994* 1013–1022.

[31] K. Sriram and W. Whitt (1986). Characterizing superposition arrival processes in packet multiplexers for voice data. *IEEE J. Select. Areas Commun.* 4:833–846

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

[33] A. Weiss (1986). A new technique for analysing large traffic systems. *J. Appl. Prob.* 18:506–532

[34] W. Whitt (1993). Tail probabilities with statistical multiplexing and effective bandwidths in multi-class queues. *Telecommunications Systems.* 2:71-107