Exponential Bounds for Queues with Markovian Arrivals

by

N.G. Duffield

School of Mathematical Sciences,
Dublin City University,
Dublin 9, Ireland.

and

School of Theoretical Physics,
Dublin Institute for Advanced Studies,
10 Burlington Road, Dublin 4, Ireland.

Abstract. Exponential bounds $\mathbb{P}[\text{queue} \geq b] \leq \varphi e^{-\gamma b}$ are found for queues whose increments are described by Markov Additive Processes. This is done application of maximal inequalities to exponential martingales for such processes. Through a thermodynamic approach the constant γ is shown to be the decay rate for an asymptotic lower bound for the queue length distribution. The class of arrival processes considered includes a wide variety of Markovian multiplexer models, and a general treatment of these is given, along with that of Markov modulated arrivals. Particular attention is paid to the calculation of the prefactor φ.

Key words: Queueing Theory, Large Deviations, Martingales, Risk Theory, Markov Additive Processes, ATM Multiplexers, Effective Bandwidths.

1. Introduction.

The problem of finding the queue length distribution in a queue with non-independent arrivals has attracted much attention recently due to applications in the design of multiplexers for the emergent asynchronous transfer mode (ATM) of data transmission in integrated services digital networks (ISDN). From the technological point of view it is required to guarantee sufficiently good quality of service: loss probabilities must be appropriately small and waiting times sufficiently short. The problem is resistant to simple exact treatment due to the nature of the arrival process. It is a superposition of sources which are typically bursty, in the sense that their activity is highly correlated into bursts rather than occurring independently at different times; and periodic (when viewed at the short time scales of the multiplexer output) either due to their origin (e.g. periodic sampling of voice traffic) or their occupation of periodic slots allocated for transmission. The goal of analysis is to provide mechanisms for design and performance prediction, and algorithms for allocation of resources during the operation of such devices. It is desirable that the results of such analysis be conservative in the sense that they should not overestimate the capacity of resources.

In this paper we present a treatment based on exponential martingales which allows us to obtain exponential upper bounds of the form $\mathbb{P}[\text{queue} \geq b] \leq \varphi e^{-\gamma b}$ for queues whose inputs are described by a Markov Additive Process (MAP). Within this framework there is some controlling Markov process $X = (X_t)$ of which the workload $W = (W_t)$ of the queue is a functional in such a manner that the pair (X, W) is also a Markov process. One can think of the states of X as labeling the states of the source of the queue, although they could also include a component describing a random service rate. One particular class of examples of MAP's is when the arrivals at the queue are deterministic functions of the Markov chain X. Within this class one can make a a large number of multiplexer models. Inhomogeneous superpositions of such processes are also included in this class. The bounds obtained generalize some previous work of Buffet and Duffield [7] where the homogeneous superposition of two-state Markov sources was considered.

The martingale used is an example of a construction which tells us that if $(U_t)_{t\in\mathbb{R}}$ is a stationary Markov process on state space Ξ, and f is an invariant integrable function on Ξ, then $f(U_t)$ is a martingale with respect to the canonical filtration generated by U. It is this observation that enables us to construct the appropriate martingale for the case where the arrivals are a function of a Markov process. The upper bound on the queue length is obtained by using maximal inequalities for positive supermartingales based on Doob's optional stopping theorem:

if $M = (M_t)_{t\in\mathbb{Z}^+}$ is a positive supermartingale with respect to some filtration, then for $m > 0$, $\mathbb{P}[\sup_{t\in\mathbb{Z}^+} M_t \geq m] \leq m^{-1}\mathbb{E}[M_0]$. Some care is needed in the application order to obtain the best (i.e. least) prefactor φ in the upper bound.

Exponential martingales were used by Kingman [18] to bound the queue length in the queue $M/G/1$. The particular martingale used in the present paper is formally similar to that used by Baccelli and Makowski to treat queues with Markov-modulated Poissonian arrivals [3,4]. The same general methodology has been used for risk theory in a finite state Markovian environment by Reinhard [23], by Björk and Grandell [5], and by Asmussen [1]. (See also the book of Grandell [12] for a comprehensive review). It turns out that in the application of our result to risk theoretic models the prefactor φ obtained in the present paper is smaller in general than those obtained hitherto.

This is an advantage, because as well as the decay rate γ, we also attach great importance to the prefactor φ in the upper bound. The motivation for this comes from the main application we have in mind: that the arrival process is a superposition of independent Markov sources. The composite arrival process is a function of the product of the underlying Markov processes of the sources, and so the results apply immediately. Such a scheme is used to model ATM multiplexers. For example, in a wide class of models with L superposed identical independent sources in a queue of service rate s we are able to obtain the prefactor in the form $\varphi = \Phi^L$, where Φ derives in an explicit manner from the prefactor of a single source in a queue with rate s/L.

The decay constant obtained is the best possible in the sense that it is also asymptotically the exponential decay rate of a *lower* bound for the queue length:

$$\liminf_{b\to\infty} b^{-1} \log \mathbb{P}[\text{queue} > b] \geq -\gamma \quad . \tag{1.1}$$

Such a result (but with a limit and equality) was obtained by explicit calculation in the (Poissonian) risk-theoretic case by Martin-Löf [20]. This thermodynamic approach relates the decay constant γ to the cumulant generating function for the arrival process, using the convexity properties of the cumulant. To be more specific, if for example the queue has fixed service rate s and the work arriving at the queue in the interval $(0, -t]$ is $A(t)$ and we define the cumulant $c(r) = \lim_{t\to\infty} t^{-1} \log \mathbb{E}[\exp(rA(t))]$ then the decay rate is $\gamma = \sup\{r \mid c(r) \leq rs\}$. In fact (1.1) relies only on the existence of large-deviation properties of the family of distributions of $\big(A(t)\big)_{t\geq 0}$. This decay rate γ has been obtained through a heuristic large deviation argument by Kesidis, Walrand and Chang [17], but as far as we

are aware, a general demonstration of the lower bound (1.1) has not been given rigorously before for non-independent arrivals. For the class of MAP's we consider, the required large deviation properties follow from work of Iscoe, Ney and Nummelin [15]. (See also the book of Bucklew [6] for a discussion of large deviations for Markov processes).

It is worthwhile to compare the applications of the present results to multiplexers with those of existing treatments. Broadly speaking we can divide these into exact calculations for all queue lengths b, asymptotic ones as $b \to \infty$, and bufferless models, i.e. $b = 0$. The queue length distribution for homogeneous superpositions a continuous-time Markov fluid-flow model has been found as an expansion in terms of eigenvalues and eigenvectors of a characteristic matrix some time ago by Anick, Mitra and Sondhi [2]. Exact treatments have been given for the corresponding heterogeneous problem by Kosten [19], and for the heterogeneous N-state Markov modulated arrival processes by Elwalid, Mitra and Stern [10]. Whereas one can recover the asymptotic decay constant γ fairly easily, further detail at finite queue lengths b seems hard to extract due to the complexity of the algorithms involved. We mention also the non-exponential Benes bounds due to Norros *et.al.* [22]. Gibbens and Hunt [11] have considered the limit $b \to \infty$ of Kosten's treatment, an approach developed further by Whitt [27]. This takes us into the regime of the *ab initio* asymptotic methods: the heuristic large deviation calculations of Weiss [26] and of Kesidis, Walrand and Chang [17], and the entropic techniques of de Veciana, Olivier and Walrand [25]. For $b = 0$, large deviation bounds for the arrival process alone are obtained by Hui [13]. The advantage of the bound of the present paper in this context is that it provides a simple estimate valid for all queue lengths b. The bound for homogeneous L-fold superpositions of On-Off Markov sources has been worked out fully by Buffet and Duffield [7]. In this case the optimal prefactor φ, which can be written as Φ^L for some $\Phi < 1$ is exactly that found by Hui. Thus our bound can be seen as a simple interpolation between the bufferless resource models $b = 0$ and the asymptotic case $b \to \infty$, taking into account the large deviation properties of the tail probabilities in terms of both b and L. This is important, since it is not clear at present whether multiplexers will operate in a short queue regime or an asymptotic one.

The paper is organized as follows. In section 2 the large deviation lower bounds for the queue-length distribution are obtained. In section 3 the exponential martingale is given and upper bounds for the queue length distribution derived. In section 4 we apply these in a number of directions: to Markov modulated arrivals (section 4.1); to arrivals which are deterministic functions of the control process X (section 4.2); to superpositions of MAP's (section 4.3); in particular heterogeneous

superpositions of arrivals, with service at a constant rate (section 4.4). In section 4.5 a special case of this is worked out in detail: the homogeneous superposition of multistate On-Off sources. Finally in section 4.6 we give a further bound for a heterogeneous superposition of differing homogeneous groups.

2. Lower Bounds.

We consider a queue with infinite buffer. Tasks arrive at the queue and are processed in the first-come first-served (FCFS) discipline. Let $A = (A_t)_{t\geq 0}$ denote the backward arrival process at the queue: $A_0 = 0$ and for $t > 0$, A_t is the work brought by the tasks which arrive in the interval $(0, -t]$. The service performed is described by an increasing positive function $S = (S_t)_{t\geq 0}$. $S_t = 0$ and for $t > 0$ S_t is the service which could be performed in the interval $(0, -t]$ if the server were never idle. For example, with a deterministic service rate s then $S_t = st$. But generally the service can be a random variable. Define the workload (or excess work) $W = (W_t)_{t\geq 0}$ by $W_t = A_t - S_t$. Then the queue length at time zero is

$$Q := \sup_{t\geq 0} W_t \quad .$$

For each $t > 0$ and $r \in \mathbb{R}$ define

$$\lambda_t(r) = t^{-1} \log \mathbb{E}[\exp(rW_t)] \qquad \text{and} \qquad \lambda(r) = \lim_{t\to\infty} \lambda_t(r) \quad , \tag{2.1}$$

where the limit exists, each possibly infinite. Note by Hölder's inequality that the λ_t and hence λ are convex functions.

Theorem 1. *Assume that the distributions of $(W_t/t)_{t>0}$ satisfy a large deviation lower bound with some rate function I, i.e.*

$$\liminf_{t\to\infty} t^{-1} \log \mathbb{P}[W_t/t > w] \geq -I(w) \quad . \tag{2.2}$$

Then

$$\liminf_{b\to\infty} b^{-1} \log \mathbb{P}[Q > b] \geq -\inf_{\tau>0} \tau I(\tau^{-1}) \quad . \tag{2.3}$$

Proof:

$$\mathbb{P}[Q > b] \geq \mathbb{P}[W_t/t > b/t]$$

for any $t > 0$ so that in particular for $t = t_b := \tau b$ (any $\tau > 0$) we have

$$\liminf_{b\to\infty} \frac{1}{b} \log \mathbb{P}[Q > b] \geq \tau \liminf_{t\to\infty} \frac{1}{t} \log \mathbb{P}[W_t/t > \tau^{-1}]$$
$$\geq -\tau I(\tau^{-1}) \quad .$$

But $\tau > 0$ is arbitrary, so the stated variational form follows.

■

As we remarked in the introduction, the variational expression in the right hand side of eq. (2.3) has been previously been proposed as the decay constant for the tail of the queue on the basis of heuristic large deviation arguments, but the argument seems not to have been made rigorous before apart from the case of Poissonian arrivals. This bound now established, we can relate the infimum in eq. (2.3) to the function λ under various hypotheses concerning λ in Theorem 2 below. (We remark that this next step was made under stronger assumptions on the arrival process A by Kesidis *et.al.* [17]).

First some terminology from convex analysis from the book of Rockafellar [24]. The effective domain of a (possibly infinite) convex function is the region where it is finite. The function λ is essentially smooth if it is differentiable on the interior of its effective domain and $|\lambda(r_n)| \to \infty$ for any sequence of points (r_n) in the interior of the effective domain converging to a point on its boundary. Finally, λ is essentially strictly convex if it is strictly convex on its effective domain. The Legendre transform λ^* of a convex function λ is defined as $\lambda^*(x) = \sup_r \big(xr - \lambda(r)\big)$. λ^* is essentially convex and essentially smooth if and only if λ is.

Theorem 2.

(1) Assume that I in Theorem 1 is the Legendre transform of λ. Then

$$\inf_{\tau>0} \tau I(\tau^{-1}) \geq \hat{\gamma} := \sup\{\gamma \mid \lambda(\gamma) \leq 0\}$$

(2) Assume λ to be essentially smooth and essentially strictly convex. Then $\lambda'(0) < 0$ iff $\hat{\gamma}$ is the unique positive solution of the equation $\lambda(\gamma) = 0$.

Proof: (1) Pick γ such that $\lambda(\gamma) \leq 0$. Then since I is the Legendre transform of λ,

$$\inf_{\tau>0} \tau I(\tau^{-1}) = \inf_{\tau>0} \tau \sup_{\gamma'} \big(\gamma'\tau^{-1} - \lambda(\gamma')\big) \geq \inf_{\tau>0} \big(\gamma - \tau\lambda(\gamma)\big) \geq \gamma \quad .$$

The bound is got by taking the supremum over all γ such that $\lambda(\gamma) \leq 0$.

(2) First, the assumptions of (1) are satisfied since the essential smoothness of λ ensures by the Gärtner-Ellis Theorem [9] that the distributions of $W(t)/t$ satisfy a Large Deviation Principle with rate function I being the Legendre transform of λ. In particular, equation (2.2) is satisfied with this I. Since λ is essentially strictly convex, if $\lambda'(0) \geq 0$ then there is no positive solution γ of $\lambda(\gamma) = 0$, since then $\lambda(\gamma)$ must be strictly positive for all $\gamma > 0$. Otherwise there is a unique solution, namely $\hat{\gamma}$.

Since λ is essentially smooth and essentially strictly convex, so is its Legendre transform I. In particular I is differentiable on the interior of its effective domain. $\tau \mapsto \tau I(\tau^{-1})$ is stationary when for the value of τ such that

$$0 = (d/d\tau)\tau I(\tau^{-1}) = I(\tau^{-1}) - \tau^{-1} I'(\tau^{-1}) \quad . \tag{2.4}$$

But since I and λ are Legendre transforms of each other, then for this value of τ, $I'(\tau^{-1}) = r$ where r is the unique solution of $\lambda'(r) = \tau^{-1}$. Thus

$$\begin{aligned} r &= \tau I(\tau^{-1}) && \text{by eq. (2.4)} \\ &= \tau\big(r\tau^{-1} - \lambda(r)\big) && \text{by definition of the Legendre transform} \\ &= r - \tau\lambda(r) \end{aligned}$$

so that for this value r is the unique solution of the equation $\lambda(r) = 0$ so that $r = \hat{\gamma}$. This stationary value is an upper bound for the infimum $\inf_{\tau>0} \tau I(\tau^{-1})$, and by part (1) is also a lower bound, and hence it equal to it.

■

3. Upper Bounds for Markovian Arrivals

In this section we restrict our attention to the case that the increments of the (time-reversed) 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. A convenient description for this is that of a Markov Additive Process. To be precise, upon some underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$, let $X = (X_t)_{t \in \mathbb{Z}^+}$ be a stationary ergodic Markov process on a state space E (with σ-field $\mathcal{E}$), and adjoin to it an additive component $W = (W_t)_{t \in \mathbb{Z}^+}$ with $W_0 = 0$ such that (X, W) is a Markov process on the state space $E \times \mathbb{R}^+$. Furthermore, for each $t \in \mathbb{N}$ 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

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

for $G \in \mathcal{E}$ and B a Borel set of $\mathbb{R}^+$. We emphasize that in accordance with the conventions of the previous section, for $t \in \mathbb{N}$, W_t is the excess work due to arrivals at times $\{-t, \ldots, -2, -1\}$ so that the queue length at time zero is $Q = \sup_{t \in \mathbb{Z}^+} W_t$.

Within the framework of Markov Additive Processes the connection with the lower bounds of the previous section follows immediately from results of Iscoe, Ney and Nummelin [15] on large deviations for MAP's which we state in Lemma 1 below. For $\gamma \in \mathbb{R}$ define the transformed kernel $\hat{P}(\gamma)$ by

$$\hat{P}(x, G; \gamma) := \int P(x, G \times dz) e^{\gamma z} \quad .$$

A technical recurrence condition for the kernel P (eq. (3.1) of [15]) is required for what follows, and we assume it to be satisfied. Recall from equation (2.1) the definition of λ as the cumulant of W. Henceforth we assume the non-trivial case that the effective domains of λ and λ_t include a common open interval about 0.

Lemma 1.

(1) For all γ in the effective domain of λ, $e^{\lambda(\gamma)}$ is the simple maximal eigenvalue of $\hat{P}(\gamma)$

(2) The corresponding eigenfunction $(v(x;\gamma) : x \in E)$ (i.e. the function such that $e^{\lambda(\gamma)} v(x;\gamma) = \int \hat{P}(x, dy; \gamma) v(y;\gamma)$) is positive and bounded above.

(3) λ and each λ_t; $t \in \mathbb{N}$ are strictly convex and essentially smooth.

(4) Let $\mathcal{F}_t$ be the σ-algebra generated by $(X_0, \ldots, X_t, W_0, \ldots, W_t)$. The sequence of functions $(M_t(\gamma))_{t \in \mathbb{Z}^+}$ defined by

$$M_t(\gamma) := e^{\gamma W_t - t\lambda(\gamma)} v(X_t; \gamma)$$

is a martingale with respect to the filtration $(\mathcal{F}_t)_{t \in \mathbb{Z}^+}$.

Proof and comments: (1) and (2) are proved in Lemma 3.1 of [15]. These can be regarded as an extension of the standard Perron-Frobenious Theorem on the maximal eigenvalue and corresponding eigenvector of matrices with positive entries, a result which would suffice for E a finite discrete set. But we do not impose this restriction. (3) and (4) are proved in Lemma 3.4 and Lemma 3.2 of [15] respectively. We repeat the simple proof of (4):

$$\begin{aligned}
\mathbb{E}[M_t(\gamma) \mid \mathcal{F}_{t-1}] &= M_{t-1}(\gamma) \left(v(X_{t-1}; \gamma) e^{\lambda(\gamma)} \right)^{-1} \mathbb{E}[e^{\gamma Z_t} v(X_t; \gamma) \mid \mathcal{F}_{t-1}] \\
&= M_{t-1}(\gamma) \left(v(X_{t-1}; \gamma) e^{\lambda(\gamma)} \right)^{-1} \int P(X_{t-1}, dx \times dz) e^{\gamma z} v(x; \gamma) \\
&= M_{t-1}(\gamma) \left(v(X_{t-1}; \gamma) e^{\lambda(\gamma)} \right)^{-1} \int \hat{P}(X_{t-1}, dx; \gamma) v(x; \gamma) \\
&= M_{t-1}(\gamma) \quad . \qquad \blacksquare
\end{aligned}$$

The basic stability criterion for the queue is that the asymptotic decay rate $\hat{\gamma}$ is strictly positive.

Lemma 2. *$\hat{\gamma} > 0$ if and only if $\mathbb{E}[Z_1] < 0$.*

Proof: Since λ (resp. λ_t) is essentially smooth $\lambda'(0)$ (resp. $\lambda_t'(0)$) exists, and since λ is the pointwise limit as $t \to \infty$ of the convex functions λ_t differentiable at zero, then by (for example) Lemma IV.6.3 of [9] $\lambda'(0) = \lim_{t\to\infty} \lambda_t'(0)$. Thus we have $\lambda'(0) = \lim_{t\to\infty} \lambda_t'(0) = \lim_{t\to\infty} t^{-1}\mathbb{E}[W_t] = \mathbb{E}[Z_1]$. Since λ is strictly convex, if $\lambda'(0) \geq 0$ then $\lambda(\gamma)$ is strictly positive for all $\gamma > 0$ and so $\hat{\gamma} = 0$. If $\lambda'(0) < 0$ then either the equation $\lambda(\gamma) = 0$ has exactly one positive solution, namely $\hat{\gamma}$, or $\lambda(\gamma) < 0$ for all $\gamma > 0$ in which case we take $\hat{\gamma} = +\infty$.

■

We now turn to the main result of this section: the upper bounds for the queue length distribution. The martingale of Lemma 1(4) is our fundamental tool for this. First we normalize the eigenfunction $v(\cdot;\gamma)$ so that $\mathbb{E}[v(X_0;\gamma)] = 1$ and hence $\mathbb{E}[M_0(\gamma)] = 1$.

Theorem 3. *Suppose $\mathbb{E}[Z_1] < 0$ and let $b > 0$. Then for any $\gamma \in [0, \hat{\gamma}]$*

$$\mathbb{P}[Q \geq b \mid \mathcal{F}_0] \leq v(X_0;\gamma)\varphi(\gamma)e^{-\gamma b} \qquad \text{and so} \qquad \mathbb{P}[Q \geq b] \leq \varphi(\gamma)e^{-\gamma b} \quad (3.1)$$

where

$$\varphi(\gamma) = \text{ess sup}\left(I_{\{Z_t>0\}}/v(X_t;\gamma)\right) \quad . \qquad (3.2)$$

Here $I_{\{Z_t>0\}}$ is the indicator function of the set $\{Z_t > 0\}$ and the essential supremum is with respect to the underlying measure $\mathbb{P}$. By stationarity this is independent of $t \geq 1$. Finally, the asymptotic decay rate of the queue length distribution:

$$\lim_{b\to\infty} b^{-1} \log \mathbb{P}[Q \geq b] = -\hat{\gamma} \quad . \qquad (3.3)$$

Proof: Let $Q_t = \max_{0\leq t'\leq t} W_{t'}$ so that since $b > 0$

$$\{Q \geq b\} = \{\sup_{t\geq 1} Q_t \geq b\} = \cup_{t\geq 1}\{Q_t \geq b, Q_{t-1} < b\} \subset \cup_{t\geq 1}\{W_t \geq b, Z_t > 0\} \quad .$$

Now for $\gamma \geq 0$, $\{W_t \geq b\} \subseteq \{e^{\gamma W_t} \geq e^{\gamma b}\} \subseteq \{M_t(\gamma) \geq e^{\gamma b}v(X_t;\gamma)\}$ since $\lambda(\gamma) \leq 0$ for $\gamma \leq \hat{\gamma}$. Therefore for $t \geq 1$,

$$\begin{aligned}\{W_t \geq b, Z_t > 0\} &\subseteq \{M_t(\gamma) \geq e^{\gamma b}v(X_t;\gamma), Z_t > 0\} \\ &\subseteq \{M_t(\gamma) \geq e^{\gamma b}/\varphi(\gamma)\} \quad .\end{aligned}$$

(Here, if necessary, we can choose a version of (X, W) for which the essential supremum in (3.2) is never exceeded). Hence by the maximal inequality for positive supermartingales (see e.g. the book of Neveu [21]),

$$P[Q \geq b \mid \mathcal{F}_0] \leq \mathbb{P}[\sup_{t \geq 1} M_t(\gamma) \geq e^{\gamma b}/\varphi(\gamma) \mid \mathcal{F}_0] \leq v(X_0;\gamma)\varphi(\gamma)e^{-\gamma b} \quad ,$$

since $M_0(\gamma) = v(X_0;\gamma)$. The second inequality of (3.1) follows by taking the expectation over X_0 using the stated normalization $\mathbb{E}[v(X_0;\gamma)] = 1$.

Choosing $\gamma = \hat{\gamma}$ from (3.1) $\limsup_{b\to\infty} b^{-1} \log \mathbb{P}[Q \geq b] \leq -\hat{\gamma}$ and so (3.3) follows by combination with Theorems 1 and 2(2).

■

Finally we note that since we work in the FCFS service discipline, our upper bound (indeed any upper bound) for $\mathbb{P}[Q \geq b]$ is in turn an upper bound for the probability of overflow for the same arrival process into a buffer of finite size b.

4. Examples.

In this section we apply Theorem 3 to various classes of queueing theoretic models. In section 4.1 we show how to calculate $\hat{\gamma}$ and φ for Markov modulated arrivals. In section 4.2 the same is done in the case that the workload is a deterministic function of the underlying process X. In the context of multiplexers we are interested in workloads which are superpositions of workloads from a number of independent sources. The service function can still be random: in the multiplexer context this could correspond to a service rate controlled externally in order to regulate output of the multiplexer into a network. Thus X would have components describing the state of the network as well as the state of the sources. This is dealt with in section 4.3, and deterministic service in section 4.4. In section 4.5 we work an example of a homogeneous superposition of L multistate On-Off sources. That is to say, for each source in the superposition, the state space of the control process X can we written as a disjoint union $E = E_0 \cup E_1$, and the increments $A_t - A_{t-1}$ of the arrival process are deterministic, being 0 or 1 according to whether the state of the control process X_t is in E_0 or E_1. The prefactor can be written $\varphi = \Phi^L$ where Φ is calculated explicitly in terms of single source eigenfunctions. Finally in section 4.6 we give an alternative prefactor, polynomial in b, which can be used in the case of heterogeneous superposition of groups of homogeneous sources.

4.1. Markov modulated increments.

Consider the subclass of MAP's in which the distribution of the increment Z_t, conditional on X_t, is independent of X_{t-1}. In other words, increment Z_t is chosen according to a distribution determined by X_t. Assuming the existence of a regular conditional distribution $Q(y, B) := \mathbb{P}[Z_t \in B \mid X_t = y]$ then we can write

$$\hat{P}(x, dy; \gamma) = R(x, dy)e^{\lambda(y,\gamma)} \tag{4.1}$$

where R is the Markov kernel for the X process alone and

$$e^{\lambda(y,\gamma)} := \int Q(y, dz)e^{\gamma z} \tag{4.2}$$

when this is finite. If it is further the case that the increments Z_t take positive values with some non-zero probability for any conditioning X_t (i.e. $Q(y, \mathbb{R}^+) > 0$ for all $y \in E$) then the prefactor simplifies to

$$\varphi(\gamma) = \sup_{x \in E} (1/v(x; \gamma)) \quad .$$

We remark than in the risk theoretic context, a larger prefactor $\sup_{x,y \in E} (v(y; \gamma)/v(x; \gamma))$ has been obtained previously [12], although the relation between the existence of a constant prefactor φ, a positive decay rate $\hat{\gamma}$, and the stability criterion $\mathbb{E}[Z_1] < 0$ seems not to have been worked out completely. An equivalent bound has been obtained for *finite* state Markov modulated Poisson processes in [1]. Translated to queueing these give the following setup. The modulating process X is a Markov chain on some finite dimensional state space E, with backward transition matrix $\{R_{xy} : x, y \in E\}$. Conditioned on the modulation $X_t = y \in E$, the individual arrivals at time t have a fixed service requirement $a(y)$ and the number of such arrivals has Poisson distribution of mean $r(y)$. The service rate is s, independent of y. In this case we find $\lambda(y, \gamma) = r(y)(e^{\gamma a(y)} - 1) - \gamma s$. $e^{\lambda(\gamma)}$ is the largest eigenvalue of the matrix $R_{xy}e^{\lambda(y,\gamma)}$. Conditioned on X_1, the expected increment in the workload is $\mathbb{E}[Z_1 \mid X_1] = \lambda'(X_1, 0) = r(X_1)a(X_1) - s$. Hence the stability condition of Lemma 2 reads $\sum_{x \in E} \rho_x \big(a(x)r(x) - s\big) < 0$, where ρ is the stationary distribution for the transition matrix R.

It is worth noting in general that in order to find the kernel for the time-reversed process (X, W), we need only find the kernel R for X in terms of its corresponding forward process.

4.2. Deterministic increments.

The following class of processes is very useful for models of multiplexers. Consider a degenerate case of Markov modulated increments in which each $Z_t = \zeta(X_t)$ for some fixed function $\zeta : E \to \mathbb{R}$. Thus we can write $Q(y, dz) = \delta_{\zeta(y)}(dz)$, the Dirac measure on $\mathbb{R}$ with support $\zeta(y)$. Applying equations (4.1) and (4.2) in this case gives

$$\hat{P}(x, dy; \gamma) = R(x, dy)e^{\gamma\zeta(y)}$$

and for the prefactor

$$\varphi(\gamma) = \sup_{x \in E:\zeta(x)>0} (1/v(x;\gamma)) \quad . \tag{4.3}$$

(This time we do *not* consider the case that the increments Z_t have non-zero probability of being positive for all X_t, for then $Z_t > 0$ with probability 1 and hence $\mathbb{E}[Z_t] > 0$, violating the stability condition.) We consider a more specific application of this result in section 4.5.

4.3. Superposed Markov Additive Processes.

Consider the case that the workload process is compose of the sum of workloads of a finite set $((X^{(\ell)}, W^{(\ell)}))^{\ell=1,\ldots,L}$ of independent MAP's, each $X_t^{(\ell)}$ taking values in some space $E^{(\ell)}$. The product process $(\bar{X}, \bar{W})$, with $\bar{X} = (X^{(1)}, \ldots, X^{(L)}) \in \bar{E} = \times_{\ell=1}^{L} E^{(\ell)}$ and $\bar{W} = \sum_{\ell=1}^{L} W^{(\ell)}$, is clearly also an MAP. Denoting by $\hat{P}^{(\ell)}(\gamma)$ the transformed kernel for $(X^{(\ell)}, W^{(\ell)})$, then the transformed kernel for $(\bar{X}, \bar{W})$ is the Kronecker product

$$\hat{P}(\gamma) = \otimes_{\ell=1}^{L} \hat{P}^{(\ell)}(\gamma) \quad ,$$

i.e. $\hat{P}(\gamma)$ is the tensor product of L copies $\hat{P}^{(\ell)}(\gamma)$, acting as a kernel for functions on the L-fold topological product $\bar{E}$. It follows that that the maximal eigenvalues $e^{\bar{\lambda}(\gamma)}$ and the corresponding eigenfunctions $\bar{v}(\bar{x};\gamma)$ of $\hat{P}(\gamma)$ are products:

$$e^{\bar{\lambda}(\gamma)} = \prod_{\ell=1}^{L} e^{\lambda^{(\ell)}(\gamma)} \qquad \text{and} \qquad \bar{v}(\bar{x};\gamma) = \prod_{\ell=1}^{L} v^{(\ell)}(x^{(\ell)};\gamma) \ , \tag{4.4}$$

where $\bar{x} = (x^{(1)}, \ldots, x^{(L)}) \in \bar{E}$, and $e^{\lambda^{(\ell)}}$ and $v^{(\ell)}$ are the eigenvalues and eigenfunctions for $\hat{P}^{(\ell)}$. Thus when the stability condition $\sum_{\ell=1}^{L} \mathbb{E}[Z_1^{(\ell)}] < 0$ is satisfied, $\hat{\gamma}$ is determined as the unique positive solution of

$$\sum_{\ell=1}^{L} \lambda^{(\ell)}(\gamma) = 0 \quad .$$

4.4. Superpositions with uniform service rates.

In applications to multiplexers we will want to consider the case that the total workload due to L independent arrival processes $(A^{(\ell)})^{\ell=1,\ldots,L}$ is serviced at a constant service rate s (i.e. the total service function is $\bar{S}_t = st$). Thus we will have Markov processes $X^{(\ell)}$ such that each $(X^{(\ell)}, A^{(\ell)})$ is an MAP. Consider again the product process $\bar{X} = (X^{(1)}, \ldots, X^{(L)})$ with summed workload process $\bar{W}$ given by

$$\bar{W} = \sum_{\ell=1}^{L} A_t^{(\ell)} - st \quad .$$

Let $c^{\ell}(\gamma) = \lim_{t\to\infty} t^{-1} \log \mathbb{E}[e^{\gamma A_t^{(\ell)}}]$ be the cumulant for $A^{(\ell)}$. Then the identification through Lemma 1(1) of the (exponential of the) cumulant of $\bar{W}$ with the maximal eigenvalue of $\hat{P}(\gamma)$ enables us to write the cumulant for $\bar{W}$ as $\bar{\lambda}(\gamma) = \sum_{\ell=1}^{L} c^{(\ell)}(\gamma) - \gamma s$. Hence when the stability condition $\sum_{\ell=1}^{L} \mathbb{E}[A_1^{(\ell)}] < s$ is satisfied, $\hat{\gamma}$ is the unique positive solution of

$$\sum_{\ell=1}^{L} c^{(\ell)}(\gamma) - s\gamma = 0 \quad . \tag{4.5}$$

The quantities $s^{(\ell)} := c^{(\ell)}(\hat{\gamma})/\hat{\gamma}$ are called the *effective bandwidths* of source ℓ. That is, $s^{(\ell)}$ is the service capacity which must be allocated to the source ℓ in order that, asymptotically as $b \to \infty$, $\mathbb{P}[Q \geq b] \leq e^{-\hat{\gamma} b}$. (See [14,16] for general discussion of effective bandwidths).

But the upper bounds hold not just in the asymptotic case $b \to \infty$. We can apply the formalism of the previous subsection, by notionally assigning to each source ℓ a service rate $s^{(\ell)}$, and setting $W_t^{(\ell)} = A_t^{(\ell)} - s^{(\ell)} t$, noting that $\sum_{\ell=1}^{L} s^{(\ell)} = s$. (Any other choice of positive service rates summing to s could be made; this one is distinguished by the fact that $\hat{\gamma}$ solves $\lambda^{(\ell)}(\gamma) = c^{(\ell)}(\gamma) - s^{(\ell)}\gamma = 0$ independently for all ℓ. See [8]). The prefactor for the superposition $\bar{\varphi}$ is then obtained using the product from of the eigenfunctions (4.4) in (3.2).

4.5. Some homogeneous deterministic multiplexer models.

In this section we consider a class of models in the intersection of those treated in sections 4.2 and 4.4. We perform the construction of section 4.4 in the special case that each of L MAP's in the L-fold superposition is an independent identical copy of some MAP (X, A), the superposition being serviced at a constant rate s. Then from eq (4.5), $\hat{\gamma}$ is the solution of $c(\gamma) = s/L$, where c is the cumulant of A. i.e. $\hat{\gamma}$ is determined from the solution of the equation $\lambda(\gamma) = 0$ for a queue with a

single source (X, A) and service rate s/L. This much is by now well known through examples. But we can now also express the prefactor $\bar{\varphi}$ of the superposition in terms of the single source MAP (X, W) where $W_t = A_t - ts/L$. For simplicity we make the following assumption (which should suffice in many multiplexer models of interest):

> the state space E of X can be written as a disjoint union $E = E_0 \cup E_1$; the increments Y of A are deterministic in the sense that $Y_t := A_t - A_{t-1} = \eta(x_t)$ for some function $\eta : E \mapsto \{0, 1\}$ with the property that
>
> $$\eta(x) = \begin{cases} 0 & \text{if } x \in E_0 \\ 1 & \text{if } x \in E_1 \end{cases} .$$

Note that with this specification of Y we will consider only the case $L > s$: otherwise the service rate exceeds the largest possible increment of the arrivals and the (stationary) queue is always empty.

The transformed kernel $\hat{P}$ for the single source MAP (X, W) is $\hat{P}(x, dy; \gamma) = \hat{R}(x, dy)e^{\gamma(\eta(y)-\sigma)}$ where $\hat{R}$ is the transition kernel for X, and we have written $\sigma := s/L$. $e^{\lambda(\gamma)}$ is the maximal eigenvalue of $\hat{P}(\gamma)$. Let $v(\cdot; \gamma)$ denote the corresponding eigenfunction, normalized so that $\mathbb{E}[v(X_0; \gamma)] = 1$, and set

$$v_0(\gamma) = \min_{x \in E_0} v(x; \gamma) \qquad \text{and} \qquad v_1(\gamma) = \min_{x \in E_1} v(x; \gamma) \quad .$$

For a configuration $\tilde{x} = (x^{(1)}, \ldots, x^{(L)})$ in the product space $\tilde{E} = E^{\times L}$ let $n(\tilde{x}) = \#\{\ell \ : \ x^{(\ell)} \in E_1\}$ be the number of sources which are on. (So $Z_t > 0$ translates to $n(\tilde{X}_t) > s$). Using the product eigenfunctions (4.4) in (3.2) then we find that the prefactor for the superposition is

$$\begin{aligned} \bar{\varphi}(\gamma) &= \sup_{\tilde{x}: n(\tilde{x}) > s} \prod_{\ell=1}^{L} \frac{1}{v(x^{(\ell)}; \gamma)} \\ &= \sup_{\tilde{x}: n(\tilde{x}) > s} \frac{1}{v_1(\gamma)^{n(\tilde{x})} v_0(\gamma)^{L - n(\tilde{x})}} \\ &\leq \begin{cases} v_1(\gamma)^{-L} & \text{if } v_0(\gamma) > v_1(\gamma) \\ v_1(\gamma)^{-s} v_0(\gamma)^{s-L} & \text{if } v_0(\gamma) \leq v_1(\gamma) \end{cases} \end{aligned}$$

Note that in either case in the last inequality, if one considers a sequence of L-fold superpositions of the same individual MAP, then increasing L whilst fixing $\sigma = s/L < 1$ (thus maintaining a constant load across all values of L), the upper bounds scale as $\mathbb{P}[Q \geq b] \leq (\Phi_{\sigma,\gamma})^L e^{-\gamma b}$ for

$$\Phi_{\sigma,\gamma} = \begin{cases} v_1(\gamma)^{-1} & \text{if } v_0(\gamma) > v_1(\gamma) \\ v_1(\gamma)^{-\sigma} v_0(\gamma)^{\sigma-1} & \text{if } v_0(\gamma) \leq v_1(\gamma) \end{cases} \tag{4.6}$$

One special case of this class of models has been worked out completely already by the methods of which those in this paper are a generalization: the case of superposed On-Off Markov arrivals. We summarize the result from [7].

For the individual sources in the superposition we take $E = \{0, 1\}$ (these states corresponding to silence and activity respectively), with $\eta(0) = 0$ (no arrival) and $\eta(1) = 1$ (a single arrival). The matrix for transition in X between silence and activity is

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

This matrix has stationary distribution $\rho = (a+d)^{-1}(d, a)$, and one shows that X is reversible: $\rho_x R_{xy} = \rho_y R_{yx}$ for all $x, y \in E$. The stability condition of Lemma 2 reads $a/(a+d) < \sigma$.

$e^{\lambda(\gamma)}$ is the maximal eigenvalue of the matrix

$$\hat{P}(\gamma) = R\, e^{\gamma(\eta(\cdot)-\sigma)} = e^{-\sigma\gamma} \begin{pmatrix} 1-a & ae^{\gamma} \\ d & (1-d)e^{\gamma} \end{pmatrix} \quad .$$

The definition of $\hat{\gamma}$ as the solution of $\lambda(\gamma) = 0$ gives $\hat{\gamma}$ as the solution of the implicit equation

$$1 = \tfrac{1}{2} e^{-\gamma\sigma} \left(e^{\gamma}(1-d) + 1 - a + \sqrt{\left(e^{\gamma}(1-d) + 1 - a\right)^2 - 4e^{\gamma}\left(1 - a - d\right)} \right) \quad .$$

We parametrize the eigenvector $v(\cdot;\gamma)$ by setting $e^{\mu(\gamma)} = v(1;\gamma)/v(0;\gamma)$ so that with normalization we have

$$v(\cdot;\gamma) = \frac{a+d}{ae^{\mu(\gamma)} + d}\left(1, e^{\mu(\gamma)}\right) \quad .$$

The dependence of μ on γ follows from the eigenvector equation giving

$$e^{\mu(\gamma)} = a^{-1}e^{-\gamma}\left(e^{\lambda(\gamma)+\sigma\gamma} + a - 1\right) \quad .$$

In Proposition 3 of [7] it is shown that

$$\mu(\gamma) > 0 \ \text{ if and only if } \ a + d < 1 \quad .$$

(This corresponding statement with inequalities replaced by equalities also holds.) In this case $v(1;\gamma)/v(0;\gamma) > 1$ so that the constant $\Phi_{\sigma,\gamma}$ in eq. (4.6) is

$$\Phi_{\sigma,\gamma} = \frac{ae^{\mu(\gamma)} + d}{(a+d)e^{\sigma\mu(\gamma)}} \quad .$$

In Theorem 2 of [7] it is furthermore shown that

$$\Phi_{\sigma,\gamma} < 1 \quad \text{for all} \quad \gamma \in (0, \hat{\gamma}] \quad .$$

The condition $a + d < 1$, which we have just seen is required to give a prefactor less than 1, can be seen as a "burstiness" condition: the correlation

$$\mathbb{E}[Y_{t+1}Y_t] - \mathbb{E}[Y_{t+1}Y_t] = (1 - a - d)ad/(a + d)^2$$

is positive iff $a + d < 1$.

Actually, $\min_{\gamma \in [0,\hat{\gamma}]} \Phi_{\sigma,\gamma}$ is not attained at $\gamma = \hat{\gamma}$ but rather at some $\gamma_{\min} < \hat{\gamma}$. Furthermore, the corresponding prefactor $(\Phi_{\sigma,\gamma_{\min}})^L$ is equal to the large deviation upper bound obtained by Hui [13] for the probability of overflow for a bufferless queue with the same arrival process. (i.e. our model with $b = 0$). This demonstrates that the value of γ for which one obtains the least upper bound on $\mathbb{P}[Q \geq b]$ depends on b, but lies in the interval $[\gamma_{\min}, \hat{\gamma}]$. Clearly the optimal choice of γ approaches $\hat{\gamma}$ as $b \to \infty$. We note also that the bound obtained using $\gamma_{\min}$ and $\mu_{\min} := \mu(\gamma_{\min})$ instead of $\hat{\gamma}$ and $\mu(\hat{\gamma})$ can be written *explicitly* in terms of the parameters a, d and σ with

$$e^{\mu_{\min}} = \frac{d\sigma}{a(1-\sigma)} \qquad \text{and} \qquad e^{\gamma_{\min}} = \frac{e^{\mu_{\min}}\left((1-a) + ae^{\mu_{\min}}\right)}{\left((1-d) + de^{\mu_{\min}}\right)} \quad .$$

This bound turns out to be extremely close numerically to those using $\hat{\gamma}$ in the case of extreme burstiness where $a + d$ is very close to 0.

4.6. Further bounds for heterogeneous superpositions.

Finally, we give another upper bound for heterogeneous superpositions of differing groups of homogeneous superpositions of MAP's. Let the superposition be of $L = \sum_{i=1}^{I} L_i$ independent sources comprising L_i MAP's of type i, the total superposition of the I groups being serviced at constant rate s. Then from section 4.4 it follows that $\hat{\gamma}$ is the solution of $\sum_i L_i c_i(\gamma) - s\gamma = 0$ where c_i is the cumulant for the arrival process of a single MAP of type i.

Let $s_i = L_i c(\hat{\gamma})/\hat{\gamma}$ so that $\sum_i s_i = s$. Denote by $A_i = (A_{i,t})_{t \geq 0}$ the summed arrivals of all MAP's of type i. Consider a MAP with workload $W_{i,t} = A_{i,t} - s_i t$. The queue length for this process is $Q_i = \sup_{t \geq 0} W_{i,t}$. Let $\bar{\varphi}_i$ denote the corresponding prefactor from section 4.4. Denoting by Q the queue length for the total superposition over all groups then clearly $Q \leq \sum_i Q_i$.

By Chebychev's inequality then for $\gamma < \hat{\gamma}$,

$$\begin{aligned}
\mathbb{P}[Q \geq b] &\leq e^{-\gamma b}\mathbb{E}[e^{\gamma Q}] \\
&\leq e^{-\gamma b}\prod_{i=1}^{I}\mathbb{E}[e^{\gamma Q_i}] \qquad \text{(independent sources)} \\
&\leq e^{-\gamma b}\prod_{i=1}^{I}\left(1+\gamma\int_0^\infty db\; e^{\gamma b}\mathbb{P}[Q_i > b]\right) \\
&\leq e^{-\gamma b}\prod_{i=1}^{I}(1+\gamma\tilde{\varphi}_i(\gamma)/(\hat{\gamma}-\gamma)) \quad .
\end{aligned}$$

Rather than minimize this expression over γ, we look for the dominant behaviour for large b which occurs by making γ close to $\hat{\gamma}$. Thus we are led to find the value of γ which minimizes $e^{-\gamma b}/(\hat{\gamma}-\gamma)^I$. This turns out to be $\gamma = \hat{\gamma} - I/b$ for b sufficiently large, yielding finally

$$\mathbb{P}[Q \geq b] \leq e^{-\hat{\gamma}b}e^I\prod_{i=1}^{I}(1+(b\hat{\gamma}/I-1)\tilde{\varphi}_i(\hat{\gamma})) \quad .$$

A better but more complex bound is obtained by an extension of the present methods for superposed Markovian On-Off sources in [8]. The method employed there also generalizes to the present case.

Acknowledgements.

Useful conversations with D.D. Botvich, E. Buffet, F.P. Kelly, G. Kesidis and J.T. Lewis are acknowledged.

References.

[1] S. Asmussen, Risk theory in a markovian environment, *Scand. Actuarial J.* (1989) 66-100

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

[3] F. Baccelli & A.M. Makowski, Dynamic, transient and stationary behavior of the M/GI/1 queue via martingales, *Ann. Prob.* 17(1989) 1691-1699

[4] F. Baccelli & A.M. Makowski, Martingale relations for the M/GI/1 queue with Markov modulated Poisson input, *Stochastic-Process. Appl.* **38**(1991) 99-133

[5] T. Björk & J. Grandell, Exponential inequalities for ruin probabilities in the Cox case., *Scand. Actuarial J.* (1988) 77-111

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

[7] E. Buffet & N.G. Duffield, Exponential upper bounds via martingales for multiplexers with Markovian arrivals, *J. Appl. Prob.* (1994) to appear.

[8] N.G. Duffield, Rigorous bounds for queue lengths in heterogeneous ATM multiplexers, Dublin preprint DIAS-STP-92-31

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

[10] A.I. Elwalid, D. Mitra & T.E. Stern, 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) 1991

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

[12] J. Grandell, *Aspects of Risk Theory*, Springer Series in Statistics, Springer, New York (1991)

[13] J.Y. Hui, Resource allocation for broadband networks, *IEEE J. Selected Areas in Commun.* **SAC-6** (1988) 1598-1608

[14] J.Y. Hui, Switching and traffic theory for integrated broadband networks, Kluwer. Boston, 1990.

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

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

[17] G. Kesidis, J. Walrand & C.S. Chang, Effective bandwidths for multiclass Markov fluids and other ATM sources, Preprint, 1992

[18] J.F.C. Kingman, A martingale inequality in the theory of queues, *Proc. Camb. Phil. Soc.* **59**(1964) 359-361

[19] L. Kosten, 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, 1988

[20] A. Martin-Löf, Entropy, a useful concept in risk-theory, *Scand. Actuarial J.* (1986)223-235

[21] J. Neveu, *Discrete parameter martingales*, North-Holland, Amsterdam (1975).

[22] I. Norros, J.W. Roberts, A. Simonain & J. Virtamo, The superposition of variable bitrate sources in ATM multiplexers, *IEEE J. Selected Areas in Commun.* **9** (1991) 378-387

[23] J.M. Reinhard, On a class of semi-Markov risks models obtained as classical risk models in a Markovian environment, *Astin Bulletin* **XIV** (1984)23-43

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

[25] G. de Veciana, C. Olivier & J. Walrand, Large Deviations for Birth Death Markov Fluids, Preprint, 1992

[26] A. Weiss, A new technique for analyzing large traffic systems, *Adv. Appl. Prob.* **18** (1986) 506-532

[27] W. Whitt, Tail probabilities with statistical multiplexing and effective bandwidths in multi-class queues, Preprint, 1992