Simple bounds for queues fed by Markovian sources: a tool for performance evaluation

Brian McGurk[1], Raymond Russell[1]

May 29, 1997

Abstract

ATM traffic is complex but only simple statistical models are amenable to mathematical analysis. We discuss a class of queuing models which is wide enough to provide models which can reflect the features of real traffic, but which is simple enough to be analytically tractable, and review the bounds on the queue-length distribution that have been obtained. We use them to obtain bounds on QoS parameters and to give approximations to the effective bandwidth of such sources. We present some numerical techniques for calculating the bounds efficiently and describe an implementation of them in a computer package which can serve as a tool for qualitative investigations of performance in queuing systems.

1 Introduction

The nature of VBR and other classes of ATM traffic is complex; different classes of traffic have very different characteristics, and the impact of the traffic on the networks designed to carry it is poorly understood. On the other hand, the behaviour of queues fed by probabilistic traffic models has been examined and analysed in some detail and a lot of results have been obtained for some simple classes of model. In this paper, we consider the case of Markov Additive Processes (MAP's), a class of traffic model which is wide enough to provide models which can capture qualitatively the features of real traffic but which is simple enough to be analytically tractable. We can construct models which reflect a particular characteristic observed in ATM traffic, such as burstiness, and apply the analytic results to them; by doing this, we can develop intuition about how the characteristic in question affects queuing systems in general.

In Section 2, we review some results from the probability literature which show that, when fed by MAP's, the distribution of the queue-length has exponential tails. This can be exploited by constructing a simple bound of the form

$$\mathbb{P}[Q > b] \leq \varphi e^{-\delta b},$$

where δ is the asymptotic decay-rate of the tail of the distribution, and φ is a constant chosen to make the bound valid for all values of b. We show how this simple bound on the queue-length distribution can be used to put bounds on different Quality of Service

[1]Dublin Institute for Advanced Studies, 10 Burlington Road, Dublin 4, Ireland.

(QoS) parameters for the queue and how this leads naturally to the concept of *effective bandwidth* of a source. This is the the minimum bandwidth that must be allocated to a source in order to guarantee a given QoS requirement.

It is important to know how easily these bounds can be computed; in Section 3, we show how to calculate φ and δ for MAP's. One attraction of the method is that the complexity of the calculation is independent of the number of sources present in the traffic stream arriving at the queue. This gives a great advantage over estimates derived from a complete solution of the model queuing problem: these generally require the analysis of matrices whose dimension is proportional to the number of sources. We present some numerical techniques for evaluating the analytical expressions efficiently, with particular emphasis on the expression for δ. Finally, we illustrate these techniques by calculating the bound for a simple two-state Markov chain in Section 4 and outline how they are implemented in an interactive computer package. Although the models we analyse may not be useful as detailed models of real ATM traffic, and we certainly do not propose them as such, this package can serve as a useful tool in the qualitative evaluation of performance of queuing systems. It is also useful as a pedagogical tool, helping to illustrate some examples of simple queues and allowing the user to visualise the general behaviour of queues, thereby building valuable intuition.

2 Theoretical background

In order to develop some intuition for the behaviour of queues in the buffers of ATM switches and multiplexors, we analyse a simple situation: the buffer is of infinite size, the service capacity is a constant s per unit time and the arrivals to the buffer have a simple (Markovian) statistical nature. The arrivals of ATM cells are not independent: if a cell arrives in one tick of the clock, it is highly likely that another cell will arrive in the next tick, or after some fixed delay. For data traffic, this is because large packets from higher level protocols must be segmented, each generating a burst of cells; for voice traffic, this is due to regular digital sampling. The simplest class of traffic models which exhibit correlations is that of Markovian arrivals. These models are flexible enough to capture the general features of ATM traffic, and yet are tractable enough to allow us calculate accurate bounds quickly.

2.1 The two-state model

Buffet and Duffield [1] considered a two-state Markov model: at time T, an input line connected to a buffer can be in one of two states. One ($X_T = 1$) corresponds to the arrival of a cell in the present clock-cycle, and the other ($X_T = 0$) to no cell arrival. The bursty nature of the arrivals is captured in the dependence of the distribution of the arrivals in the present clock-cycle on what happened in the previous clock-cycle. If a cell arrived just previously, then the probability of another cell arriving is high, close to 1; if, however, no cell arrived, then the probability of a cell arrival is small. We express this dependence precisely in the transition matrix:

$$P = \begin{pmatrix} 1-a & a \\ d & 1-d \end{pmatrix}, \qquad \text{where} \quad \begin{aligned} a &= \mathbb{P}[X_T = 1 \,|\, X_{T-1} = 0] \\ d &= \mathbb{P}[X_T = 0 \,|\, X_{T-1} = 1] \end{aligned}$$

The closer a and d are to zero, the burstier the model is. Buffet and Duffield analysed the queue formed when a superposition of these arrivals at a buffer is served at a constant rate s. Using martingale techniques, they obtained a simple upper bound on the queue-length distribution:

$$\mathbb{P}[Q > b] \leq \varphi e^{-\delta b}.$$

Fig. 1 shows the typical form of the logarithm of the queue-length distribution, and the corresponding Duffield-Buffet bound.

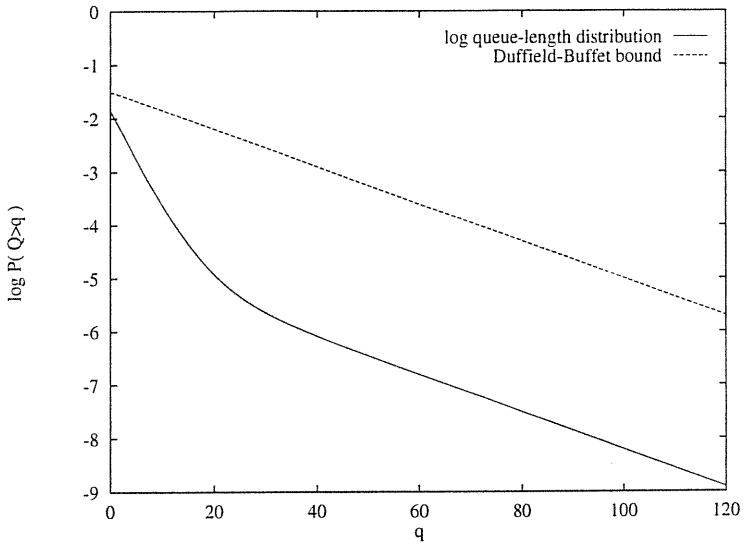

Fig. 1. The Duffield-Buffet bound

φ and δ are determined by the parameters a, d and s through a single transcendental equation. The equation is simple to solve numerically in a few iterations, yielding a fast bound on the probability of the queue exceeding any given length.

2.2 General Markov models

Duffield [2] extended these results to any queue driven by a Markov Additive Process (MAP). The workload W_T of the queue is defined to be the total arrivals up to time T less the total service available up to time T. With a MAP there is some controlling Markov process X_T and the activity of the source in time-slot T is $a(X_T)$, so

$$W_T = \sum_{t=1}^{T} a(X_t) - sT$$

Duffield again uses martingale techniques to derive an upper bound of the form

$$\mathbb{P}[Q > b] \leq \varphi e^{-\delta b}, \tag{1}$$

and shows that the decay constant δ is optimal in that it also provides an asymptotic lower bound:

$$\lim_{b \to \infty} \frac{1}{b} \log \mathbb{P}[Q > b] \geq -\delta.$$

The structure of the prefactor is important in allowing us to derive our bounds quickly: in the case of some models, if there are a large number, L, of independent and identically distributed sources feeding into the buffer, then

$$\varphi = e^{-\mu L}$$

where μ is characterised by the prefactor for a queue fed by a single source and served at rate s/L. This allows us to extend the bounds derived in the simple single source case to any number of sources without further computational effort. In most models, the prefactor for a homogeneous superposition is not exactly exponential in L but it is always true that, if L is large, φ can be well approximated by $e^{-\mu L}$ where μ is determined by the statistics of a single source served at rate s/L. Incidentally, it also illustrates the economies of scale available through statistical multiplexing: if $\mu > 0$, then adding another source, and increasing the service rate to maintain constant load, reduces the probability that the queue exceeds any buffer size by a factor of $e^{-\mu}$. See [3] and [4] for more details. We can, therefore, characterise the general bound for the queue fed by a large number L of sources,

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

by just two constants, μ and δ, where the problem of determining them is independent of the size of the system.

2.3 Queues in finite buffers

These bounds hold for queues with infinite waiting space, but the upper bounds are also useful for the finite buffer case. If we denote the queue in an infinite buffer by Q_∞, and the queue in a finite buffer of size B by Q_B, then $\mathbb{P}[Q_B > b] \leq \mathbb{P}[Q_\infty > b]$, and so we can use any upper bounds on Q_∞ for Q_B too:

$$\mathbb{P}[Q_B > b] \leq \varphi e^{-\delta b}.$$

For large buffer size B, these bounds will obviously be as good as for the infinite buffer case; for small buffers, however, they may not be tight enough. Toomey [5] has studied the problem of MAP's queuing in finite buffers, and has shown that the distribution of a queue with integer arrivals and service has the general form

$$\mathbb{P}[\text{overflow}] = c_0 e^{-\delta_0 B} + c_1 e^{-\delta_1 B} + \ldots$$

Each of the decay constants δ_i is an eigenvalue of a certain operator, and the coefficients may be determined by solving for the corresponding eigenvector. The smallest eigenvalue, δ_0, corresponds to the decay constant δ of the Duffield-Buffet formula (equation 1). This suggests a practical procedure of starting with the smallest eigenvalue, and solving for as many as are necessary to refine the bound to the desired degree.

2.4 The effective bandwidth approximation

In ATM networks, the buffer sizes are generally fixed and the service available is variable. It is natural, then, to ask questions about how much service we need to allocate to guarantee a certain quality of service. Since the size of the fixed buffer determines the maximum cell delay variation the problem is to ensure that the cell-loss ratio will be less than some target

value. We can obtain bounds on the cell-loss ratio in a finite buffer using our bounds on the queue length distribution.

The probability that a queue in a finite buffer of size B overflows is bounded by the probability that the corresponding queue in an infinite buffer exceeds length B:

$$\mathbb{P}\left[Q_B \text{ overflows}\right] \leq \mathbb{P}\left[Q_\infty > B\right] \leq \varphi e^{-\delta B}.$$

The expected number of cells lost per clock-cycle due to buffer overflow is given by

$$\mathbb{E}\left[\text{no. of cells lost}\right] = \mathbb{E}\left[\text{no. of cells arriving while } Q_B \text{ overflows}\right] \mathbb{P}\left[Q_B \text{ overflows}\right].$$

To a very good approximation, the arrivals are independent of the state of the queue, and so the expected number of cells arriving while Q_B overflows is approximately the mean activity of the sources. The cell-loss ratio is the ratio of the number of cells lost to the total number of cells arriving, or

$$\begin{aligned} \text{C.L.R.} &= \frac{\mathbb{E}\left[\text{no. of cells lost}\right]}{\mathbb{E}\left[\text{no. of cells arriving}\right]} \\ &= \frac{\mathbb{E}\left[\text{no. of cells lost}\right]}{\text{mean activity}} \end{aligned}$$

giving

$$\text{C.L.R.} \approx \mathbb{P}\left[Q_B \text{ overflows}\right] \leq \varphi e^{-\delta B}.$$

We want to try to bound the minimum service rate required to guarantee that the cell-loss ratio will be less than some acceptable target ratio.

$$\text{minimum required service} = \min\left\{ s \,:\, \text{C.L.R.}(s) \leq t \right\}$$

We can approximate this minimum by using the bound on the C.L.R.; in the case where φ is close to 1, this yields the effective bandwidth function σ

$$\begin{aligned} \sigma(t) &= \min\left\{ s \,:\, e^{-\delta B} \leq t \right\} \\ &= \min\left\{ s \,:\, \delta(s) \geq -\log(t)/B \right\} \end{aligned}$$

If φ is significantly less than 1, we improve our approximation and define the refined effective bandwidth function by

$$\sigma_{\text{ref}}(t) = \min\left\{ s \,:\, \log\varphi(s) - \delta(s)B \leq \log(t) \right\}$$

In either case, the effective bandwidth gives a conservative bound on the minimum required service.

3 Calculation techniques

3.1 Calculating δ

The queue process is completely determined by the arrivals process and the service rate, and therefore, not surprisingly, we calculate the asymptotic decay rate of the queue-length

distribution, δ, from the asymptotics of the distribution of the arrivals. First, we define the scaled cumulant generating function λ of the arrival process, as

$$\lambda(\theta) := \lim_{T\to\infty} \frac{1}{T} \log \mathbb{E}\left[e^{\theta \sum_{t=1}^{T} a(X_t)}\right].$$

$\lambda(\theta)$ is, by construction, a convex function. It is easy to verify that the slope of λ at $\theta = 0$ is the mean arrival rate, and that the asymptotic slope is the maximum achievable arrival rate. For the queue to be stable, the service rate must be greater than the mean arrivals. Furthermore, for the queue to be non-empty, the maximum arrivals must exceed the service rate. δ is found by solving for the positive root of the equation

$$\lambda(\theta) = s\theta. \tag{2}$$

Since $\lambda(\theta)$ is a convex function, the stability conditions for the queue ensure that such a root will exist (see Fig. 2) and will be unique.

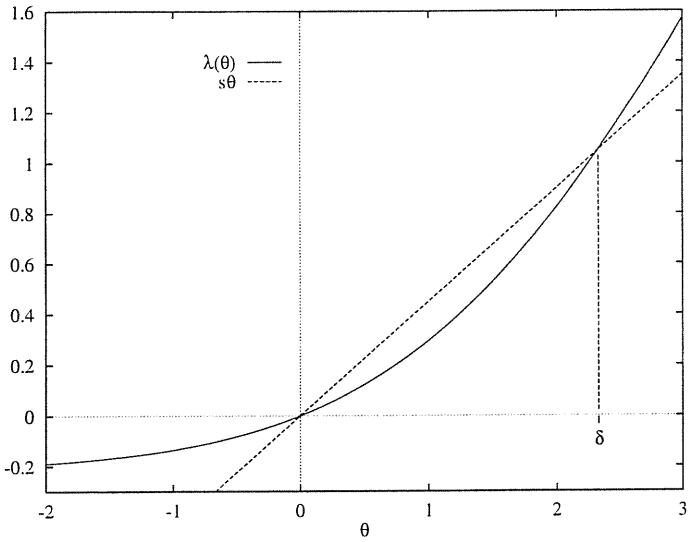

Fig. 2. The scaled cumulant generating function

We start solving this equation by examining the structure of the expectation for a finite state MAP. Let X_T be the controlling Markov chain, N be the number of states of X_T, and $a(x)$ be the increment to the arrivals at the queue when X_T is in state x. Consider $\mathbb{E}\left[e^{\theta \sum_{t=1}^{T} a(X_t)}\right]$: because of the Markovian property, we may write

$$\mathbb{E}\left[e^{\theta \sum_{t=1}^{T} a(X_t)}\right] = \sum_{x_1=1}^{N} \cdots \sum_{x_T=1}^{N} e^{\theta \sum_{t=1}^{T} a(x_t)} \prod_{t=2}^{T} \mathbb{P}\left[X_t = x_t \,|\, X_{t-1} = x_{t-1}\right] \mathbb{P}\left[X_1 = x_1\right],$$

where x_t labels the state of the chain at time t. We now pair each of the exponential factors $e^{\theta a(x_t)}$ with the corresponding transition probability by defining

$$(\tilde{P}_\theta)_{x_t x_{t-1}} := e^{\theta a(x_t)} \mathbb{P}\left[X_t = x_t \,|\, X_{t-1} = x_{t-1}\right], \qquad (\pi_\theta)_{x_1} := e^{\theta a(x_1)} \mathbb{P}\left[X_1 = x_1\right].$$

The T summations of the product of these factors is nothing other than $T-1$ matrix multiplications written out explicitly; the expectation may now be written in matrix notation as

$$\mathbb{E}\left[e^{\theta\sum_{t=1}^{T}a(X_t)}\right] = \pi_\theta\left(\tilde{P}_\theta\right)^{T-1}1^\dagger, \tag{3}$$

where $1^\dagger$ is the transpose of a vector containing 1's in each column. The matrix $\tilde{P}_\theta$ is called the *twisted transition matrix* because it is the transition matrix P twisted by the exponential factors $e^{\theta a(x_t)}$. Thus we have that λ is given by

$$\begin{aligned}\lambda(\theta) &= \lim_{T\to\infty}\frac{1}{T}\log\left[\pi_\theta\left(\tilde{P}_\theta\right)^{T-1}1^\dagger\right]\\ &= \log\rho\left(\tilde{P}_\theta\right),\end{aligned}$$

where $\rho(\tilde{P}_\theta)$ is the spectral radius of $\tilde{P}_\theta$. If X_T is ergodic (stationary, recurrent and irreducible), then $\rho(\tilde{P}_\theta)$ is the maximum modulus of the eigenvalues of $\tilde{P}_\theta$. The problem of determining δ is then the following: find the unique $\theta > 0$ such that

$$\log\rho\left(\tilde{P}_\theta\right) = s\theta$$

We have developed a number of different ways of solving this problem, outlined as follows.

The Powell method

We may take a direct approach, using techniques from linear algebra to evaluate the largest eigenvalue of $\tilde{P}_\theta$ as a function of θ. The equation $\lambda(\theta) = s\theta$ is then readily solved using a simple bisection algorithm. It turns out that one competitive method for determining the spectral radius of a matrix is the Powell method. The spectral radius $\rho(A)$ of a matrix A is defined by $\rho(A) := \sup_v |Av|/|v|$, so that after a large number n of iterations of A, $|A^n v| \approx \rho(A)^n|v|$. To find $\rho(A)$, one starts with a random initial vector, v_0 say, and forms the iterates of A applied to it:

$$v_1 := Av_0, \quad v_2 := Av_1, \quad \ldots \quad v_k := Av_{k-1} = A^k v_0.$$

$\rho(A)$ is then estimated as the ratio of the moduli of successive vectors in the sequence: $\rho(A) = |v_k|/|v_{k-1}|$. If A has other eigenvalues close in modulus to $\rho(A)$, then this estimate will only converge poorly. In this case, a better estimate is $\rho(A) = \sqrt{|v_{k+1}|/|v_{k-1}|}$. The choice of v_0 can also strongly affect the convergence of this method. For example, if v_0 is an eigenvector of A which contains no component in the direction of the eigenvector corresponding to $\rho(A)$, then the method will not converge at all. For practical purposes, a good choice of initial vector is suggested by Equation 3: the vector π_0 is the stationary measure of the controlling chain X_t, and hence the eigenvector of $\tilde{P}_0$ corresponding to eigenvalue 1. By the Perron-Frobenius theorem, 1 is the largest eigenvalue of $\tilde{P}_0$ and so, for small values of θ, π_θ will be close to the eigenvector of $\tilde{P}_\theta$ of corresponding to eigenvalue $\rho(\tilde{P}_\theta)$. The Powell method will then converge rapidly, quickly yielding a good estimate of $\lambda(\theta)$.

The determinant method

An alternative approach is to start with the eigenvalue equation for $\tilde{P}_\theta$: α is an eigenvalue of $\tilde{P}_\theta$ iff $\det(\tilde{P}_\theta - \alpha I) = 0$, where I is the identity matrix. We are looking for the value

of θ which gives $\log \rho(\tilde{P}_\theta) = s\theta$, i.e. $\rho(\tilde{P}_\theta) = e^{s\theta}$. Since $\rho(\tilde{P}_\theta)$ is an eigenvalue of $\tilde{P}_\theta$, we could also look for solutions to

$$\det(\tilde{P}_\theta - e^{s\theta} I) = 0. \tag{4}$$

In general, there will be many values of θ such that $\tilde{P}_\theta$ has eigenvalue $e^{s\theta}$, but we know that δ will be the smallest positive such θ. Since calculating determinants is a lot cheaper numerically than calculating eigenvalues, the roots of Equation 4 are a lot easier to find than the root of Equation 2. However, we have very little information about the form of $\det(\tilde{P}_\theta - e^{s\theta} I)$ as a function of θ. It is difficult to know precisely how many zeros it has, and whether or not any particular solution to Equation 4 is the smallest positive one. We can, however, test any solution θ_0 found by using a single evaluation of λ using the Powell method: $\delta = \theta_0 \quad$ iff $\quad \lambda(\theta_0) = s\theta_0$.

The root-tracking method

This method is based on the observation that the eigenvalues of $\tilde{P}_\theta$ are smooth functions of θ, and the knowledge that, for Markov chains, the eigenvalue of largest modulus at $\theta = 0$ is the eigenvalue of largest modulus for all values of θ. Let us call this eigenvalue $\alpha(\theta)$. It satisfies the eigenvalue equation

$$f(\theta; \alpha(\theta)) = 0, \qquad \text{where} \quad f(\theta; \alpha) = \det(\tilde{P}_\theta - \alpha I),$$

and is a smooth function of θ, and so

$$\frac{\partial f}{\partial \theta}(\theta; \alpha(\theta)) + \frac{\partial f}{\partial \alpha}(\theta; \alpha(\theta)) \frac{d\alpha}{d\theta} = 0.$$

Noting also that $\alpha(0) = 1$, we may calculate $\alpha(\theta)$ by solving the first order O.D.E.

$$\frac{d\alpha}{d\theta} = -\frac{\partial f}{\partial \theta}(\theta; \alpha(\theta)) / \frac{\partial f}{\partial \alpha}(\theta; \alpha(\theta))$$

starting with the initial value $\alpha(0) = 1$. The attraction of this method is that the numerical solution of O.D.E.'s is a subject which has attracted much attention. Because of this, there are a great many powerful and well-tested methods for solving them; see Press et al. [6] for an illuminating review and more references. In practice, the accuracy of the solution need not be that great; it is sufficient to track $\alpha(\theta)$ approximately as it initially decreases with increasing θ, until it exceeds $e^{s\theta}$. (Recall that $\log \alpha(\theta) = \lambda(\theta)$ and look again at Fig. 2.) The value of θ at which this occurs may then be used as an initial point in a Newton-Raphson solution to Equation 4 from the previous method. The expressions for the partial derivatives of f are, in general, quite cumbersome, and so this method is best suited to models in which the determinant in f may be explicitly evaluated.

3.2 Calculating the prefactor φ

We saw how the scaled cumulant generating function λ is the logarithm of the largest eigenvalue of the twisted transition matrix $\tilde{P}_\theta$. [2] shows that we can calculate the prefactor φ from the corresponding eigenvector of $\tilde{P}_\theta$ as follows.

Let $\mathbf{v}(s)$ be the eigenvector of $\tilde{P}_{\delta(s)}$ of eigenvalue $e^{s\delta(s)}$:

$$\mathbf{v}(s) \tilde{P}_{\delta(s)} = e^{s\delta(s)} \mathbf{v}(s).$$

This is a vector with a real component $v_i(s)$ corresponding to each state i ($1 \leq i \leq N$) of the controlling Markov chain; we take $\mathbf{v}(s)$ to be normalised so that $v_1(s)+\ldots+v_N(s) = 1$. We denote by E those states of the Markov chain such that the activity of the source while in those states exceeds the service rate s:

$$E = \{\, x_i \,:\, a(x_i) > s \,\}.$$

The prefactor φ is then simply

$$\varphi(s) = \max_{x_i \in E} \frac{1}{v_i(s)}.$$

3.3 Homogeneous superpositions

Suppose that the arrivals consist of a homogeneous superposition of sources, that is, the total arrivals in a time slot come from the sum of the activities of L independent and identically distributed Markov chains $X_T^{(1)}, \ldots X_T^{(L)}$:

$$a_{\text{total}}(X_T) = a(X_T^{(1)}) + \ldots + a(X_T^{(L)}).$$

In this case, the state space of the controlling Markov chain X_T is the product space $\{\, 1, 2, \ldots N \,\}^L$ and the transition matrix is the L-fold tensor product of the common transition matrix of each source.

Calculating δ

Consider the scaled cumulant generating function of the arrivals process. The arrivals can be written as the sum of the arrivals from each of the L sources:

$$\sum_{t=1}^{T} a(X_t) = \sum_{t=1}^{T} a(X_t^{(1)}) + \ldots + \sum_{t=1}^{T} a(X_t^{(L)}).$$

Since the L Markov chains are independent, the expectation in the scaled cumulant generating function breaks up into a product of L terms,

$$\mathbb{E}\left[e^{\theta \sum_{t=1}^{T} a(X_t)}\right] = \mathbb{E}\left[e^{\theta \sum_{t=1}^{T} a(X^{(1)}}\right] \ldots \mathbb{E}\left[e^{\theta \sum_{t=1}^{T} a(X^{(L)}}\right]$$

and, since they are identically distributed, we have

$$\mathbb{E}\left[e^{\theta \sum_{t=1}^{T} a(X_t)}\right] = \left(\mathbb{E}\left[e^{\theta \sum_{t=1}^{T} a(X^{(1)}}\right]\right)^L,$$

giving

$$\log \mathbb{E}\left[e^{\theta \sum_{t=1}^{T} a(X_t)}\right] = L \cdot \log \mathbb{E}\left[e^{\theta \sum_{t=1}^{T} a(X^{(1)}}\right].$$

Thus the scaled cumulant generating function for the total arrivals is

$$\lambda(\theta) = L\lambda^{(1)}(\theta),$$

where $\lambda^{(1)}$ is the common scaled cumulant generating function of all the sources. To find δ for the superposition, we need to solve Equation 2 which is, in this case, equivalent to solving

$$\lambda^{(1)}(\theta) = (s/L)\theta.$$

Calculating φ

Since the transition matrix has a product structure, so do the twisted transition matrix and its eigenvector $\mathbf{v}(\mathbf{s})$ of maximal eigenvalue. We can exploit this structure to obtain a prefactor φ which is itself an L-fold product:

$$\varphi = e^{-L\mu},$$

where μ can be determined from $\mathbf{v}^{(1)}$, the eigenvector (of maximal eigenvalue) of the twisted transition matrix of a single source. Details are given in [2].

Thus, to calculate our simple two-parameter bound in the case of a homogeneous superposition of L sources, we need only solve Equation 2 and calculate the corresponding eigenvector for the case of one source.

3.4 Calculating effective bandwidths

As we have seen, $\delta(s)$ is the unique positive solution of the equation

$$\lambda(\theta) = s\theta \tag{5}$$

so that,

$$\frac{\lambda(\delta(s))}{\delta(s)} = s$$

Now, the effective bandwidth function is defined by

$$\sigma(t) = \min\{\, s \,:\, \delta(s) \geq -\log(t)/B \,\}$$

where t is the highest acceptable loss-ratio. If δ_t is the fraction on the right hand side of the inequality then $\sigma(\delta_t)$ is the value of s for which

$$\delta(s) = \delta_t$$

and so,

$$\sigma(\delta_t) = \lambda(\delta_t)/\delta_t.$$

In general, the refined effective bandwidth function is difficult to evaluate explicitly since both δ and μ are functions of s; however, it is still readily calculated numerically.

4 Implementation

4.1 Calculations for a 2-State Markov Model

For the case of a 2-state Markov model, calculating μ and δ reduces to a numerically solvable transcendental equation. We need to examine the maximum modulus of the eigenvalues of the twisted transition matrix. For the 2-state model this is a simple problem. Assuming the activity in state 1 be 0 and in state 2 to be 1, and the transition probabilities to be as defined in section 2.1 , the matrix in question is

$$\tilde{P}_\theta = \begin{pmatrix} (1-a) & e^\theta a \\ d & e^\theta(1-d) \end{pmatrix} \tag{6}$$

We can easily solve the eigenvalue equation for this matrix and determine the largest eigenvalue :

$$\begin{vmatrix} (1-a)-\alpha & e^{\theta}a \\ d & e^{\theta}(1-d)-\alpha \end{vmatrix} = 0$$

$$\alpha^2 - (1-a+e^{\theta}(1-d))\alpha + e^{\theta}(1-a-d) = 0$$

Solving this equation for α, we get

$$\begin{aligned} \alpha &= \frac{1}{2}\left[1-a+(1-d)e^{\theta} \pm \sqrt{(1-a+(1-d)e^{\theta})^2 - 4(1-a-d)e^{\theta}}\right] \\ \lambda(\theta) &= \log(\alpha_{\max}) \\ &= \log\left[1-a+(1-d)e^{\theta} \pm \sqrt{(1-a+(1-d)e^{\theta})^2 - 4(1-a-d)e^{\theta}}\right] - \log 2 \end{aligned}$$

The effective bandwidth is now easily calculated by using this expression for $\lambda(\theta)$ in equation 3.4. In order to find $\delta(s)$, we must solve Equation 5 numerically. As discussed in section 3.2, $\varphi(s)$ is found from the eigenvector, $\mathbf{v}(s)$, of the maximal eigenvalue of $\tilde{P}_{\delta(s)}$. Since we are dealing with a simple on-off model, we have that

$$\varphi(s) = \frac{1}{v_2(s)}.$$

We know from equation 5 that the maximal eigenvalue of $\tilde{P}_{\delta(s)}$ is $e^{s\delta(s)}$; thus we are looking for the value of v_2 in the equation

$$(v_1 \;\; v_2)\tilde{P}_{\delta(s)} = e^{s\delta(s)}(v_1 \;\; v_2).$$

Taking the equation corresponding to the first column of the matrix, we get

$$v_2 = \frac{e^{s\delta(s)} - 1 + a}{d} v_1,$$

and, using the normalisation that v_1 and v_2 sum to 1, we find that

$$\varphi(s) = \frac{e^{s\delta(s)} + a - 1}{e^{s\delta(s)} + a + d - 1}.$$

4.2 A proposal for an interactive tutorial package

The numerical techniques for evaluating δ and φ which we outlined in Section 3, and illustrated above for the two-state model, are very efficient. We have tested them by implementing them in C on a 66MHz Intel 486 for various different MAP's and, even for moderately large state-spaces (N=100), δ and φ can be evaluated in a negligible amount of time (less that 1s). This suggested that an interactive package could be built around these routines; such a package could exploit the excellent graphical capabilities which even modest PC's possess today. We have designed such a package: it allows the user choose from a range of Markov models, allows them to specify the parameters of the model (such as mean activity, burstiness and so on) and the service rate of the queue and displays the bound on the queue-length distribution and various QoS parameters. Since the calculations are effected almost instantaneously, the user can play around with many different scenarios, allowing them to develop intuition about what the impact of the different characteristics of the traffic is on its queuing behaviour. The package is licensed for free use and is available for ftp from `ftp://ftp.stp.dias.ie/DAPG/`

5 Conclusion

In this paper, we considered the queuing behaviour of arrivals processes called MAP's which have an underlying Markov structure. We reviewed some results from the probability literature which show that, when fed by MAP's, the distribution of the queue-length has exponential tails. We exploited this by constructing a simple bound of the form

$$\mathbb{P}[Q > b] \leq \varphi e^{-\delta b},$$

where δ is the asymptotic decay-rate of the tail of the distribution, and φ is a constant chosen to make the bound valid for all values of b. We showed how this simple bound on the queue-length distribution can be used to put bounds on different Quality of Service (QoS) parameters for the queue and how the concept of effective bandwidth arises naturally.

We showed how to calculate φ and δ for MAP's and presented some numerical techniques for evaluating the analytical expressions efficiently. We illustrated these techniques in the case of a two-state Markov chain and outlined how these techniques have been used to implement an interactive tutorial package.

References

[1] E. Buffet and N.G. Duffield (1992) Exponential upper bounds via martingales for multiplexers with Markovian arrivals. *J. Appl. Prob.* **31** 1049–1061

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

[3] D.D. Botvich and N.G. Duffield Large deviations, the shape of the loss curve, and economies of scale in large multiplexers. Preprint DIAS-APG-94-12 Accepted for Queueing Systems subject to revision.

[4] D.D.Botvich, T.J. Corcoran, N.G. Duffield, P.Farrell Economies of scale in long and short buffers of large multiplexers. Proceedings of 12th UK Teletraffic Symposium.

[5] F. Toomey (1994) Queues in finite buffers with Markovian arrivals: an application to bursty traffic. Preprint.

[6] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery *Numerical Recipes in C* Cambridge University Press, Cambridge (1992)

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