Evaluating the Queue Length Distribution of an ATM Multiplexer with Multiple Time Scale Arrivals

Predrag R. Jelenković and Aurel A. Lazar
Department of Electrical Engineering
and
Center for Telecommunications Research
Columbia University, New York, NY 10027-6699
{predrag, aurel}@ctr.columbia.edu, Tel: 854-2399

Abstract

For an ATM multiplexer we develop a *recursive* asymptotic expansion method for approximating the queue length distribution and investigate the radius of convergence of the queue asymptotic expansion series. The analysis focuses on "small" to "moderate" buffer sizes under the conditions of strictly stable multiple time scale arrivals. For a class of examples we ***analytically*** determine the radius of convergence using methods of linear operator theory. We also give general sufficient conditions under which the radius converges to zero; this shows roughly what situations have to be avoided for the proposed method to work well. We combine the asymptotic expansion method with the EB approximation, and give an approximation procedure for the buffer probabilities for all buffer ranges. The procedure is tested on extensive numerical examples. We suggest this procedure for efficient admission control in ATM networks.

1 Introduction

Numerous investigations have shown that the arrival processes (sources) that arise in ATM networks (like voice and video) have a very complex statistical structure; especially troublesome characteristic is the high statistical dependency (e.g. see [13]). Modeling of this high dependency usually leads to analytically very complex statistical characteristics; the associated evaluation of the queue length distribution is typically intractable. However, because of the stringent QOS requirements in ATM, only the tail of the queue length distribution in the domain of very small probabilities is needed. This has motivated many researchers to find simple approximations of the asymptotic behavior of the queue length distribution.

More formally, let $\{A_t, t \geq 0\}$ be an integer valued, discrete time, stationary, and ergodic process (on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$), and let $c \in \mathbb{N}$ be the capacity of the server. Then, for any initial random variable Q_0, the following (Lindley's) equation

$$Q_{t+1} = (Q_t + A_t - c)^+ \qquad (1)$$

completely defines the queue length process $\{Q_t\}$. Queues of this type represent a natural model for ATM multiplexers. According to the classical result of Loynes's [16], if $\mathbb{E}A_t < c$ (and $\{A_t, t \geq 0\}$ is stationary and ergodic) $\{Q_t\}$ couples with the unique stationary solution $\{Q_t^s\}$ of the recursion (1) for any initial condition Q_0; in particular $\mathbb{P}[Q_t \geq x] \to \mathbb{P}[Q_0^s \geq x]$ as $t \to \infty$ (for simplicity we will refer to Q_t^s simply as Q).

To go beyond the existence and uniqueness of the stationary queue length distribution $\mathbb{P}[Q > x]$ one has to impose more restrictive assumptions than merely stationarity and ergodicity. Using the Theory of Large Deviations (see [19]), under the general assumptions (in addition to stationarity and ergodicity), of the Gärtner-Ellis (Cramér) type, one can show that

$$\lim_{x \to \infty} -\frac{\log \mathbb{P}[Q > x]}{x} = \theta^*, \qquad (2)$$

for some positive constant θ^*, called the ***asymptotic decay rate*** (or the equivalent bandwidth constant) [1, 7]. (For the *non Cramér/Subexponential* queueing asymptotic analysis we refer the reader to [12] and references therein.) Also, in the case of (finite) Markov modulated arrivals the following stronger result holds: $\mathbb{P}[Q > x] \sim \alpha e^{-\theta^* x}$ as $x \to \infty$, where α and θ^* are positive constants. Some numerical calculations for simple arrival processes (like On-Off Markov sources) have shown that the constant α is "usually" of the order one. This led many authors to believe that the simple approximation $\mathbb{P}[Q > x] \approx e^{-\theta^* x}$ holds; this approximation is commonly referred as [2] the effective bandwidth (EB) approximation. Based on this result admission control policies based on the concept of effective bandwidth have been developed; see [1, 6, 8, 7, 14].

However, as discussed in [2], the EB approximation may often be very inaccurate. This is usually the situation when many sources (N) are multiplexed; in this case it was shown that $\alpha \approx e^{-\gamma N}$ for some constant γ. More formal analysis for the multiplexing of the large number of sources is given in [5].

In [10, 11] we have investigated the impact of mul-

0743-166X/96 $5.00 © 1996 IEEE

tiple time scales on the queue length distribution. The main motivation for the work done in [11] as well as in the present paper is that many traffic sources, such as variable bit rate video (VBR), have a multiple time scale structure [15] that spans from a few ns to few hours (one hour = 3.6 $10^{12}ns$). The impact of the multiple time scales on the mutiplexer performance has been independently reported in [18].

This paper is the natural continuation of the work done in [11]. Therefore, for reasons of completeness, in section 2 we give a short summary of the main results and issues presented in [11]. The rest of the paper is organized as follows.

In section 3 we present a *recursive* asymptotic expansion method for approximating of the queue length distribution for "small" to "moderate" buffer ranges under the condition of strictly stable multiple time scale arrivals. In section 4 we discuss the radius of convergence of the queue asymptotic expansion series. Using linear operator theory methods (for a class of examples) we *analytically* determine the radius of convergence. We also give general sufficient conditions under which the radius converges to zero. This result directs our numerical investigations for deriving further conclusions about the qualitative behavior of the radius of convergence. Section 5 discusses the error of asymptotic expansion approximation, and gives a simple criterion for its evaluation. Finally, by combining the asymptotic expansion method with the EB approximation, we propose an approximation for the buffer probabilities for all buffer ranges. This approximation is tested in section 6 on extensive numerical examples. The paper is concluded in section 7.

2 Summary of Previous Work

In this section we give a short summary of the main results obtained in [11] where equation (2) was extended to the case of arrival processes that converge in a cumulant sense. Based on this result, the *asymptotic independence of the EB constant on the slow time scale statistics* was shown. (For the precise formulation of these results the reader is referred to [11].) The EB constant turned out to be the same as the one for which traffic sources are continuously producing traffic at their "peak" rate. The immediate implication of this result is that an equivalent bandwidth based admission control (which solely depends on the EB constant) may significantly underutilize the system resources.

In the same paper ([11]) numerous numerical examples are given that illustrate the "polygonal" behavior of the queue length distribution on the logarithmic scale. The polygonal behavior is due to the multiple time scale nature of the arrival processes (for illustration see solid lines in Figure 1). Intuitively, different time scales are responsible for building up various regions of the buffer. As the different time scales of the arrival processes start to mix, they result in different buffer decrease rates for different buffer sizes. Eventually on the largest time scale the decrease rate becomes equal to the EB constant. Also the actual distribution obtained by statistical multiplexing of the six different parts of the Star Wars video on the slice level is shown in Figure 2. One can *clearly see that the queue length distribution can not be well approximated with a single line (exponential)!*

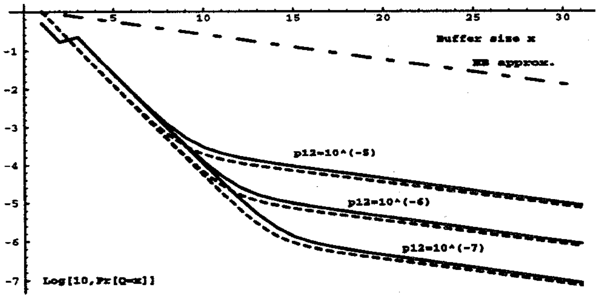

Figure 1: Graph of $\log_{10} \mathbb{P}[Q^\epsilon = x]$; for details see [JLA95b], Example 1.

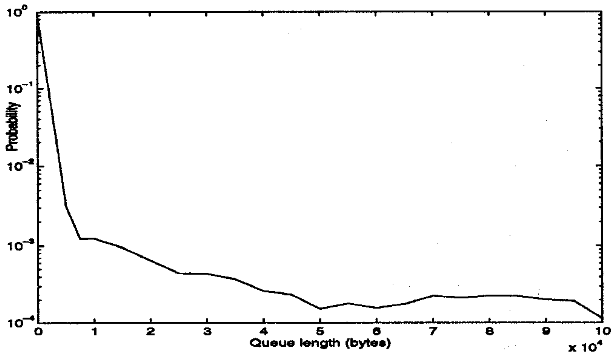

Figure 2: Queue length distribution for multiplexing 6 parts of the Star Wars video sequence on the slice level. The total length of the multiplexed sequence is 1000,000 slices (≈ 23 min).

In what follows, we present a general *superposition theorem* for ATM multiplexers ([11]). This theorem sets the stage for the work presented in this paper. We consider an arrival process of the form $A_t^\epsilon \overset{def}{=} X_t(B_t^\epsilon)$. Assume that this process is Markov modulated, i.e., for each $\epsilon > 0$, the process B^ϵ is an irreducible aperiodic Markov Chain with state space $\{1, \ldots, K\}$, and transition probabilities $\{p_{jk}^\epsilon = O(\epsilon), j \neq k, 1 \leq j, k \leq K\}$, such that the (unique) stationary distribution $\{\pi_j^\epsilon\} \to \{\pi_j^0\}, 1 \leq j \leq K$, as $\epsilon \to 0$ (Note that some of π_i^0 may be equal to zero). The conditional processes $X(\cdot)$ are assumed to be stationary, ergodic processes independent of B_t^ϵ. We call this class of processes *Nearly*

Decomposable Markov Modulated Stationary Processes (NDMMSP). Let $\mathbb{P}_j[Q^\epsilon > x] \overset{def}{=} \lim_{t\to\infty} \mathbb{P}[Q_t^\epsilon > x, B_t^\epsilon = j]$ (note that we assume that this limit exists). Then the following theorem holds.

Theorem 1 *Assume that $X(j), 1 \le j \le K$, are stationary, and ergodic, and that the strict stability condition $\mathbb{E}X(j) < c, 1 \le j \le K$ is satisfied. Then for any x*

$$\lim_{\epsilon\to 0} \mathbb{P}_j[Q^\epsilon > x] = \pi_j^0 \mathbb{P}[Q(j) > x], \qquad (3)$$

where $\mathbb{P}[Q(j) > x]$ represents the queue length distribution given that the arrival process is $X(j)$.

Proof: Given in [10].

Remarks: **(i)** This theorem shows that for a small ϵ, $\mathbb{P}[Q^\epsilon > x] \approx \sum_{j=1}^K \pi_j^0 \mathbb{P}[Q(j) > x]$. Borrowing from linear system theory language, we call this relationship a *superposition principle*. **(ii)** For small buffer sizes, the queue length distribution does not depend on the long term dependency structure.

Remark (ii) explains why in [15], due to the stringent time delay constraints (small buffer), it has been concluded that the dominant effect on the queue performance was the short term autocorrelation, and that the long term autocorrelation is negligible. This effect was also noted in [17], where nearly decomposable Markov Modulated Poisson processes are used to model video traffic (in our model this is obtained by specializing $X(j)$ to be Poisson).

As suggested in remark **(i)** we can approximate the queue length distribution with: **a)** $\mathbb{P}[Q^\epsilon > x] \approx \sum_{j=1}^K \pi_j^0 \mathbb{P}[Q(j) > x]$, or the even simpler approximation **b)** $\mathbb{P}[Q^\epsilon > x] \approx \sum_{j=1}^K \pi_j^0 e^{-\theta^*(j)}$, where $e^{-\theta^*(j)}$ is the EB approximation for $\mathbb{P}[Q(j) > x]$. The latter approximation is exemplified in Figure 1 (graphs with dashed lines). However, this approximation may give poor results for the moderate or large buffer sizes, see Figure 3; the dashed line represents the approximation and the solid line the true probabilities. Note that the error has a tendency to increase with the buffer size.

If we consider $\mathbb{P}[Q^\epsilon = x]$ as a function of ϵ then Theorem 1 gives us the first coefficient in the asymptotic expansion of $\mathbb{P}[Q^\epsilon = x]$ with respect to epsilon. Naturally, in order to get a better approximation than the one given by Theorem 1 we have to obtain high order expansion coefficients of this asymptotic expansion. This will be explored in the rest of the paper.

3 Asymptotic Expansion

In this section we will investigate the problem of expanding the probabilities $\mathbb{P}[Q^\epsilon = x]$, under the strict stability conditions, into a series with respect to ϵ. To do this we need to assume more structured arrivals to the queue (than NDMMSPs that we used in the previous section).

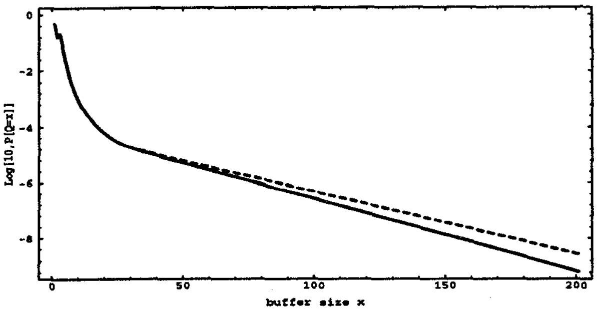

Figure 3: Graph of $\log_{10} \mathbb{P}[Q^\epsilon = x]$ from Example 2.

Assume that the arrival process is Nearly Decomposable Markov Modulated I.I.D. (NDMMIID), i.e., the modulating process B_t^ϵ is Markovian with decomposed state space $\mathcal{S} = \cup_{k=1}^K \mathcal{S}_k$, and transition matrix $\{p_{ij}^\epsilon\}, i, j \in \mathcal{S}, \epsilon \in (0,1)$. It is assumed that p_{ij}^ϵ are infinitely differentiable with respect to ϵ, and $\lim_{\epsilon\to 0} p_{ij}^\epsilon = 0$ for $i \in \mathcal{S}_k, j \in \mathcal{S}_l, k \ne l$, and $\lim_{\epsilon\to 0} p_{ij}^\epsilon = p_{ij}^0$ for $k = l$. Each matrix $P_k^0 = \{p_{ij}^0\}_{i,j\in\mathcal{S}_k}$ is assumed to be irreducible. Let π^ϵ be the steady state distribution of the chain B^ϵ, and $\pi^0 = \lim_{\epsilon\to 0} \pi^\epsilon$; we assume that each $\pi^0(\mathcal{S}_k) > 0$. Furthermore, when in state j, the source is generating I.I.D. arrivals with the moment generating function $a_j(z) = \sum_{k=0}^\infty a_{j,k} z^k, |z| \le 1$. Assume that $a_j(z), j \in \mathcal{S}$ satisfy Cramér conditions, i.e., there exists $\delta \in \mathbb{R}, \delta > 1$ such that $a_j(\delta) < \infty$ for all $j \in \mathcal{S}$. Define $A(\epsilon, z) = \{p_{ij}^\epsilon a_j(z)\}_{i,j\in\mathcal{S}}^T$, and $Q(\epsilon, z) = (Q_1(\epsilon, z), \ldots, Q_K(\epsilon, z))^T$, where $Q_j(\epsilon, z) = \sum_{k=0}^\infty q_{j,k}^\epsilon z^k$, $q_{j,k}^\epsilon = \mathbb{P}[Q = k, B^\epsilon = j]$, and $(\cdot)^T$ stands for transposition. In practice the recursion $Q_{t+1} = (Q_t - c)^+ + A_t$ is often used instead of (1) because of its simpler boundary condition. For that reason we will also use it in all our numerical examples. Other than that all the results, including the next theorem (in the appropriate rephrased form), are valid for recursion (1) as well. Using the classical z-transform technique, the solution for $Q(\epsilon, z)$ is given by

$$Q(\epsilon, z) = [Iz^c - A(\epsilon, z)]^{-1} A(\epsilon, z) \Pi(\epsilon), \qquad (4)$$

where $\Pi(\epsilon) = \{\Pi_j^\epsilon\}_{1\le j\le N}^T$, and $\Pi_j^\epsilon = \sum_{k=0}^{c-1} q_{jk}^\epsilon (z^c - z^k)$, $q_{jk} \overset{def}{=} \lim_{t\to\infty} \mathbb{P}[Q_t = k, B_t = j]$. Because of the smoothness assumption on p_{ij}^ϵ it follows that $A(\epsilon, z) = \sum_{k=0}^\infty A^k \epsilon^k$, $Q(\epsilon, z) = \sum_{k=0}^\infty Q^k \epsilon^k$, and $\Pi(\epsilon) = \sum_{k=0}^\infty \Pi^k \epsilon^k$. Then, the following result holds for the expansion coefficients Q^k.

Theorem 2 *The Taylor expansion coefficients of $Q(\epsilon, z)$ with respect to ϵ are recursively given by: $Q^0 =$*

$[Iz^c - A^0]^{-1}A^0\Pi^0$, *and*

$$Q^k = [Iz^c - A^0]^{-1}\left\{A^0\Pi^k + \sum_{l=0}^{k-1} A^{k-l}(Q^l + \Pi^l)\right\}, \quad (5)$$

$k \geq 1$. *Boundary vectors* $\Pi^k, k \geq 0$, *are uniquely obtained from the analyticity of the functions* $Q^k(z)$ *in the unit circle, and the equation* $Q^k(1) = \lim_{\epsilon \to 0}(\frac{\partial}{\partial \epsilon})^k \pi^\epsilon$.

Proof: Given in [9].

Note: Nearly decomposable Markov queueing processes with a *finite* state space have been previously considered in the literature (see [3]). However, we are unaware of any work which considers an *infinite state space*, or a moment generating function approach, and the determination of the boundary conditions.

It is important to observe that for the exact solution of equation (4) one has to solve for all the roots in the unit circle of the polynomial $\det[Iz^c - A(\epsilon, z)]$. Usually for the case of statistical multiplexing this polynomial is of a very large degree, and the number of roots is very large, i.e, often impossible to solve. *However, if there is enough structure, the matrix* A^0 *may be of a diagonal or block diagonal form, and the polynomial* $\det[Iz^c - A^0]$ *may be decomposed into a product, which is incomparably easier to solve. Also, the large number of roots, (if not all, see equation (6) and section 6) may vanish.*

4 Radius of Convergence

In this subsection we attempt to find the radius of convergence of the queueing series $\sum_{k=0}^{\infty} Q^k \epsilon^k$. In general this is a very difficult problem. However, under more restrictive assumptions than in the previous section, we will be able to give some answers to this problem.

More precisely, we define the radius of convergence of the queueing series as $r = \sup\{\epsilon : |\sum_{k=0}^{\infty} Q^k \epsilon^k| < \infty, \; |z| \leq 1\}$. We assume that the modulating chain is totally decomposable, i.e., each $\mathcal{S}_k$ is just one point. This implies that the matrix A^0 is diagonal with diagonal elements $\{a_j(z)\}$ and $A(\epsilon, z) = A^0(z)P(\epsilon)^T$, $P(\epsilon) = \{p_{ij}^\epsilon\}$. We also assume that $P(\epsilon)$ has a finite asymptotic expansion, i.e., $P(\epsilon) = \sum_{k=0}^{n} P^k \epsilon^k$, that π^ϵ is a constant vector independent of ϵ, and $c = 1$. Then vectors $Q^k, k \geq 1$, have the form $Q^k(z) = (z-1)q^k(z)$. Further, $Q^0 = (z-1)[Iz - A^0]^{-1}A^0[I - A^{0'}(1)]\pi$, i.e., $\Pi^0 = (z-1)[I - A^{0'}(1)]\pi$; boundary conditions $\Pi^k, 1 \leq k \leq n$ are obtained from Theorem 2; and for all $k > n$,

$$q^k(z) = [Iz - A^0]^{-1}\sum_{l=1}^{n} A^l(q^{k-l}(z) - q^{k-l}(1)). \quad (6)$$

Now, with $f^k \overset{def}{=} (q^k, q^{k-1}, \ldots, q^{k-n+1})$, we can construct a linear operator Λ such that

$$f^k = \Lambda f^{k-1}, \quad (7)$$

where f_1^k is given by recursion 6 and $f_j^k = f_{j-1}^{k-1}, 2 \leq j \leq n$. Thus, the radius of convergence of the queueing series is given by $1/r = \lim_{m\to\infty} ||\Lambda^m f^n||^{1/m}$, where $||\cdot||$ can be chosen to be the Euclidian norm. One way of finding the preceding limit is to find all the eigenvalues (λ) and eigenvectors (e_λ) of the operator Λ, and then possibly find the integral representation of the function f^n as

$$f^n = \int e_\lambda d\mu(\lambda) + g^n, \quad (8)$$

for an appropriately chosen spectral measure μ, and function g^n belonging to the null space of the operator Λ. If D is the set of points that support measure μ, then (under appropriate conditions) $1/r = \sup\{|\lambda| : \lambda \in D\}$. In order to get more intuition about the problem let us work out the following example. (Our intention here is not to rigorously answer the questions regarding linear operators, rather to use the spectral concept to obtain the desired answer; for the general theory of linear operators see [4].)

Example 1 Let's assume that the arrival process is two state modulated with the transition probabilities $\epsilon p_{01}, \epsilon p_{10}$; when in state one the source is producing 2 packets with probability $(1-a)$ and zero packets with probability a, and when in state 0 there are no arrivals, i.e., $a_0(z) = 1, a_1(z) = a + (1-a)z^2$. After some algebra, by applying Theorem 2 and recursion (6) one finds that $Q_0^0 = p_{10}/(p_{01}+p_{10})$, $Q_1^0 = \frac{(-1+2a)p_{01}(a+z^2-az^2)}{(p_{01}+p_{10})(a-z+az)}$, and $q_0^1(z) = \frac{a\,p_{01}\,p_{10} - a^2\,p_{01}\,p_{10}}{(-1+2a)(p_{01}+p_{10})(a-z+az)}$. Also, for each $k \geq 1$, $\frac{q_1^k(z)}{q_0^k(z)} = \frac{-c_{10}-z^2+c_{10}z^2}{c_{10}-z+c_{10}z}$. This implies that we only need to calculate q_0^k, that is recursively given, for $k \geq 1$, by

$$q^{k+1}(z) = \frac{q^k(1)\,d\,(a-z+az) + (-1+2a)\,q^k(z)\,b(z)}{(-1+2a)\,(-1+z)\,(a-z+az)}, \quad (9)$$

where $d = (-p_{01} + 2a\,p_{01} + p_{10})$, and $b(z) = (-a\,(p_{01}+p_{10}) + p_{01}\,(1-a)\,z - p_{10}\,(1-a)\,z^2)$.

Now that we have all the expansion coefficients at our disposal, we will take a linear operator approach to determine the radius of convergence of the queue distribution asymptotic series. As we mentioned before the above recursion can be viewed as a linear operator Λ that acts on the set of real functions that are finite in the interval $[-1, 1]$. Then, for each real λ, there is an eigenfunction e_λ that satisfies the equation $\Lambda e_\lambda = \lambda e_\lambda$. These eigenfunctions are given by

$$e_\lambda(z) = \frac{z\,(-p_{01} + 2a\,p_{01} + p_{10})(a-z+az)}{(-1+2a)\,(\lambda(-a+z-z^2+az^2) - b(z))},$$

where $b(z)$ is defined by (9), and $x \in \mathbb{R}$. Assuming that there exists the integral representation (8) for q_0^1, the support of the spectral measure $\mu(\cdot)$ can be determined as follows. Observe that q_0^1 has a simple pole at $z = a/(1-a)$. Therefore, the only eigenfunctions that can be used to approximate q_0^1 are either ones that have the same pole $z = a/(1-a)$, or the ones that do not have real poles at all. After relatively simple analysis it can be found that there is no eigenfunction with the pole $z = a/(1-a)$. The exclusion of all the eigenfunctions with real poles from the set of eigenfunctions gives the support of the spectral measure μ to be the real interval (λ_1, λ_2), where $\lambda_i, i = 1, 2$, are analytically given by

$$\lambda_i = \Big(p_{01} - 3\,a\,p_{01} + 2\,a^2\,p_{01} - 4\,a\,p_{10} + 4\,a^2\,p_{10} \pm 2\,\sqrt{(-1+a)\,a\,d\,p_{10}}\Big)\,(-1+2\,a)^{-2},$$

where $d = p_{01} - 2\,a\,p_{01} - p_{10}$. (This argument is similar with one arguing that for the spectral approximation of the even periodic functions one can only consider $\{\cos(n\omega x), n \geq 0\}$) If we assume that the spectral measure for the integral representation of q_0^1 is supported on the whole interval (λ_1, λ_2), it follows that $r = 1/\max(|\lambda_1|, |\lambda_2|)$.

We graphically investigate this radius of convergence. The results are plotted in Figure 4 for different values of parameters: $p_{10} = 1$, $p_{01} = 1/10, 1/100, 1/1000$, and $a \in (0.5, 1)$ (or equivalently for the utilization going from one to zero). The graph appears to be virtually the same for different values of p_{01}; however, it heavily depends on the utilization parameter a (utilization $\rho = 2(1-a)$). We see that for a very close to 0.5 (utilization one) the radius of convergence is very small, and approaches zero as ρ goes to one. For $a > 0.935, (\rho < 0.129)$ the series converges for all $\epsilon \in (0, 1)$.

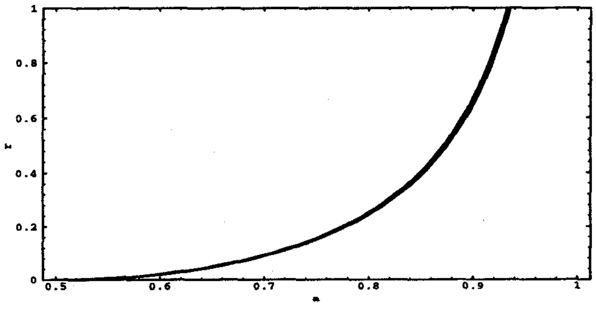

Figure 4: Radius of convergence r versus load parameter a (utilization $\rho = 2(1-a)$, for three different values of $p_{01} = 10^{-i}, i = 1, 2, 3$.

Although the previous analysis gives the precise result on the radius of convergence, it is usually not feasible to determine it for more complex cases. However, the fact that radius of convergence goes to zero as the utilization in at least of one of the states approaches the capacity of the channel can be proven in much greater generality. We state that result in the following theorem.

Theorem 3 *The radius of convergence of the queueing asymptotic series whose expansion coefficients $q^k, k \geq n$, are given by recursion (6) converges to zero if at least in one of the states i, the utilization $\rho_i = a_i'(1)$ converges to the capacity ($c = 1$) of the channel.*

Proof: Given in [9].

We illustrate this theorem in Figure 5; different curves illustrate different numerical examples whose detailed description is given in [9]. From numerous numerical experiments we have observed that the radius of convergence is typically larger than 10^{-3} as long as the utilization of the subchains is smaller than 0.95 and $X(i)$ does not have a heavy tail (see [9]). The dependence of the radius of convergence on the heavy tail of $X(i)$ is illustrated in Figure 6; the dashed line represents the heavy tail case. Since the ratio between the time scales ($\approx \epsilon$) is usually smaller than 10^{-3}, we conclude that the asymptotic expansion will be a useful tool for approximating the small buffer queue length distribution.

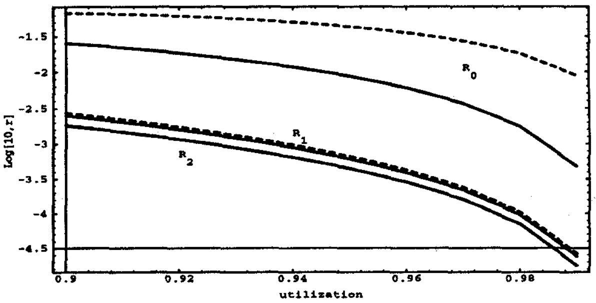

Figure 5: Dependence of the radius of convergence on the conditional utilization.

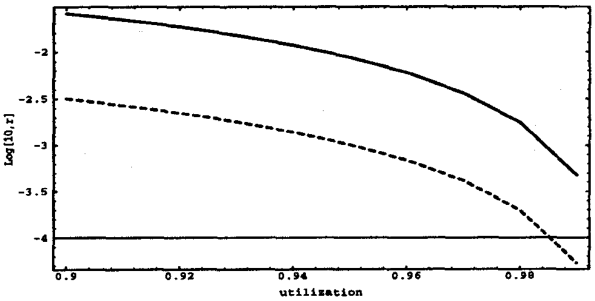

Figure 6: The dependence of the radius of convergence the the heavy tail of $X(i)$; the dashed line represents the heavy tail case.

5 Expansion Error

In order to evaluate the expansion error one has to (i) estimate the error of the approximation of the analytical function $Q(z,\epsilon)$ with its finite approximation $\hat{Q}(z,\epsilon) \overset{def}{=} \sum_{k=0}^{K} Q^k \epsilon^k$; and (ii) translate this approximation into the error of the queue length distribution, which means to translate $|Q(z,\epsilon) - \hat{Q}(z,\epsilon)|$ into the error of its expansion coefficients with respect to z. It is clear that the error function $|Q(z,\epsilon) - \hat{Q}(z,\epsilon)|$ has to be of the order of $O((\epsilon/r)^{K+1})$ in the unit circle. However, how this error translates into the error on the queue length distribution needs to be investigated. For one thing we know that it strongly depends on the ratio ϵ/r. Numerical illustrations of this dependency are given in [9].

Regardless of all the difficulties for arriving at a precise theoretical analysis of the radius of convergence and expansion error, we have found that these quantities exhibit rather nice numerical behavior. The following simple heuristic criterion usually gives good estimates of the radius of convergence

$$\hat{r} = \left| \frac{s_{i_{max}}^{K}}{s_{i_{max}}^{K-1}} \right|,$$

where K is the largest index for which we calculated Q^k, s_j^k is obtained by inverse z-transform of Q^k, and i_{max} is the maximum index for which this transform is computed. An even more important quantity is the relative error of the probabilities, ideally defined as

$$e_r^k(i) \overset{def}{=} \left| \frac{p_i - \hat{p}_i^k}{p_i} \right|,$$

where p_i is the true probability, and the $\hat{p}_i^k$'s are its approximations obtained with the k term expansion. Obviously, the p_i's are not available; replacing them with $\hat{p}_i^{k+1}$, we obtain the estimate

$$\hat{e}_r^k(i) \overset{def}{=} \left| \frac{\hat{p}_i^{k+1} - \hat{p}_i^k}{\hat{p}_i^{k+1}} \right|.$$

Excellent numerical agreement between $\hat{e}_r^k$ and e_r^k is exemplified in Figure 7 for $k = 1, 8$ (true error is plotted with solid lines, and the estimated one with dashed lines.) Thus, should the relative error not exceed δ, using the k term expansion we can estimate the buffer range B_{max} for which our approximation is within the specified error range, i.e.,

$$B_{max}(k) = \max\{i : \hat{e}_r^{k-1}(i) < \delta\}. \qquad (10)$$

Finally, once we find the desired number of the expansion coefficients k and the buffer range $B_{max}(k)$, we can estimate the tail probabilities $p_i, i > B_{max}(k)$, with EB approximation, i.e.,

$$p_i = p_{B_{max}} e^{-(i-B_{max})\theta^*},$$

where θ^* is the EB constant. This procedure is going to be tested in the following section with numerical examples.

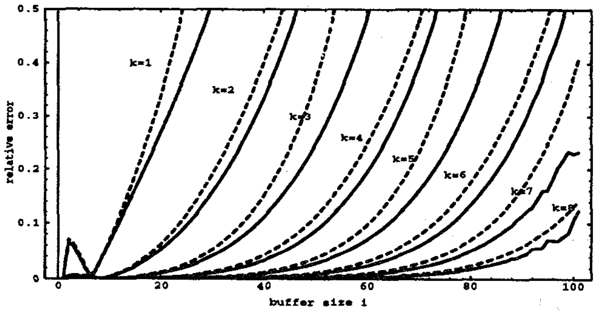

Figure 7: $e_r(k)$ (solid lines), and estimated relative error $\hat{e}_r(k)$ (dashed lines) versus the buffer size i; the graphs are parameterized by the number of terms k used in the asymptotic expansion.

6 Statistical Multiplexing: Numerical Examples

Consider n heterogeneous NDMMIID sources with m.g. matrices given by $A_i(z,\epsilon) = \sum_{k=0}^{m} A^k \epsilon^k$. Then, the m.g.m. for the aggregate process is given as

$$A(z,\epsilon) = \bigotimes_{i=1}^{n} A_i(z,\epsilon), \qquad (11)$$

where $\bigotimes$ stands for the Kronecker product. Since Kronecker product is a distributive operation with respect to addition, by identifying the coefficients with respect to ϵ in equation (11) we can calculate the coefficients of the asymptotic expansion of $A(z,\epsilon) = \sum_{k=0}^{m^n} A^k e^k$. If each of the sources happens to be of the form $A_i(z,\epsilon) = A_i^0 + \epsilon A_i^1$, i.e., consists of only two time scales, than $A^k, 1 \le k \le n$ are simply given by

$$A^k = \sum_{1\le l_1<\ldots<l_k\le n} \bigotimes_{i=1}^{l_1-1} A_i^0 \otimes A_{l_1}^1 \otimes \bigotimes_{i=l_1+1}^{l_2-1} A_i^0 \otimes A_{l_2}^1 \otimes \ldots \otimes A_{l_k}^1 \otimes \bigotimes_{i=l_k+1}^{n} A_i^0. \qquad (12)$$

We will use the formula above in the following numerical examples. Consider statistical multiplexing of the *heterogeneous* On-Off traffic sources; each source i is characterized with transition probabilities p_{i01}, p_{i10}. When in state zero, the source is producing no arrivals and when in state one the source sends i.i.d. arrivals with m.g.f. $a_{i1}(z) = d_{i0} + (1 - d_{i0)})z^b$; note that for different values of b we can experiment with source burstiness. (There is no particular reason for choosing sources like this, except that it makes them easy to parameterize.)

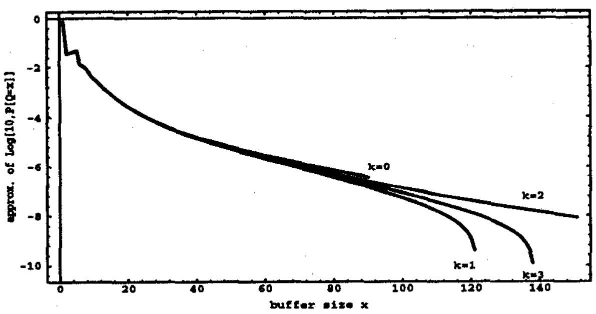

Figure 8: Approximate "probabilities" from example 5 for $k = 0, 3$

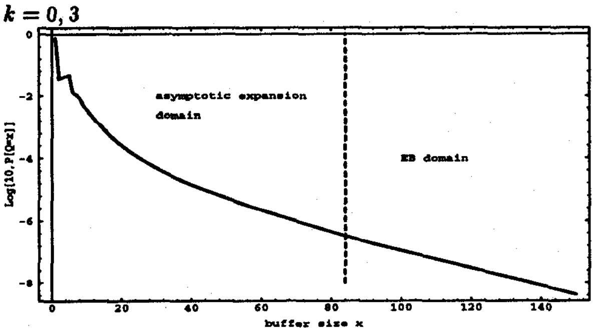

Figure 9: Total queue distribution approximation, obtained by combination of the asymptotic expansion method and EB approximation.

Example 4 In this example we consider SMUX of 5 heterogeneous On-Off sources with the following set of parameters: $p_{i01} = 1/(4+i), p_{i10} = 1/2, i = 1,5$, $d_{i0} = 1 - 9/(40\,(2+i)), i = 1,4, d_{50} = 791/800, b = 4$, $\epsilon = 5.5\ 10^{-3}$. The aggregate $A(\epsilon, z)$ matrix has a dimension 32 x 32! We have calculated first four expansion functions $Q^k(z), k = 0,3$, and its z-transform inverses; the results are plotted in Figure 8. For the relative error bound $\delta = 0.2$ we have estimated the reliable buffer region $B_{max} = 83$ ($\hat{r} \approx 7\ 10^{-4}$). For buffer sizes $i > B_{max}$ we have used the EB approximation ($\theta^* = 0.0657$). The combination of asymptotic expansion and EB approximation is plotted in Figure 9. We see (Figure 9) that the transition between the two approximations is smooth. Therefore, although we have no error estimate in the EB domain, from the smoothness of fit we can expect that the approximation is excellent in the EB domain too. This *smoothness of fit* can be used as a heuristic criterion for the overall goodness of the approximation.

Let us now compare this approximative method with the classical exact z-transform inversion. The first step in order to obtain the exact solution is to find the inverse of $[Iz - A(\epsilon, z)]$, then to find 31 roots of the polynomial $\det[Iz - A(\epsilon, z)]$ in the unit circle; to use these roots to obtain boundary probabilities $q_{j0}, 1 \leq j \leq 32$, and at last to find the inverse z-transform of the vector $Q(\epsilon, z)$. For comparison, we have not been able to complete even the first step, i.e., to find $[Iz - A(\epsilon, z)]^{-1}$ after 24 hours, when we stopped the program. (Computation was attempted with Mathematica 2.2 on (150MHz, 64M RAM, 100M virtual memory) on SGI machine. However, (using the same environment Mathematica + SGI) we obtained $\hat{p}_i^2, 1 \leq i \leq 150$, in less than an hour. This clearly shows the efficacy of the asymptotic expansion method. The second example that we have chosen is even of a large size (64 x 64).

Example 5 In this example we multiplex 6 On-Off sources. The dimensionality of the problem is 64 x 64. It is needless to say that it is completely hopeless to find the exact solution. The parameters are: $p_{i01} = 1/(4+i)$, $p_{i10} = 1/2, i = 1,6$, $d_{10} = 1 - 9/160$, $d_{20} = 1 - 9/200$, $d_{30} = 1 - 9/240$, $d_{40} = 1 - 9/240$, $d_{50} = 1 - 9/600$, $d_{60} = 1 - 9/480$, $b = 2$, $\epsilon = 6.7\ 10^{-4}$. Again we have calculated the first four coefficients $Q^k(z), k = 0,3$. The inverses of $\hat{Q}^k(z), k = 0,3$ are plotted in Figure 10. Using the same method we calculate: $\hat{r} \approx 9\ 10^{-4}$, B_{max} ($\delta = 0.2$), $\theta^* = 0.1298$. And with the combination of the EB approximation we present the solution in Figure 11. Again, we can observe the smoothness of the fit between the asymptotic expansion and the EB domain.

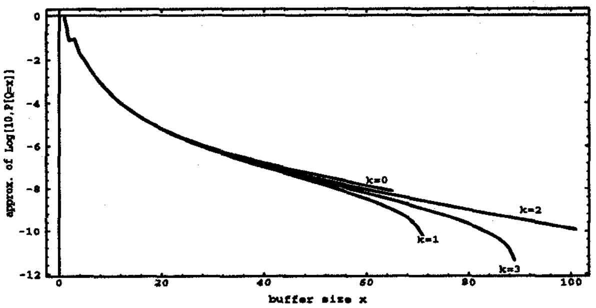

Figure 10: Approximate "probabilities" from example 5 for $k = 0, 3$.

We have already seen that using the asymptotic expansion method we were able to solve rather sizable examples, for which the exact solution is either very difficult or almost impossible to find. Having in mind that all the calculations are done with Mathematica 2.2 which is known to be slow for intensive numerical problems, we expect that this method is going to be much faster when implemented in C. With this in mind we expect that our method will be very useful for large practical problems that arise in ATM admission control.

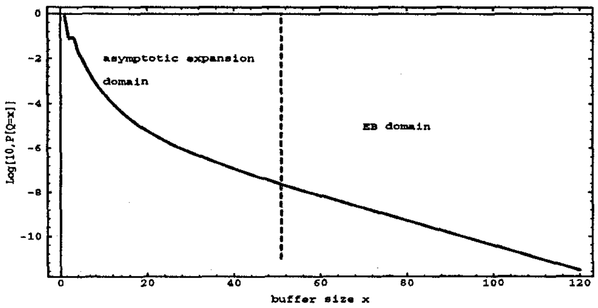

Figure 11: Total queue distribution approximation, obtained by combination of the asymptotic expansion method and EB approximation.

7 Conclusion

In this paper we have developed a recursive asymptotic expansion for approximating the queue length distribution of "small" to "moderate" buffer ranges under the condition of strictly stable multiple time scale arrivals. We discussed the radius of convergence of the asymptotic expansion series. Using methods of linear operator theory (for a class of examples) we were able to *analytically* solve for the radius of convergence. We also gave general sufficient conditions under which the radius converges to zero; this roughly shows which situations have to be avoided for the method to work well. Further investigation of the radius of convergence was done numerically. The numerical analysis indicated that the radius of convergence is reasonably large as long as the arrivals utilization is not very close to one (capacity), or exhibit heavy tail. Also, we numerically showed that a simple probability error estimation gives reliable results. At last, combining the asymptotic expansion method with the EB approximation, we suggested an approximation for the buffer probabilities of all buffer ranges. In short, our procedure can be summarized as follows:

- *calculate the desired number k of expansion coefficients, and estimate the queue distribution by inverting $\hat{Q}^k$.*
- *estimate the reliable buffer range B_{max}.*
- *use EB constant for approximating the probabilities of the buffer sizes greater than B_{max}.*

This procedure has been tested on extensive numerical examples. We suggest to apply it to efficient admission control in ATM networks.

References

[1] C. S. Chang. Stability, queue length and delay of deterministic and stochastic queueing networks. *IEEE Transactions on Automatic Control*, 39:913–931, 1994.

[2] Gagan L. Choudury, David M. Lucantoni, and Ward Whitt. Squeezing the most of ATM. *to appear in IEEE Trans. on Communications*, 1995.

[3] P. J. Courtois. *Decomposability, Queueing and Computer System Applications*. Academic Press, 1977.

[4] N. Dunford and J. T. Schwartz. *Linear Operators I, II, III.* John Wiley and Sons, 1958-71.

[5] A. Elwalid, D. Heyman, T. V. Lakshman, D. Mitra, and A. Weiss. Fundamental bounds and approximations for atm multiplexers with applications to video teleconferencing. *IEEE Journal on Selected Areas in Communications*, 13(6):1004–1016, August 1995.

[6] A. I. Elwalid and D. Mitra. Effective bandwidth of general markovian traffic sources and admission control of high speed networks. *IEEE/ACM Trans. on Networking*, 1(3):329–343, June 1993.

[7] P. V. Glynn and W. Whitt. Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *Studies in Appl. Prob.*, 1994.

[8] R. Guerin, H. Ahmadi, and M. Nagshineh. Equivalent capacity and its application to bandwidth allocation in high-speed networks. *IEEE J. Select. Areas Commun.*, 9:968–981, 1991.

[9] P. R. Jelenković and A. A. Lazar. Evaluating the queue length distribution of an ATM multiplexer with multiple time scale arrivals. CTR Technical Report CU/CTR/TR 420-95-26, Columbia University, July 1995. (www: http: //www.ctr.columbia.edu/comet/publications).

[10] P. R. Jelenković and A. A. Lazar. On the dependence of the queue tail distribution on multiple time scales of ATM multiplexers. CTR Technical Report CU/CTR/TR 400-95-06, Columbia University, March 1995. (www: http: //www.ctr.columbia.edu/comet/publications).

[11] P. R. Jelenković and A. A. Lazar. On the dependence of the queue tail distribution on multiple time scales of ATM multiplexers. In *Conference on Information Sciences and Systems*, pages 435–440, Baltimore, MD, March 1995. (www: http: //www.ctr.columbia.edu/comet/publications).

[12] P. R. Jelenković and A. A. Lazar. Subexponential asymptotics of a markov-modulated G/G/1 queue. *Submitted to Journal of Appl. Prob.*, November 1995.

[13] P. R. Jelenkovic and B. Melamed. Algorithmic modeling of tes processes. *IEEE Transactions on Automatic Control*, 40(7):1305–1312, July 1995.

[14] F. P. Kelly. Effective bandwidths at multi-class queues. *Queueing Systems*, 9:5–16, 1991.

[15] A. A. Lazar, G. Pacifici, and D. E. Pendarakis. Modeling video sources for real-time scheduling. *Multimedia Systems*, 1(6):253–266, 1994.

[16] R. M. Loynes. The stability of a queue with non-independent inter-arrival and service times. *Proc. Cambridge Philos Soc.*, 58:497–520, 1968.

[17] Paul Skelly, Mischa Schwartz, and Sudhir Dixon. A histogram-based model for video traffic behavior in an ATM multiplexer. *IEEE/ACM Transactions on Networking*, 1(4):446–456, 1993.

[18] D. Tse, R. Gallager, and J. Tsitsiklis. Statistical multiplexing of multiple time-scale markov stream. *IEEE, Selected Areas in Communications*, August 1995.

[19] A. Weiss and A. Shwartz. *Large deviations for performance analysis : queues, communications, and computing.* New York: Chapman & Hall, 1995.