SQUEEZING THE MOST OUT OF ATM

Gagan L. Choudhury,[1] David M. Lucantoni[2] and Ward Whitt[3]

Abstract—**Even though ATM seems to be clearly the wave of the future, one performance analysis indicates that the combination of stringent performance requirements (e.g., 10^{-9} cell blocking probabilities), moderate-size buffers and highly bursty traffic will require that the utilization of the network be quite low. That performance analysis is based on asymptotic decay rates of steady-state distributions used to develop a concept of** *effective bandwidths* **for connection admission control. However, we have developed an** *exact numerical algorithm* **that shows that the effective-bandwidth approximation can overestimate the target small blocking probabilities by several orders of magnitude when there are many sources that are more bursty than Poisson. The** *bad news* **is that the appealing simple connection-admission-control algorithm using effective bandwidths based solely on tail-probability asymptotic decay rates may actually not be as effective as many have hoped. The** *good news* **is that the statistical multiplexing gain on ATM networks may actually be higher than some have feared. For one example, thought to be realistic, our analysis indicates that the network actually can support** *twice as many sources* **as predicted by the effective-bandwidth approximation. That discrepancy occurs because for a large number of bursty sources the asymptotic constant in the tail probability exponential asymptote is extremely small. That in turn can be explained by the observation that the asymptotic constant decays exponentially in the number of sources when the sources are scaled to keep the total arrival rate fixed. We also show that the effective-bandwidth approximation is** *not always conservative***. Specifically, for sources less bursty than Poisson, the asymptotic constant grows exponentially in the number of sources (when they are scaled as above) and the effective-bandwidth approximation can greatly underestimate the target blocking probabilities. Finally, we develop** *new approximations* **that work much better than the pure effective-bandwidth approximation.**

Keywords—**ATM, connection admission control, statistical multiplexing, effective bandwidths, performance analysis, queueing models, asymptotic decay rates, numerical transform inversion**

I. INTRODUCTION AND SUMMARY

Great energy is being devoted to studying the promising new *asynchronous transfer mode* (ATM) technology for supporting multi-service high-speed communication networks; e.g., see Roberts [39]. As indicated in [39], interest in ATM is stimulated by two factors: first, by new technology making it possible to transmit and switch at very high bandwidths and, second, by growing demand for more sophisticated and powerful communication services.

Even though ATM seems to be well on its way to widespread use, a performance analysis based on asymptotic decay rates of steady-state distributions indicates that the combination of stringent performance requirements (e.g., 10^{-9} cell blocking probabilities), moderate-size buffers and highly bursty traffic associated with these new services will necessitate operating the networks at very low utilizations. (The basic model represents an ATM switch receiving fixed-size ATM cells from several sources and transmitting them over an output channel in a first-in-first-out fashion. The basic question is: How many sources can be admitted for a fixed buffer size with a specified small cell blocking probability?) A main message of this paper is that *ATM networks may be able to achieve higher utilizations than that asymptotic analysis indicates.* In other words, there seems to be more potential for statistical multiplexing gain.

The approximation based on asymptotic decay rates of tail probabilities is very appealing because it supports a concept of *effective bandwidths.* From an engineering perspective, the notion of effective bandwidths is very natural. The idea is to assign each source an effective bandwidth requirement, and then consider any subset of sources feasible (admissible) if the sum of the required effective bandwidths is less than the total available bandwidth. Thus, a suitable notion of effective bandwidths could go a long way toward solving the connection admission control problem. For instance, once effective bandwidths have been assigned, we can approach engineering and design problems using multi-rate loss network models, e.g., as in Choudhury, Leung and Whitt [12,13] and references therein.

1 AT&T Bell Laboratories, Room 1L-238, Holmdel, NJ 07733-3030; email: gagan@buckaroo.att.com

2 IsoQuantic Technologies, LLC, 10 Oak Tree Lane, Wayside, NJ 07712; email:dml@cnj.digex.net

3 AT&T Bell Laboratories, Room 2C-178, Murray Hill, NJ 07974-0636; email: wow@research.att.com

A large-deviations asymptotic analysis provides strong support for the simple effective-bandwidth procedure, because it shows that it is asymptotically correct as the buffer size gets large and the tail probabilities get small, and because it provides a basis for assigning actual effective bandwidth values to different sources (voice, data, video, etc.); e.g., see Hui [31], Gibbens and Hunt [26], Guerin, Ahmadi and Naghshineh [28], Kelly [32], Sohraby [41], [42], Chang [10], Whitt [45], Elwalid and Mitra [23], Kesidis, Walrand and Chang [33], Glynn and Whitt [27], Courcoubetis, Kesidis, Ridder, Walrand and Weber [20], and Chang, Heidelberger, Juneja and Shahabuddin [11].

The additive nature of effective bandwidths is clearly appropriate if we let the effective bandwidths be either the source *peak rates* or the source *average rates.* However, it seems intuitively clear that working with peak rates is far too conservative, while working with average rates is far too optimistic. (We show this later in our examples.) Most of the recent work has been aimed at finding appropriate effective bandwidths in between the peak and average rates.

Unfortunately, however, from the outset, teletraffic engineering experience suggests that there may be a flaw in the effective bandwidth concept because *it corresponds to having no traffic smoothing from multiplexing.* In particular, it is known that when many separate bursty sources are multiplexed (superposed), the total is less bursty than the components; i.e., there is a basis for more statistical multiplexing gain than with Poisson sources. This is theoretically supported by the classical limit theorem stating that superpositions of arrival processes, suitably scaled, converge to a Poisson process as the number of component arrival processes increases; e.g., see Çinlar [19]. See Heffes and Lucantoni [30], Sriram and Whitt [43] and Fendick, Saksena and Whitt [25] for related performance studies. In contrast, with the effective-bandwidth approximation, the burstiness of n superposed independent and identically distributed sources is the same as for a single source (e.g., see p. 76 of [45]). This implies that the effective-bandwidth approximation will predict greater congestion for any fixed arrival rate than it should.

Nevertheless, there is a case for the effective bandwidths, because previous teletraffic analysis (such as in [25], [30], [43]) did not focus on extremely small loss probabilities such as 10^{-9}. Such very small loss probabilities naturally suggest that appropriate asymptotics should provide what we want. And the asymptotic analysis associated with effective bandwidths indicates no traffic smoothing.

The question, then, is what will actually happen in real systems? Will there be significant traffic smoothing or not? Is the asymptotic analysis supporting effective bandwidths sufficiently accurate or is it not?

These questions have plagued researchers in recent years because the natural models are very difficult to analyze. It is difficult to calculate very small tail probabilities in models with many independent bursty sources, by either exact analytical formulas or by computer simulation.

Our main contribution is a new algorithm for a queueing model that enables us to compute the desired small tail probabilities exactly when there are many bursty sources. From this exact analysis, we find that in some cases the effective-bandwidth approximation is excellent, but in other cases (which we think are realistic), it is not. Thus, *we conclude that the notion of effective bandwidths based directly on large-deviation asymptotics may not be effective for connection admission control.*

However, from an engineering perspective, the notion of effective bandwidths remains very appealing. Thus it is natural to look for modifications of the proposed effective bandwidth approximation which will work well; e.g., as in Guerin and Gun [29]. It is important to note that our analysis does not rule out the general notion of effective bandwidths. Instead, our analysis only indicates that there may be serious difficulties with the implementation based directly on a particular kind of asymptotics, without refinements.

A second main contribution of this paper is to develop modifications of the basic effective-bandwidth approximation for tail probabilities that are more accurate (namely, (1.1), (1.5) and (1.7) below). However, these new approximations do *not* support the simple notion of effective bandwidths discussed above. Nevertheless, the approximate tail probabilities can be used for admission control.

The key to our analysis here is the exact numerical solution of the $\sum_{i=1}^{n} G_i/G/1$ queueing model, which has a single server, unlimited waiting room, the first-in first-out service discipline and i.i.d. (independent and identically distributed) service times that are independent of a superposition arrival process. The arrival process is the superposition of n independent component general arrival processes. Our algorithm permits these component arrival processes to be heterogeneous *Markovian arrival processes* (MAPs), as in Lucantoni [35,36]; see Section 8 here. (The general theory also allows batch arrivals, but in this paper we consider only single arrivals.) Since superpositions of independent MAPs are again MAPs, it suffices to consider the MAP/G/1 queue. However, the computational effort increases when the number of sources increases. Our algorithm is based on Lucantoni [35]. *A major new idea is the combination of the matrix-analytic results in [35] with numerical transform inversion.* In particular, we use the Fourier-series method in Abate and Whitt [5]. We also use a scheme for accuracy enhancement for computing small probabilities as described in [15]. Moreover, we use a number of techniques for speeding up computation, in order to treat a large number of sources. These techniques are described in Section 8. For the computation of the asymptotic parameters we use algorithms in [1], [2], [14].

We have also extended the numerical transform inversion algorithms to multidimensional transforms [15] and applied the multidimensional inversion to calculate *transient* performance measures in the MAP/G/1 queue [37]. Thus, we are in a position to study the *transient behavior* as well as the steady-state behavior. With the ability to calculate time-dependent tail probabilities, we can study whether or not transient analysis is needed in the admission control problem. Here, however, we only discuss steady-state behavior.

In our examples here, we consider only homogeneous sources, but in Section 7 we also indicate how to treat large systems with heterogeneous sources approximately. We primarily consider sources that are more bursty than a Poisson process, in particular, *Markov modulated Poisson processes* (MMPPs) with two-state Markov environment processes. (The MMPP is a Poisson process whose rate is itself a continuous-time Markov chain.) For the main example, we consider on-off sources in which the arrival rate in one environment state is zero, i.e., interrupted Poisson processes, which are known to be renewal processes having hyperexponential interarrival-time distributions [34]. These bursty source models are often used to model ATM traffic. They can represent quite bursty traffic. We have also obtained similar results for other (non-renewal) MMPP sources and multiple numbers of more than one kind (two or three) of MMPP sources.

We also consider sources that are *less bursty* than a Poisson process. In particular, we consider renewal processes with E_2 (Erlang of order 2) interarrival times. Such less bursty sources might appear in ATM networks as a consequence of traffic shaping at the edge of the network. Contrary to conventional wisdom, we show that the effective-bandwidth approximation *underestimates* the exact tail probabilities with these less bursty sources, so that *the effective bandwidth approximation need not be conservative.*

As a surrogate for the steady-state blocking probability with a finite buffer, we consider the tail probability of the steady-state waiting time in our infinite-capacity queue. When the service times are deterministic with mean 1, as is the case with ATM cells (with the appropriate time units), the least integer above the steady-state waiting time coincides with the steady-state queue length or buffer content seen by an arrival. (We can also directly compute the queue-length distribution in the MAP/G/1 queue.) We are thus approximating the steady-state blocking probability in the finite-capacity system by the steady-state probability that the buffer content exceeds that capacity level in the infinite-capacity system at an arrival epoch.

Hence, we let W be the steady-state waiting time until beginning service and we focus on the tail probabilities $P(W > x)$. It turns out that in considerable generality these waiting-time tail probabilities are *asymptotically exponential*, i.e.,

$$P(W > x) \sim \alpha e^{-\eta x} \text{ as } x \to \infty , \qquad (1.1)$$

where η is a positive constant called the *asymptotic decay rate*, α is a positive constant called the *asymptotic constant*, and $f(x) \sim g(x)$ as $x \to \infty$ means that $f(x)/g(x) \to 1$ as $x \to \infty$; see Abate, Choudhury and Whitt [2]. In standard single-source examples the approximation provided by (1.1) is often remarkably accurate [2]–[4]. Moreover, numerical experience indicates that for bursty sources the approximation provided by (1.1) with α replaced by 1 often tends to be conservative, i.e.,

$$P(W > x) \leq e^{-\eta x} \text{ for all } x . \qquad (1.2)$$

This is a basis for the *simple one-parameter approximation*

$$P(W > x) \approx e^{-\eta x} \qquad (1.3)$$

The asymptotic decay rate η in (1.1)–(1.3) is the basis for the concept of effective bandwidths. The simple one-parameter approximation (1.3) is appealing because the key parameter η in (1.1)–(1.3) is relatively easy to determine, exactly or approximately. Indeed, it is typically much easier to obtain than the asymptotic constant α. Moreover, the resulting admission-control algorithm using effective bandwidths based on (1.3) is remarkably simple. Hence, *we call* (1.3) *the effective-bandwidth approximation*.

It is important to note that much of the effective-bandwidth literature has not arrived at (1.3) via (1.1). The large-deviation result supporting (1.3) is the weaker limit

$$x^{-1} \log P(W > x) \to -\eta \text{ as } x \to \infty . \qquad (1.4)$$

General sufficient conditions for (1.4) are given in Chang [10] and Glynn and Whitt [27]. Clearly (1.1) implies (1.4), but not conversely. Indeed, the weaker limit (1.4) yields no information about the asymptotic constant α in (1.1).

We have indicated that most of the effective-bandwidth literature has arrived at approximation (1.3) via the limit (1.4) or related bounds such as (1.2). However, several people have focused on (1.1) too, namely, Abate et al. [2], Baiocchi [7], Elwalid and Mitra [23,24] and Neuts [38]. Given (1.1), (1.3) is a convenient simple approximation because α is much harder to obtain than η, and because (1.3) is consistent with the notion of effective bandwidths.

As discussed in [3], for standard single-source models approximation (1.3) is often a reasonable substitute for approximation (1.1)–(1.4) when we are primarily interested in percentiles of the distribution, rather than the probabilities themselves. If the asymptotic constant α in (1.1) is not too different from 1, then it plays a relatively small role in higher percentiles.

It should be noted that the asymptotic decay rate η in (1.1) is identical to the asymptotic decay rate for the blocking probability in the corresponding finite-buffer model as the buffer size gets large; see Baiocchi [7]. Hence, our analysis of the infinite-capacity model is directly relevant to the finite-capacity model with large buffer sizes.

In order to do better than (1.1), without calculating the exact tail probabilities, we have also developed a *refined three-term approximation* of the form

$$P(W > x) \approx \alpha_1 e^{-\eta_1 x} + \alpha_2 e^{-\eta_2 x} + \alpha_3 e^{-\eta_3 x} , \qquad (1.5)$$

where α_1 and η_1 are the asymptotic constant and asymptotic decay rate in (1.1), while α_2, α_3 η_2 and η_3 are chosen to match the probability of delay $P(W > 0)$ and the first three moments EW, $E(W^2)$ and $E(W^3)$, with η_2 and η_3 required to satisfy $\eta_1 \leq \min\{\eta_2, \eta_3\}$; see Choudhury, Lucantoni and Whitt [17].

We thus have four ways to ''calculate'' the tail probability $P(W > x)$: (1.3), (1.1), (1.5) and the full numerical algorithm. Each successive method tends to be more accurate, but each successive method requires substantial more computational effort. The effective-bandwidth approximation (1.3) is by far the easiest to compute, with there being virtually no limit to the size of the model; e.g., see [45].

Using our exact numerical algorithm for the MAP/G/1 queue, we have investigated the three approximations in (1.3), (1.1) and (1.5). For many standard single-source models, the refinement in (1.1) is remarkably accurate even for x not too large, e.g., for the 90th percentile of the distribution. For standard single-source models for which (1.1) is excellent at the 90th percentile and beyond, (1.5) typically is excellent for the entire distribution.

However, we have found that *the story changes dramatically when the arrival process is the superposition of many component processes.* Then the effective bandwidth approximation (1.3) can perform very badly. It is even possible for (1.1) and (1.5) to perform badly, but there is a substantial region where (1.3) performs badly and (1.1) performs well. This is the basis for our new improved approximations beyond (1.3).

We emphasize that, for the model being considered, the limits in (1.1) and (1.4) do indeed hold, and we are indeed interested in very small tail probabilities, and relatively large x. However, there is a serious difficulty (evidently pointed out for the first time here): *The asymptotic constant* α *in (1.1) can be very different from 1 when the number of component arrival processes is large.*

We also explain *why* the asymptotic constant α can be very different from 1 with many sources. To do so, we consider the case of n identical sources with fixed total rate. As n increases, the total rate is kept fixed by properly scaling the individual streams. With this structure, it is known that the asymptotic decay rate η in (1.1) is actually independent of n. We give numerical examples showing that the asymptotic constant with n sources (and this scaling), α_n, is itself *asymptotically exponential* in n, i.e.,

$$\alpha_n \sim \beta e^{-n\gamma} \text{ as } n \to \infty , \qquad (1.6)$$

where $\beta > 0$ and, for sources more bursty than Poisson, $\gamma \geq 0$, while for sources less bursty than Poisson, $\gamma \le 0$. Moreover, $|\gamma|$ in (1.6) tends to be larger when the burstiness gets further from Poisson and the traffic intensity decreases, which is a likely operating condition for ATM networks.

Combining (1.1) and (1.6), we obtain the refined asymptotic approximation (with the scaling for (1.6))

$$P(W > x) \approx \beta e^{-n\gamma} e^{-\eta x} . \tag{1.7}$$

Our numerical results supporting (1.6) and (1.7) naturally suggest a limit theorem, on which we were working but others have delivered. The idea is still to consider large-deviation asymptotics, but now letting both n and x go to infinity. Since this paper was originally submitted, large deviation limits supporting (1.6) and (1.7) have been reported by several authors, the first known to us being A. Dembo and O. Zeitouni (talk at National Meeting of the Operations Research Society of America, Boston, April 1994). Large-deviation papers supporting (1.6) and (1.7) have been written by Botvich and Duffield [9], Courcoubetis and Weber [21], Simonian [40] and Tse, Gallager and Tsitsiklis [44]. To state their result, let W^n be the steady-state waiting time with n sources. Paralleling (1.4), their result is

$$n^{-1} \log P(W^n > nx) = I(x) \text{ as } n \to \infty , \tag{1.8}$$

for an appropriate function $I(x)$, yielding the approximation

$$P(W^n > x) \approx e^{-nI(x/n)} . \tag{1.9}$$

Clearly (1.8) provides strong theoretical support for (1.6) and (1.7), but (1.8) is substantially weaker than (1.1) and does not yield the asymptotic constant β in (1.6) and (1.7). Nevertheless, the form of the limit function $I(x)$ provides useful insights, as can be seen from the references.

We propose (1.7) as a basis for developing useful approximations for the tail probabilities when there are large numbers of each of a few types of sources. In particular, we suggest using (1.7) to estimate β and γ by extrapolating using (1.6) and exact calculations for small n; see Section 7 for further discussion.

Approximation (1.7) can be regarded as an approximation to the true asymptote (1.1). Our numerical experience indicates that for sources more bursty than Poisson the true asymptote (1.1) is a good approximation at 10^{-9} if α is not too small, e.g., $\alpha \ge 10^{-4}$. For a wide range of problems, this is the region of interest. Our experience also indicates that the true asymptote (1.1) is a poor approximation at 10^{-9} if α is too small, e.g., $\alpha \le 10^{-8}$. For estimated values of α in between, e.g., $10^{-8} < \alpha < 10^{-4}$, the true asymptote is only moderately good as an approximation. In this region, we propose using the true asymptote with α replaced by 10^{-4} as a rough conservative heuristic approximation. For sources less bursty than Poisson, (1.1) tends to be a good (bad) approximation at a target blocking probability 10^{-9} if $\alpha < 10^4$ ($\alpha > 10^8$).

Here is how the rest of the paper is organized. In Section 2 we consider a specific example with bursty sources to demonstrate that the effective-bandwidth approximation (1.3) can perform poorly. In Section 3 we consider the case of n identical sources scaled so that the total rate is independent of n, and give examples supporting (1.6). Next in Section 4 we present modifications of the example in Section 2 with different numbers of sources in which the two asymptotic approximations in (1.1) and (1.3) are, first, both good and, second, both bad. These cases seem identifiable by looking at the observed value of the asymptotic constant α. In Section 5 we investigate the impact of changing other variables. We show that the asymptotic approximations tend to get worse as the number of sources increases, the buffer size decreases, the channel utilization decreases, the target blocking probability increases, and the source gets further from Poisson, either more bursty or less bursty.

In Section 6 we consider sources less variable than Poisson. For these sources, we show that the effective-bandwidth approximation need not be conservative; i.e., (1.2) does not hold. In Section 7 we show how to do approximations in large systems with heterogeneous sources. In Section 8 we briefly describe the computational algorithm. Finally, in Section 9 we state our conclusions.

II. AN EXAMPLE WHERE THE EFFECTIVE-BANDWIDTH APPROXIMATION FAILS

In this section we consider an example with homogeneous on-off sources. (The on-off property is not essential for our model or for the conclusions we draw from it. A source can have multiple states, each characterized by a different arrival rate. We use such sources in later examples.) Our primary purpose is to show that the effective-bandwidth approximation (1.3) can be a bad approximation for a realistic example and can lead to seriously underestimating the number of sources the queue can accommodate.

We let the service times be i.i.d. with mean 1. Thus the unit of time is the time required to transmit one ATM cell. We let the service time have an Erlang distribution of order 16 (E_{16}), which has *squared coefficient of variation* (SCV, variance divided by the square of the mean) $c_s^2 = 1/16$. This distribution is similar to the deterministic (D) distribution of ATM cells, but better behaved for numerical calculations. (It is possible to do calculations for D distributions, but it requires more computational effort to achieve comparable precision.) In fact, we have computed using the E_{1024} distribution and have found for the examples in this paper with MMPP sources that there is little difference between E_{16} and E_{1024}, supporting our hypothesis that D is well approximated by E_{16}.

The on-off source has exponentially distributed on and off periods. During the on periods, cells arrive according to a Poisson process; during the off periods there are no arrivals. Each on-off source is characterized by three parameters, the mean on period ω, the mean off period ζ and the (peak) rate during the on period, p. Here we let the on and off periods have means $\omega = 436.36$ and $\zeta = 4363.63$, respectively. We let the peak rate be $p = 0.1375$. The overall average rate is thus $\lambda = p\omega/(\omega + \zeta) = 0.0125$. The ratio of peak to mean rates is thus $p/\lambda = 11.0$. The mean number of cells during an on period is $p\omega = 60.0$.

We feel that the level of source burstiness considered above can happen quite naturally in bursty data sources. Furthermore, the example is chosen mainly to illustrate that bad behavior can occur. Later we discuss how the basic behavior of superposed sources changes with change in burstiness, buffer size and other parameters.

The above on-off source (an interrupted Poisson process) can also be characterized as a renewal process with a hyperexponential (H_2, mixture of two exponentials) interarrival time distribution [34]. The mean interarrival time for one source is thus $1/\lambda = 80$ and the SCV is $c_a^2 = 100.2$. Scaled to have mean 1, the first four moments of an interarrival time are: $m_1 = 1$, $m_2 = 101.2$, $m_3 = 1.68 \times 10^4$ and $m_4 = 3.73 \times 10^6$. This is another way to see that the individual sources are quite bursty. (For a simple exponential random variable with mean 1, the n^{th} moment is $n!$)

Since the arrival process is a renewal process, The SCV coincides with the normalized asymptotic variance or *limiting index of dispersion for counts* for the counting process $N(t)$, i.e.,

$$I_c(\infty) \equiv \lim_{t\to\infty} \frac{\mathrm{Var}\, N(t)}{EN(t)} = c_a^2 = 100.2 \, , \qquad (2.1)$$

see [25,43]. The value $I_c(\infty) = 100.2$ is large, so that each source is indeed quite bursty.

We let the buffer size be 600 cells, which falls in the range considered by switch manufacturers. The buffer size 600 is ten times the mean number of cells in an on period; i.e., the buffer contains 10 bursts. The quality of the approximations depends strongly on the buffer size. (We discuss this further in Section 5.) Currently buffer size is limited by the availability of fast memory. We regard 600 as a representative moderate buffer size.

As is customary in ATM, we let the target blocking probability be 10^{-9}. Hence, we choose the number of sources so that

$$P(W > 600) \approx 10^{-9} \, . \qquad (2.2)$$

Our exact numerical results indicate that (2.2) is attained with $n = 24$ sources, yielding a utilization of 30%.

Figure 1 displays the exact tail probabilities $P(W > x)$ and the approximations (1.1), (1.3) and (1.5). In addition, Figure 1 displays the Poisson approximation, which is the exact tail probability in the $M/E_{16}/1$ queue, computed by numerical transform inversion. The tail probability $P(W > x)$ is our approximate probability of buffer overflow with x being the buffer size.

Insert Figure 1 about here.

The (exact) asymptotic parameters in (1.1) are $\eta = 0.01809$ and $\alpha = 1.453 \times 10^{-5}$. The three-term approximation is also fully determined, based on the exact probability of delay $P(W > 0) = 0.4242$ and first three moments $EW = 0.6924$, $E(W^2) = 8.573$ and $E(W^3) = 579.3$.

Of course, there is the question of numerical accuracy of the exact results reported in Figure 1. We verified them using a built-in accuracy check in our numerical procedures, as explained in [15, 37] and Section 8 here. In the related case of 64 sources with exponential service times, numerical results by Elwalid and Mitra [24] confirmed our results.

Figure 1 indicates that (1.1) and (1.5) are still in error by a factor of 3.9 at $x = 600$. This error is relatively small, though, compared to the error in (1.3), which is by a factor of 10^5. The Poisson approximation obviously is even worse than (1.3).

It is interesting to consider Figure 1 in relation to previous discussions of ''cell-level congestion'' and ''burst-level congestion,'' e.g., in [39]. To a large extent, Figure 1 indicates two nearly linear regions for the exact curve. As discussed on p. 19 of [39], there is an initial period of steep nearly linear decline corresponding to ''cell-level congestion'' and a later period of more gradual nearly linear decline corresponding to ''burst-level congestion.'' This burst level congestion corresponds to the true asymptote (1.1). However, the transition between these two regimes is smooth and fairly long. As in [39], much of the ATM literature considers *separate models* to represent cell-level and burst-level congestion. In contrast, we represent *both within a single model*. Indeed, we consider precisely the multiple-MMPP-source model described as difficult on p. 185 of [39]. In this context, a major contribution here is to point out that the intercept with the y-axis of the asymptote, which is the asymptotic constant α in (1.1), may well be very small.

Now suppose, instead, that we use approximation (1.3). Let η_n be η in (1.1) as a function of n (without any rescaling of individual sources). As indicated above, $\eta_{24} = 0.01809$, but η increases as n (and ρ) decreases. Figure 2 displays η_n as a function of n (without any rescaling of individual sources). It turns out that

$$e^{-600\eta_n} \le 10^{-9} \text{ for all } n \le 12 \, , \qquad (2.3)$$

but not for $n \ge 13$; $\eta_{12} = 0.03749$ and $\eta_{13} = 0.03354$. Hence, using the effective-bandwidth approximation (1.3), we conclude that the queue can accommodate only 12 sources with criterion (2.2). Hence, (1.3) indeed significantly underestimates the capacity. *The actual capacity is two times that predicted by (1.3).*

Insert Figure 2 about here.

Table 1 compares seven different procedures for determining the number of sources that can be supported:

(i) The exact tail probabilities

(ii) the effective-bandwidth approximation (1.3)

(iii) the full asymptotic approximation (1.1)

(iv) the three-term approximation (1.5)

(v) the Poisson approximation

(vi) average-rate engineering

(vii) peak-rate engineering.

Insert Table 1 about here.

From Figure 1 and Table 1, we see that even though the true asymptote is not too accurate for the tail probability (being off by a factor of 4), it produces a good estimate for the number of sources (being off by only one).

Since the average rate is $\lambda = 0.0125$, if we could *size by average rate,* then the queue could accommodate 80 sources. Since the peak rate is $p = 0.1375$, if we *sized by peak rate,* then the queue can accommodate 7 sources. The numbers 7 and 80 help put the realized improvement from 12 to 24 going from approximation (1.3) to the exact numerical result in perspective.

There is a simple explanation for the poor performance of approximation (1.3). For $n = 24$, the asymptotic constant in (1.1) is $\alpha_{24} = 1.453 \times 10^{-5}$. As can be seen from Figure 1, in this case the one-term asymptotic approximation $\alpha e^{-\eta x}$ based on (1.1) is a good approximation, but replacing α by 1 introduces a large error.

In customary models with a single source, the asymptotic constant α is usually not too different from 1. Hence, it is interesting to see how α_n depends on n (again, without rescaling the individual sources). This is shown in Figure 3. There we see that α_n declines from $\alpha_1 = 0.340$ to a minimum value of $\alpha_{11} = 6 \times 10^{-8}$ and then increases toward 1 again as the traffic intensity approaches 1. From heavy-traffic theory, we anticipate that $\alpha_n \to 1$ as $\rho_n \to 1$. Assuming that α_n stays well above the target value 10^{-9} for all n, as it does in Figure 3, we can anticipate that the asymptote (1.1) will be a reasonably good approximation, but we clearly cannot simply replace the asymptotic constant α by 1.

Insert Figure 3 about here.

Instead of fixing the buffer size at 600 and asking how many sources the queue can accommodate, we could instead fix the number of sources at 24 and ask what size buffer is required. The exact numerical results indicate a buffer size of 600. The approximations (1.1) and (1.5) indicate that a buffer size of 530 is required, approximation (1.3) indicates 1146 and the Poisson approximation indicates only about 20. These results are also displayed in Table 1. This is another way to look at the weakness of the effective-bandwidth and Poisson approximations. This view also shows that approximations (1.1) and (1.5) are reasonably good.

As shown in [4], the asymptotics (1.1) for the steady-state waiting time are closely related to corresponding asymptotics for other steady-state quantities, such as the workload, sojourn time and queue length (at arrivals and at arbitrary times). For this example, the queue length decay rate parameter σ in [4] is $\sigma = 0.9821$. Although we consider only the waiting time here, our experience indicates that a detailed analysis of one of the other steady-state distributions tells essentially the same story. For the example considered in this section, if we look at the queue length at an arbitrary time (instead of at the arrival instant) then both the exact value and approximation (1.1) drop roughly by a factor of 3, but the effective-bandwidth approximation (1.3) remains unchanged (another weakness of (1.3)). However this further difference by a factor of 3 is small compared to the already-established difference by a factor of 10^5.

III. ASYMPTOTICS FOR THE ASYMPTOTIC CONSTANT IN SCALED SUPERPOSITION PROCESSES

As shown in [10], [18], [23], [24], [27], [45], the asymptotic decay rate η_n in the $\sum_{i=1}^{n} G_i/G/1$ model with the superposition of n i.i.d. arrival processes is independent of n when we fix the total arrival rate by scaling the component arrival processes. In particular, let $\{N(t) : t \geq 0\}$ be the arrival counting process with only one source. The proper scaling is achieved with n sources by letting each component arrival process be distributed as $\{N(t/n) : t \geq 0\}$. If the arrival rate of $N(t)$ is λ, then the arrival rate of $N(t/n)$ is λ/n, so that the total arrival rate with n sources is λ for all n. (This scaling is also discussed previously in [43] and 25].)

Of course, when we add sources in a real network, we do not do any rescaling, so we did not do any rescaling in Section 2. However, since η_n is independent of n with this rescaling, the rescaling helps us understand what is happening with the various approximations. As discussed in [43], a key theoretical reference point is the fact that, with the scaling, the superposition process approaches a Poisson process as $n \to \infty$. Since η_n does not change with n, we see that the two limits $x \to \infty$ and $n \to \infty$ do not interchange. This is a source of our difficulties.

Our numerical experience indicates that with this rescaling as $n \to \infty$ the asymptotic constant α_n in (1.1) approaches 0 for sources more bursty than Poisson and approaches infinity for sources less bursty than Poisson. More precisely, α_n appears to decay or grow exponentially as in (1.6) as $n \to \infty$. Numerical evidence supporting (1.6) is given in Figures 4 and 5 for sources more bursty than Poisson. There we display α_n as a function of n in log scale for different examples. Figure 4 displays α_n as a function of n for the example in Section 2, while Figure 5 displays α_n as a function of n in four other examples. For Figure 4, the reference case is the case with $n = 24$ sources and $\rho = 0.30$. All other cases in Figure 4 are rescaled to have this same ρ. In each example we rescale the arrival processes as n changes, so that η_n is independent of n. The linearity in Figures 4 and 5 for n not too small provides strong support for (1.6). (Figure 12 in Section 6 provides similar support for sources less bursty than Poisson.)

Insert Figures 4 and 5 about here.

The four examples depicted in Figure 5 represent all combinations of two different traffic intensities, $\rho = 0.5$ and $\rho = 0.8$, and two MMPP arrival processes. The service-time distributions always are E_{16} with mean 1. As before, these represent nearly deterministic service times. The two arrival processes are two-state MMPPs with asymptotic variance constants in (2.1) of 2.8 and 28. These represent moderately bursty and substantially more bursty sources, respectively (but less bursty than the example in Section 2). These examples have positive arrival rates in both environment states, so that the component MMPPs are *not* renewal processes.

In particular, the MMPPs are characterized by four parameters, one of which can be taken to be the arrival rate. The ratio of the arrival rates in the two environment states is fixed at 4. The expected numbers of arrivals during visits to the two environment states are equal, 5 in the less bursty example with $I_c(\infty) = 2.8$ and 75 in the more bursty example with $I_c(\infty) = 28$.

From Figure 5, we see that the key decay rate parameter γ in (1.6) is decreasing in the traffic intensity ρ but increasing in the burstiness. This appears to be a general tendency. This means that the phenomenon in Section 2 is most likely to occur with low ρ and high burstiness, which is what we anticipate for ATM. The phenomenon might have been missed by others, because they focused on higher ρ and less bursty sources. For instance, the examples in Section 5 of Elwalid and Mitra [23] all have high ρ (above 0.75) and lower burstiness, so that $\alpha > 10^{-2}$.

IV. THREE REGIMES FOR THE ASYMPTOTIC APPROXIMATIONS

The example in Section 2 was one for which the asymptotic approximation $\alpha e^{-\eta x}$ in (1.1) is pretty good, but the simple approximation $e^{-\eta x}$ in (1.3) is bad (because $\alpha \approx 10^{-5}$). In this section we present modifications of this example in which, first, *both* approximations (1.1) and (1.3) are *good* and, second, *both* approximations (1.1) and (1.3) are *bad.*

To obtain these alternative cases, we simply modify the number of sources in the example with $n = 24$ in Section 2, scaling the arrival processes as in Section 3 to keep the total arrival rate fixed at $\lambda = 0.3$ and η fixed at $\eta = 0.01809$. In particular, the two alternative regimes are obtained by letting $n = 2$ and $n = 60$. The results are displayed in Figures 6 and 7.

Since η is independent of n, approximation (1.3) is the same for the three cases $n = 2$, $n = 24$ and $n = 60$. However, $\alpha_2 = 0.505$, $\alpha_{24} = 1.45 \times 10^{-5}$ and $\alpha_{60} = 2.22 \times 10^{-12}$. (For $n = 1$, $\alpha_1 = 0.992$.) From Figures 1, 6 and 7, it is evident that we have the three regimes as claimed. In Figure 6 with $n = 2$, approximations (1.1), (1.3) and (1.5) are all very close, while in Figure 7 it is evident that approximations (1.1), (1.3) and (1.5) are very far apart.

Insert Figures 6 and 7 about here.

Since $\alpha_{60} \approx 10^{-12} < 10^{-9}$ when $n = 60$, it should come as little surprise that the asymptotics (1.1) have not kicked in by the time the tail probabilities reach 10^{-9} in this case. As a rough rule of thumb, it appears that approximation (1.1) tends to be good only when α_n is greater than the desired tail probability $P(W > x)$. To a large extent, it appears that we are able to understand when the various approximations will be sufficiently accurate by computing only the asymptotic constant α. (This computation is possible, even for very large n, by computing α_n for small values of n and then extrapolating using (1.6).)

From Figure 7, we also see that the three-term approximation (1.5) is not accurate at 10^{-9}. The performance of approximation (1.5) in Figure 7 with $n = 60$ is much worse than in Figure 1 with $n = 24$. It is reassuring that its poor performance is signalled by the fact that it is not possible to find parameters exactly matching the third moment in this case; see [17]. Moreover, the three-term approximation performs pretty well even in Figure 7 for tail probabilities above 10^{-4}, for which it was originally designed. The poor performance at 10^{-9} in Figure 7 suggests that alternative fitting procedures should be considered for extremely small probabilities such as 10^{-9}. (This is being investigated.)

V. CHANGING OTHER PARAMETERS

In Section 4 we saw what happens as we changed the number of sources, scaling them in the manner of Section 3 so that the total arrival rate remains unchanged. We saw that the quality of the asymptotic approximations decline as n increases.

More generally, we conclude that the asymptotic approximations tend to get worse as the number of sources increases, the buffer size decreases, the channel utilization decreases, the target blocking probability increases, and the source gets further from Poisson. However, we would also like to point out that there is significant difference in the accuracy and the region of applicability of the effective bandwidth asymptotic (1.3) and the other asymptotic approximations (1.1) and (1.5). Roughly speaking, (1.3) is not bad when $10^{-2} < \alpha < 10^{2}$, whereas (1.1) and (1.5) are not bad when $10^{-8} < \alpha < 10^{8}$. We expect that in practical engineering situations with multiple sources (1.1) and (1.5) will typically be a reasonably good approximation, while (1.3) may be quite bad.

In this section we consider the effects of buffer size and burstiness. Figures 8 and 9 compare the approximations with exact values when the buffer size is 6000 and 60, respectively, instead of 600 as in Section 2. Here we keep the number of sources fixed at $n = 24$. We scale the sources in the manner of Section 3 until the blocking probability at the indicated capacity is 10^{-9}.

Insert Figures 8 and 9 about here.

So far, we have focused on the steady-state waiting time at arrival epoch. Similar results hold for the steady-state workload at an arbitrary time (the virtual waiting time). To demonstrate this, the remaining numerical results in this paper, including those in Figures 8 and 9, are for the steady-state workload. (The differences between the waiting time and workload are negligible compared to the main phenomena being discussed.)

When the buffer size is increased to 6000, the buffer holds 100 bursts instead of 10. Figure 8 shows that all the approximations perform very well with this larger buffer, just as in Figure 6. Now the utilization is 0.835. The effective-bandwidth approximation works very well; it predicts that the system can support 23 sources, which is within 1 of the exact value.

In contrast, Figure 9 shows that the approximations get even worse when we decrease the buffer size from 600 to 60, which corresponds to just one burst. Figure 9 parallels Figure 7. Now the utilization is only 0.18. The effective-bandwidth approximation would now only admit 10 sources.

Suppose now we keep the buffer size at the high value of 6000 and the number of sources at $n = 24$, but increase the burstiness. Suppose that we increase the burstiness by multiplying the mean number of bursts in an on period by 5. We keep the arrival rate fixed by multiplying the ratio of the off period to the on period by 5. This makes the mean number of arrivals in an on period 300, which means that the buffer now holds $6000/300 = 20$ bursts.

Figure 10 compares the approximations with exact values in this case. Figure 10 shows that there is once again a big difference between the effective-bandwidth approximation and the exact result, just as in Figure 1. For this example, the effective-bandwidth approximation would admit only 11 sources. Hence, the advantage of the larger buffer of size 6000 is offset by the larger burstiness. The actual performance of the approximations obviously depends on the combination of variables that actually prevails.

Insert Figure 10 about here.

VI. SOURCES LESS BURSTY THAN POISSON

In this section we consider sources that are less bursty than Poisson. In particular, we assume that each source is renewal with interarrival times that are E_2. As before, we assume that the overall arrival process is the superposition of independent versions of the single source process. It seems unlikely that the sources in an ATM network will actually be less bursty than Poisson, but this is a possibility, due to traffic shaping at the network edge.

In the earlier bursty model we approximated the deterministic cell-length distribution by an E_{16} distribution and commented that E_{16} and D give about the same result. However, this is no longer true with less bursty arrival processes. Therefore, in order to remain pretty close to D, in this section we assume an E_{1024} service time distribution.

Figure 11 displays the approximations and exact tail probabilities for this case assuming 24 sources and a buffer size of 8. In order to meet the 10^{-9} buffer overflow probability, the channel utilization has to be 29%. In this case the buffer overflow probability is greatly *underestimated* by the effective-bandwidth approximation (1.3). For this case of less bursty sources, the effective-bandwidth approximation (1.3) would admit 39 sources in order to meet the buffer overflow requirement with a buffer of size 8, instead of the proper number of 24.

Insert Figure 11 about here.

Figure 11 also shows the probability of buffer overflow for Poisson arrivals. For large tail probabilities, the exact result is close to the Poisson prediction and, for small tail probabilities, the exact result approaches the true asymptote. It is also interesting to observe that the true asymptote is about ten orders of magnitude higher than the effective bandwidth approximation. This is in sharp contrast with the earlier highly bursty examples where the true asymptote was always smaller than the effective bandwidth approximation. Also, note that there is a qualitative change in the shape of the exact curve. It is concave in Figure 11 as opposed to being convex for the more bursty sources.

The behavior of the tail probabilities in Figure 11 may be understood by plotting, in log scale, the asymptotic constant α_n as a function of n, the number of sources. This we do in Figure 12 for the E_2 sources at channel utilizations 0.3 and 0.7 respectively. This is similar to what we did in Figure 5, and indeed Figures 5 and 12 look similar in the sense that in both cases the logarithm of α_n is asymptotically linear with n, i.e., α_n changes exponentially with n. However, the striking difference is that with more bursty sources (Figure 5) α_n decays exponentially with n and approaches zero, while in the less bursty E_2 case, α_n grows exponentially with n and approaches infinity. The growth rate increases as the channel utilization decreases. We have also plotted how the α_n changes with n for Poisson sources with channel utilizations of 0.3 and 0.7 respectively. Of course, α_n does not change with n in the Poisson case. The Poisson case places in perspective the spectacular growth rate of α_n with n in the non-Poisson case.

Insert Figure 12 about here.

As before, the effective-bandwidth approximation (1.3) performs better with larger buffer sizes. This is shown in Figure 13 where the buffer size is 100 instead of 8. As before, the number of sources is $n = 24$. The sources have been scaled in the manner of Section 3 so that $P(W > 100) = 10^{-9}$. This drives the channel utilization up to 0.949. Figure 13 shows that all the approximations are close to the exact values in this case.

Insert Figure 13 about here.

VII. APPROXIMATIONS FOR LARGE SYSTEMS WITH HETEROGENEOUS SOURCES

In this section we propose a method for obtaining useful approximations for tail probabilities in large systems. Approximations are needed, because our exact algorithm cannot handle large numbers of sources, since the number of phases of the superposed Markovian Arrival Process (MAP) grows rapidly with the number of sources. Specifically, it can be shown that if there are L types of sources, k_i sources of type i, and each source of type i has m_i phases, then the total number, P, of phases of the superposed MAP is given by

$$P = \prod_{i=1}^{L} \begin{bmatrix} k_i + m_i - 1 \\ m_i - 1 \end{bmatrix} . \tag{7.1}$$

Note that for Poisson sources $m_i = 1$ and $\begin{bmatrix} k_i + m_i - 1 \\ m_i - 1 \end{bmatrix} = 1$, so that we can add any number of them without increasing P. To run the exact model in reasonable time, we need P to be at most about 100. In all the numerical examples in this paper we assumed $L = 1$ and $m_i = 2$ so that we could treat up to about $K_1 = 99$ sources. (We actually considered up to 60 sources). However, with $L = 2$ and 3 (with m_i still fixed at 2) we can treat only up to $K_i = 9$ and $K_i = 3$ sources of each type.

Now we specify our approximation procedure in terms of two examples. First, suppose that there are three classes, each with 100 homogeneous sources. We can calculate the asymptotic decay rate η exactly by considering the three-class system with one source in each class, with each source scaled appropriately, as in Section 3, so that the arrival rate of each single source equals the total arrival rate for all 100 sources, for each class, in the original system.

In order to approximately determine the asymptotic constant α in the original system, we calculate the exact asymptotic constants in the systems with three classes and k sources in each class, again scaled to be consistent with the original system according to Section 3, for several feasible k, e.g., $k = 1$, 2 and 3. (Note that here $L = 3$ and $m_i = 2$, so that $k_i = 3$ for all i is feasible.) Then, assuming (1.6), we obtain an approximate asymptotic constant for the original system by fitting β and γ in (1.6) to the data. If the resulting estimated asymptotic constant α is substantially greater than the target tail probability, then we can apply approximation (1.1) with confidence. If the estimated asymptotic constant α is less than the target tail probability, then we note that (1.1) is probably not appropriate. If the estimated asymptotic constant is greater than the target tail probability, but not much greater, then we might use the heuristic suggested in Section 1; i.e., we might use (1.1) with a higher value of α as a rough conservative estimate.

The approach we have just described is satisfactory if all classes have many sources, as when there are 100 sources from each of the three classes. However, in actual applications, e.g., with video sources, there may be only a few sources from some classes. To illustrate, suppose that we have four classes, with 1 source in the first class, 2 sources in the second class and 100 sources each in the last two classes. Let all sources have two states. The method we have just applied does not work for this example, but a modification does.

We have found that the asymptotic relation (1.6) still holds if we divide the sources into two groups, and hold one group of sources fixed, while we multiply the number of sources in the second group by n. As before, we scale the sources in the second group, so that the total arrival rate for each class remains fixed, independent of n.

As before, (1.1) holds for each n and the asymptotic decay rate η is independent of n. Moreover, numerical experience indicates that the asymptotic relation (1.6) still holds, but now β in (1.6) is a function of the fixed sources. Now to estimate α in the original system we at first estimate β and γ in (1.6) using $n = 2$ and 3 and then estimate α from those using $n = 100$. Note that here $L = 4$, $m_i = 2$ for each i, $k_1 = 1$, $k_2 = 2$, $k_3 = k_4 = n$ and hence $n = 2$ and 3 are feasible for the exact model.

More generally, the asymptotic result just described suggests an approximation for m classes with multiplicities $n_1, n_2, \ldots, n_m$ of the form

$$\alpha_{n_1, \ldots, n_m} \approx \beta e^{-(\gamma_1 n_1 + \ldots + \gamma_m n_m)} . \tag{7.2}$$

For multiple classes, approximation (7.2) is convenient because we can determine the asymptotic decay rates γ_i by changing one class at a time. However, more work is needed on this.

VIII. THE ALGORITHM

In this section we describe our algorithm for numerically computing the exact tail probabilities. To compute the tail probabilities, we draw heavily on Lucantoni [35]; see [36] for a review. We model each source as a *Markovian Arrival Process* (MAP), which is a two-dimensional Markov process $\{N(t), J(t)\}$ on the state space $\{(i,j): i \geq 0, 1 \leq j \leq m\}$ with an infinitesimal generator having the structure

$$Q = \begin{bmatrix} D_0 & D_1 & 0 & 0 & \ldots \\ 0 & D_0 & D_1 & 0 & \ldots \\ 0 & 0 & D_0 & D_1 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \end{bmatrix} \tag{8.1}$$

where 0, D_0 and D_1 in (8.1) are $m \times m$ matrices; D_0 has negative diagonal elements and nonnegative off-diagonal elements; and $D \equiv D_0 + D_1$ is an irreducible infinitesimal generator. We also assume that $D \neq D_0$, which ensures that arrivals will occur. The variable $N(t)$ counts the number of arrivals in the interval $[0,t]$, while $J(t)$ gives the arrival ''phase'' at time t.

In this paper we consider the superposition of many sources. If the individual sources are characterized by matrices D_{0i} and D_{1i} for $i = 1,2,\ldots,K$, where K is the number of sources, then it can be shown [35] that the superposed system is also a MAP characterized by matrices

$$D_0 = D_{01} \oplus D_{02} \oplus \ldots \oplus D_{0K} \tag{8.2}$$

$$D_1 = D_{11} \oplus D_{12} \oplus \ldots \oplus D_{1K} \tag{8.3}$$

where $\oplus$ is the Kronecker sum. We do all computations on the superposed MAP characterized by D_0 and D_1.

As explained in [35], the *Laplace-Stieltjes transform* (LST) of the steady-state waiting-time cumulative distribution function (CDF) is given by

$$\hat{W}(s) = (\pi d)^{-1} s(1-\rho) g[sI + D(\hat{h}(s))]^{-1} e\ , \tag{8.4}$$

where e is a vector of all 1's, $\hat{h}(s)$ is service time LST, ρ is the traffic intensity (the mean arrival rate times the mean service time), $d = D_1 e$, π is the stationary probability vector of the Markov process with generator D (i.e., π satisfies $\pi D = 0$ and $\pi e = 1$), g satisfies the equations $gG = g$ and $ge = 1$, and G satisfies the equation

$$G = \int_0^\infty e^{(D_0 + D_1 G)x} dH(x)\ , \tag{8.5}$$

where $H(x)$ is the service-time CDF and

$$D(\hat{h}(s)) = D_0 + D_1 \hat{h}(s)\ . \tag{8.6}$$

Our computation is based on, first, fast and accurate computation of $\hat{W}(s)$ using (8.4) and, second, fast and accurate transform inversion to get the waiting time CDF from $\hat{W}(s)$. To do the inversion, we use the Fourier-series method in [5] along with the round-off error control procedure in [15]. The procedure in [15] also provides a self-contained accuracy check by doing the calculation twice with different round-off control parameters. This amounts to using different contours for complex inversion integration. With this procedure, we achieve high accuracy even at the tail probability of 10^{-9}.

Next we describe two techniques for greatly speeding up the computation of the LST $\hat{W}(s)$. First, note that all individual sources are two-phase sources. The Kronecker sum operations in (8.2) and (8.3) makes the superposed source a 2^K-phase source. This would prevent us from using large K, but we can take advantage of the fact that the sources are homogeneous to greatly reduce the dimensionality of the superposed source. We can define the phase of the superposed source as one plus the number of component sources in phase 1. Therefore, the total number of phases is $K + 1$ instead of 2^K. The D_0 and D_1 matrices of the new superposed sources are given in terms of the component D_{01} and D_{11} matrices of the (identical) component sources as follows:

$$\begin{aligned} (D_1)_{i,i} &= (i-1)(D_{11})_{1,1} + (K-i+1)(D_{11})_{2,2} \\ &\quad \text{for} \quad i = 1,2,\ldots,K+1 \\ (D_1)_{i,i+1} &= (K-i+1)(D_{11})_{2,1} \quad \text{for} \quad i = 1,2,\ldots,K \\ (D_1)_{i,i-1} &= (i-1)(D_{11})_{1,2} \quad \text{for} \quad i = 2,3,\ldots,K+1 \end{aligned} \tag{8.7}$$

with $(D_1)_{i,j} = 0$ for all other pairs (i,j),

$$\begin{aligned} (D_0)_{i,i+1} &= (K-i+1)(D_{01})_{2,1} \quad \text{for} \quad i = 1,2,\ldots,K \\ (D_0)_{i,i-1} &= (i-1)(D_{01})_{1,2} \quad \text{for} \quad i = 2,3,\ldots,K+1 \end{aligned} \tag{8.8}$$

with $(D_0)_{i,j} = 0$ for all other pairs (i,j) with $i \neq j$, and

$$\begin{aligned} (D_0)_{i,i} = -[&(D_0)_{i,i-1} + (D_0)_{i,i+1} \\ &+ (D_1)_{i,i} + (D_1)_{i,i-1} + (D_1)_{i,i+1}]\ . \end{aligned} \tag{8.9}$$

For the service-time distribution, we approximate the deterministic distribution by an Erlangian distribution of order k (E_k). A significant computational burden is the computation of G from the matrix integral equation (8.5). The uniformization procedure recommended in [35] for general service-time distributions can be significantly improved for Erlang distributions by noting that for the E_k distribution with mean 1 (8.5) reduces to

$$G = [I - k^{-1}(D_0 + D_1 G)]^{-k}\ . \tag{8.10}$$

Equation (8.10) can easily be solved by successive substitution, i.e.,

$$G_{n+1} = [I - k^{-1}(D_0 + D_1 G_n)]^{-k}\ , \tag{8.11}$$

where G_0 is chosen to be a stochastic matrix. Note that since we choose $k = 2^m$ (specifically 2^4 for bursty sources and 2^{10} for smooth sources), each iteration in (8.11) involves only a single matrix inversion and m matrix multiplications.

In order to compute the asymptotic parameters α and η in (1.1), we use two independent algorithms (that cross-check each other): the algorithm in [2] and the moment-based procedure in [14] and [1].

IX. CONCLUSIONS

Our first main conclusion here is that the effective-bandwidth approximation (1.3) can break down when there is a large number of independent sources. Approximation (1.3) tends to get worse as the number of sources increases, the channel utilization decreases, the buffer size decreases and the source gets further from Poisson, either more bursty or less bursty.

If the sources are more bursty than Poisson, as is anticipated for ATM networks, then the effective-bandwidth approximation (1.3) is conservative. When the approximation is bad, there may be substantially more statistical multiplexing gain than approximation (1.3) predicts. On the other hand, if the sources are less bursty than Poisson, then the effective-bandwidth approximation is no longer conservative, and may also be bad. In general, contrary to many statements, the effective-bandwidth approximation need not be conservative.

Typically, the exact tail probabilities lie between the effective-bandwidth approximation (1.3) and the true asymptote (1.1). If these two curves are very close to each other (differ by less than a factor of 10), then (1.3) is usually reasonably accurate. If the two curves are far from each other, but not too far (differ by less than 10^8), then (1.3) is bad, but (1.1) and (1.5) are reasonably accurate. Finally, if the two curves are extremely far apart (differ by more than 10^8), then even (1.1) and (1.5) are not accurate. Our second main conclusion is that, even though (1.3) may not be a good approximation, the true asymptote (1.1) often is a good approximation. Moreover, from calculations of the asymptotic parameters α and η, we have a way to estimate whether or not the approximations will be good.

A reason for the degradation of the effective-bandwidth approximation as the number n of sources increases is that the asymptotic constant α_n in (1.1) is itself asymptotically exponential in n. For sources more bursty than Poisson, α_n decrease to 0 exponentially fast, as shown in Figures 4 and 5. For sources less bursty than Poisson, α_n increases to infinity exponentially in n, as shown in Figure 12. The Poisson case is the reference case, because α_n there does not change with n. The asymptotically exponential form for α_n in (1.6) allows us to compute it approximately for arbitrary n by extrapolating based on computed values for small n. For heterogeneous sources with different multiplicities, we exploit (7.1). Having a way to approximate the asymptotic constant α is important, not only because we can use it in approximations (1.1) and (1.5), but also because the value of α indicates whether or not the approximations will be good.

Finally, in addition to gaining a better understanding of the effective-bandwidth approximation (1.3), we have provided bases for more refined analysis tools via our exact MAP/G/1 numerical algorithm, the refined approximation (1.5) [17] and the exponential relation for α_n in (1.6). We have indicated how (1.6) and (7.1) can be combined with the true asymptote (1.1) to get an approximation that is almost as simple as the effective-bandwidth approximation (1.3), but is applicable in a substantially wider region.

Acknowledgments. This work was done when David Lucantoni was at AT&T Bell Laboratories. We thank our colleagues Bharat Doshi, Ted Eckberg, Dave Houck and Pat Wirth for helpful comments.

REFERENCES

[1] J. Abate, G. L. Choudhury, D. M. Lucantoni and W. Whitt. Asymptotic analysis of tail probabilities based on the computation of moments. *Ann. Appl. Prob.*, to appear.

[2] J. Abate, G. L. Choudhury and W. Whitt. Asymptotics for steady-state tail probabilities in structured Markov queueing models. *Stochastic Models*, 10 (1994) 99-143.

[3] J. Abate, G. L. Choudhury and W. Whitt. Exponential approximations for tail probabilities in queues, I: waiting times. *Opns. Res.*, to appear.

[4] J. Abate, G. L. Choudhury and W. Whitt. Exponential approximations for tail probabilities in queues, II: sojourn time and workload. *Opns. Res.*, to appear.

[5] J. Abate and W. Whitt. The Fourier series method for inverting transforms of probability distributions. *Queueing Systems* 10 (1992) 5-88.

[6] D. Anick, D. Mitra and M. M. Sondhi. Stochastic theory of a data handling system with multiple sources. *Bell System Tech. J.*, 61 (1982) 1871-1894.

[7] A. Baiocchi. Asymptotic behavior of the loss probability of the MAP/GI/1/K queue, Part I: theory. INFOCOM Dept., University of Rome ‘‘La Sapienza,’’ 1992.

[8] A. Baiocchi, N. Blefari, M. Listanti, A. Roveri and R. Winkler. Loss performance analysis of an ATM multiplexer loaded with high-speed ON-OFF sources. *IEEE J. Sel. Areas Commun.* 9 (1991) 388-393.

[9] D. D. Botvich and N. G. Duffield, Large deviations, the shape of the loss curve, and economies of scale in large multiplexers, Dublin City University, Dublin, Ireland.

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

[11] C.-S. Chang, P. Heidelberger, S. Juneja and P. Shahabuddin. Effective bandwidth and fast simulation of ATM intree networks, IBM RC 18586, IBM T. J. Watson Research Center, Yorktown Heights, NY, 1992.

[12] G. L. Choudhury, K. K. Leung and W. Whitt. An algorithm for product-form loss networks based on numerical inversion of generating functions. *Proc. IEEE Globecom ‘94.* (1994) 1123-1128.

[13] G. L. Choudhury, K. K. Leung and W. Whitt. An inversion algorithm for loss networks with state-dependent rates. *Proc. IEEE Infocom ‘95,* 513-521.

[14] G. L. Choudhury and D. M. Lucantoni. Numerical computation of moments with application to asymptotic analysis. *Opns. Res.*, to appear.

[15] G. L. Choudhury, D. M. Lucantoni and W. Whitt. Multidimensional transform inversion with applications to the transient M/G/1 queue. *Ann. Appl. Prob.* 4 (1994) 719-740.

[16] G. L. Choudhury, D. M. Lucantoni and W. Whitt. On the effectiveness of effective bandwidths for admission control in ATM networks. *The Fundamental Role of Teletraffic in the Evolution of Telecommunication Networks, Proceedings of ITC14,* J. Labetoulle and J. Roberts, eds., 1a (1994) 411-420.

[17] G. L. Choudhury, D. M. Lucantoni and W. Whitt. Refined approximations for G/G/1 queues, in preparation.

[18] G. L. Choudhury and W. Whitt. Heavy-traffic approximations for the asymptotic decay rate in BMAP/G/1 queues. *Stochastic Models* 10 (1994) 453-498.

[19] E. Çinlar. Superposition of point processes, pp. 549-606 in *Stochastic Point Processes: Statistical Analysis, Theory and Applications*, ed. P. A. W. Lewis, Wiley, New York, 1972.

[20] C. Courcoubetis, G. Kesidis, A. Ridder, J. Walrand and R. R. Weber. Admission control and routing in ATM networks using inferences from measured buffer occupancy. *IEEE Trans. Commun.*, to appear.

[21] C. Courcoubetis and R. R. Weber. Buffer overflow asymptotics for a switch handling many traffic sources, University of Cambridge, 1994.

[22] A. Dembo and O. Zeitouni. presentation at the national meeting of the Operations Research Society of America, Boston, April 1994.

[23] A. I. Elwalid and D. Mitra. Effective bandwidth of general Markovian traffic sources and admission control of high speed networks, *IEEE/ACM Trans. Networking* 1 (1993) 329-343.

[24] A. I. Elwalid and D. Mitra. Markovian arrival and service communication systems: spectral expansions, separability and Kronecker-product forms, 1993.

[25] K. W. Fendick, V. R. Saksena and W. Whitt. Investigating dependence in packet queues with the index of dispersion for work. *IEEE Trans. Commun.*, 39 (1991) 1231-1244.

[26] R. J. Gibbens and P. J. Hunt. Effective bandwidths for the multi-type UAS channel. *Queueing Systems* 9 (1991) 17-28.

[27] P. W. Glynn and W. Whitt. Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *Studies in Applied Probability, Papers in Honour of Lajos Takács,* J. Galambos and J. Gani (eds.), Applied Probability Trust, Sheffield, England, 1994, 131-156.

[28] R. Guerin, H. Ahmadi and M. Naghshineh. Equivalent capacity and its application to bandwidth application in high-speed networks. *IEEE J. Sel. Areas Commun.* 9 (1991) 968-981.

[29] R. Guerin and L. Gün. A unified approach to bandwidth allocation and access control in fast packet-switched networks. INFOCOM ’92, Florence, Italy.

[30] H. Heffes and D. M. Lucantoni. A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance. *IEEE J. Sel. Areas Commun.* SAC4 (1986) 856-868.

[31] J. Y. Hui. Resource allocation for broadband networks. *IEEE J. Select. Areas Commun.* SAC-6 (1988) 1598-1608.

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

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

[34] A. Kuczura. The interrupted Poisson process as an overflow process. *Bell System Tech. J.*, 52 (1973) 437-448.

[35] D. M. Lucantoni. New results on the single server queue with a batch Markovian arrival process. *Stochastic Models* 7 (1991) 1-46.

[36] D. M. Lucantoni. The BMAP/G/1 queue: a tutorial. *Models and Techniques for Performance Evaluation of Computer and Communication Systems*, L. Donatiello and R. Nelson, eds., Springer-Verlag, 1993, 330-358.

[37] D. M. Lucantoni, G. L. Choudhury and W. Whitt. The transient BMAP/G/1 queue. *Stochastic Models* 10 (1994) 145-182.

[38] M. F. Neuts. The caudal characteristic curve of queues. *Adv. Appl. Prob.* 18 (1986) 221-254.

[39] J. W. Roberts. *Performance Evaluation and Design of Multiservice Networks, COST 224 Final Report,* Commission of the European Communities, Luxembourg, 1992.

[40] A. Simonian and J. Guilbert. Large deviations approximation for fluid queues fed by a large number of on-off sources, preprint, 1994.

[41] K. Sohraby. On the asymptotic behavior of heterogeneous statistical multiplexer with applications. *IEEE INFOCOM ‘92,* Florence, Italy.

[42] K. Sohraby. On the theory of general on-off sources with applications in high speed networks. *IEEE INFOCOM ‘93,* San Francisco, CA.

[43] K. Sriram and W. Whitt. Characterizing superposition arrival processes in packet multiplexers for voice and data. *IEEE J. Sel. Areas. Commun.* SAC-4 (1986) 833-846.

[44] D. N. C. Tse, R. G. Gallager and J. N. Tsitsiklis. Statistical multiplexing of multiple time-scale Markov streams. Laboratory for Information and Decision Sciences, M.I.T. 1994.

[45] W. Whitt. Tail probabilities with statistical multiplexing and effective bandwidths for multi-class queues. *Telecommunication Systems*, 2 (1993) 71-107.

Gagan L. Choudhury received the B.Tech. degree in Radio Physics and Electronics from the University of Calcutta, India in 1979 and the MS and Ph.D. degrees in Electrical Engineering from the State University of New York (SUNY) at Stony Brook in 1981 and 1982, respectively. Currently he is a technical manager at the AT&T Bell Laboratories, Holmdel, New Jersey, USA. His group is responsible for the performance assessment of various AT&T products and services based on modeling and simulation. His main research interest is in the development of multi-dimensional numerical transform inversion algorithms and their application to the performance analysis of telecommunication and computer systems.

David M. Lucantoni is Vice-President and Chief Technical Officer for IsoQuantic Technologies, LLC. IQ Tech develops network-level software solutions for the wireless and wireline telecommunications industry. Dr. Lucantoni received a B.S in Mathematics, an M.S. in Statistics and a Ph.D. in Operations Research from the University of Delaware in 1976, 1978, and 1981, respectively. He joined AT&T Bell Laboratories in 1981 and over the next 13 years worked on the performance analysis of various telecommunication systems. He has published over forty professional papers ranging from multiplexer design to broadband congestion control algorithms to state-of-the-art solution techniques to complex stochastic models. He was the co-recipient of the IEEE Steven O. Rice Prize paper award in the area of Communication Theory in 1986 and was promoted to Distinguished Member of the Technical Staff in 1987.

Following a position at Motorola’s Satellite Communications Division as a performance analyst for the IRIDIUM(tm) Low Earth Orbit satellite system, Dr. Lucantoni co-founded IsoQuantic Technologies, LLC, in 1994.

Ward Whitt received the A.B. degree in mathematics from Dartmouth College in 1964 and the Ph.D. degree in operations research from Cornell University in 1969. He taught in the Department of Operations Research at Stanford University in 1968-1969, and in the Department of Administrative Sciences at Yale University from 1969 to 1977. Since 1977, he has been employed by AT&T Bell Laboratories. He currently works in the Network Services Research Laboratory, Murray Hill, NJ. His research interests include queueing theory, stochastic processes, stochastic models in telecommunications and numerical inversion of transforms.

Dr. Whitt is a member of the Operations Research Society of America and the Institute of Mathematical Statistics.

Figure 1. A comparison of approximations and exact values of the probability of buffer overflow (approximated by the tail probability $P(W > x)$) as a function of the buffer size x for the example in Section 2.

Figure 2. The decay rate η_n in (1.1) as a function of the number n of sources (without scaling the sources) for the example in Section 2.

Figure 3. The asymptotic constant α_n in (1.1) as a function of the number n of sources (without scaling the sources) for the example in Section 2.

Figure 4. The asymptotics constant α_n as a function of the number n of sources (with scaling) for the example in Section 2.

Figure 5. The asymptotic constant α_n as a function of the number n of sources (with scaling) for four examples in Section 3.

Figure 6. A comparison of approximations and exact values of the probability of buffer overflow (approximated by the tail probability $P(W > x)$ as a function of the buffer size x for the example with $n = 2$ in Section 4.

Figure 7. A comparison of approximations and exact values of the probability of buffer overflow (approximated by the tail probability $P(W > x)$ as a function of the buffer size x for the example with $n = 60$ in Section 4.

Figure 8. A comparison of approximations and exact values of the probability of buffer overflow (approximated by the workload-tail probability) as a function of the buffer size x for the example with $n = 24$ in Section 2 and higher buffer capacity 6000 in Section 5.

Figure 9. A comparison of approximations and exact values of the probability of buffer overflow (approximated by the workload-tail probability) $n = 24$ in Section 2 and lower buffer capacity 60 in Section 5.

Figure 10. A comparison of approximations and exact values of the probability of buffer overflow (approximated by the workload-tail probability) $n = 24$ in Section 2, higher buffer capacity 6000 and higher burstiness in Section 5.

Figure 11. A comparison of approximations with exact values of the probability of buffer overflow (approximated by the workload tail probability) as a function of the buffer size x for the example in Section 6 with E_2-renewal-process sources, $n = 24$ and buffer size 8.

Figure 12. The asymptotic constant α_n as a function of the number n of sources (with scaling for the E_2 examples in Section 6.

Method of Computation	Number of Sources Allowed for Buffer Size $x = 600$	Buffer Size Required to Support $n = 24$ Sources
exact	24	600
effective-bandwidth approximation (1.3)	12	1146
(1.1) and (1.5)	25	530
Poisson	78	20
Average-Rate engineering	80	not applicable
peak-rate engineering	7	not applicable

Table 1.
A comparison of different methods for determining (i) the number of sources for the fixed buffer size $x = 600$ and (ii) the buffer size required to support $n = 24$ sources for the example in Section 2.

Even though ATM seems to be clearly the wave of the future, one performance analysis indicates that the combination of stringent performance requirements (e.g., 10^{-9} cell blocking probabilities), moderate-size buffers and highly bursty traffic will require that the utilization of the network be quite low. That performance analysis is based on asymptotic decay rates of steady-state distributions used to develop a concept of *effective bandwidths* for connection admission control. However, we have developed an *exact numerical algorithm* that shows that the effective-bandwidth approximation can overestimate the target small blocking probabilities by several orders of magnitude when there are many sources that are more bursty than Poisson. The *bad news* is that the appealing simple connection-admission-control algorithm using effective bandwidths based solely on tail-probability asymptotic decay rates may actually not be as effective as many have hoped. The *good news* is that the statistical multiplexing gain on ATM networks may actually be higher than some have feared. For one example, thought to be realistic, our analysis indicates that the network actually can support *twice as many sources* as predicted by the effective-bandwidth approximation. That discrepancy occurs because for a large number of bursty sources the asymptotic constant in the tail probability exponential asymptote is extremely small. That in turn can be explained by the observation that the asymptotic constant decays exponentially in the number of sources when the sources are scaled to keep the total arrival rate fixed. We also show that the effective-bandwidth approximation is *not always conservative*. Specifically, for sources less bursty than Poisson, the asymptotic constant grows exponentially in the number of sources (when they are scaled as above) and the effective-bandwidth approximation can greatly underestimate the target blocking probabilities. Finally, we develop *new approximations* that work much better than the pure effective-bandwidth approximation.

Figure 13. A comparison of approximations with exact values of the probability of buffer overflow (approximated by the workload tail probability) as a function of the buffer size x for the example in Section 6 with E_2-renewal-process sources, $n = 24$ and buffer size 100.

I. Introduction and Summary

Great energy is being devoted to studying the promising new *asynchronous transfer mode* (ATM) technology for supporting multi-service high-speed communication networks; e.g., see Roberts [39]. As indicated in [39], interest in ATM is stimulated by two factors: first, by new technology making it possible to transmit and switch at very high bandwidths and, second, by growing demand for more sophisticated and powerful communication services.

Even though ATM seems to be well on its way to widespread use, a performance analysis based on asymptotic decay rates of steady-state distributions indicates that the combination of stringent performance requirements (e.g., 10^{-9} cell blocking probabilities), moderate-size buffers and highly bursty traffic associated with these new services will necessitate operating the networks at very low utilizations. (The basic model represents an ATM switch receiving fixed-size ATM cells from several sources and transmitting them over an output channel in a first-in-first-out fashion. The basic question is: How many sources can be admitted for a fixed buffer size with a specified small cell blocking probability?) A main message of this paper is that *ATM networks may be able to achieve higher utilizations than that asymptotic analysis indicates.* In other words, there seems to be more potential for statistical multiplexing gain.

The approximation based on asymptotic decay rates of tail probabilities is very appealing because it supports a concept of *effective bandwidths.* From an engineering perspective, the notion of effective bandwidths is very natural. The idea is to assign each source an effective bandwidth requirement, and then consider any subset of sources feasible (admissible) if the sum of the required effective bandwidths is less than the total available bandwidth. Thus, a suitable notion of effective bandwidths could go a long way toward solving the connection admission control problem. For instance, once effective bandwidths have been assigned, we can approach engineering and design problems using multi-rate loss network models, e.g., as in Choudhury, Leung and Whitt [12,13] and references therein.

A large-deviations asymptotic analysis provides strong support for the simple effective-bandwidth procedure, because it shows that it is asymptotically correct as the buffer size gets large and the tail probabilities get small, and because it provides a basis for assigning actual effective bandwidth values to different sources (voice, data, video, etc.); e.g., see Hui [31], Gibbens and Hunt [26], Guerin, Ahmadi and Naghshineh [28], Kelly [32], Sohraby [41], [42], Chang [10], Whitt [45], Elwalid and Mitra [23], Kesidis, Walrand and Chang [33], Glynn and Whitt [27], Courcoubetis, Kesidis, Ridder, Walrand and Weber [20], and Chang, Heidelberger, Juneja and Shahabuddin [11].

The additive nature of effective bandwidths is clearly appropriate if we let the effective bandwidths be either the source *peak rates* or the source *average rates.* However, it seems intuitively clear that working with peak rates is far too conservative, while working with average rates is far too optimistic. (We show this later in our examples.) Most of the recent work has been aimed at finding appropriate effective bandwidths in between the peak and average rates.

Unfortunately, however, from the outset, teletraffic engineering experience suggests that there may be a flaw in the effective bandwidth concept because *it corresponds to having no traffic smoothing from multiplexing.* In particular, it is known that when many separate bursty sources are multiplexed (superposed), the total is less bursty than the components; i.e., there is a basis for more statistical multiplexing gain than with Poisson sources. This is theoretically supported by the classical limit theorem stating that superpositions of arrival processes, suitably scaled, converge to a Poisson process as the number of component arrival processes increases; e.g., see Çinlar [19]. See Heffes and Lucantoni [30], Sriram and Whitt [43] and Fendick, Saksena and Whitt [25] for related performance studies. In contrast, with the effective-bandwidth approximation, the burstiness of n superposed independent and identically distributed sources is the same as for a single source (e.g., see p. 76 of [45]). This implies that the effective-bandwidth approximation will predict greater congestion for any fixed arrival rate than it should.

Nevertheless, there is a case for the effective bandwidths, because previous teletraffic analysis (such as in [25], [30], [43]) did not focus on extremely small loss probabilities such as 10^{-9}. Such very small loss probabilities naturally suggest that appropriate asymptotics should provide what we want. And the asymptotic analysis associated with effective bandwidths indicates no traffic smoothing.

The question, then, is what will actually happen in real systems? Will there be significant traffic smoothing or not? Is the asymptotic analysis supporting effective bandwidths sufficiently accurate or is it not?

These questions have plagued researchers in recent years because the natural models are very difficult to analyze. It is difficult to calculate very small tail probabilities in models with many independent bursty sources, by either exact analytical formulas or by computer simulation.

Our main contribution is a new algorithm for a queueing model that enables us to compute the desired small tail probabilities exactly when there are many bursty sources. From this exact analysis, we find that in some cases the effective-bandwidth approximation is excellent, but in other cases (which we think are realistic), it is not. Thus, *we conclude that the notion of effective bandwidths based directly on large-deviation asymptotics may not be effective for connection admission control.*

However, from an engineering perspective, the notion of effective bandwidths remains very appealing. Thus it is natural to look for modifications of the proposed effective bandwidth approximation which will work well; e.g., as in Guerin and Gun [29]. It is important to note that our analysis does not rule out the general notion of effective bandwidths. Instead, our analysis only indicates that there may be serious difficulties with the implementation based directly on a particular kind of asymptotics, without refinements.

A second main contribution of this paper is to develop modifications of the basic effective-bandwidth approximation for tail probabilities that are more accurate (namely, (1.1), (1.5) and (1.7) below). However, these new approximations do *not* support the simple notion of effective bandwidths discussed above. Nevertheless, the approximate tail probabilities can be used for admission control.

The key to our analysis here is the exact numerical solution of the $\sum_{i=1}^{n} \mathrm{G}_i/\mathrm{G}/1$ queueing model, which has a single server, unlimited waiting room, the first-in first-out service discipline and i.i.d. (independent and identically distributed) service times that are independent of a superposition arrival process. The arrival process is the superposition of n independent component general arrival processes. Our algorithm permits these component arrival processes to be heterogeneous *Markovian arrival processes* (MAPs), as in Lucantoni [35,36]; see Section 8 here. (The general theory also allows batch arrivals, but in this paper we consider only single arrivals.) Since superpositions of independent MAPs are again MAPs, it suffices to consider the MAP/G/1 queue. However, the computational effort increases when the number of sources increases. Our algorithm is based on Lucantoni [35]. *A major new idea is the combination of the matrix-analytic results in [35] with numerical transform inversion.* In particular, we use the Fourier-series method in Abate and Whitt [5]. We also use a scheme for accuracy enhancement for computing small probabilities as described in [15]. Moreover, we use a number of techniques for speeding up computation, in order to treat a large number of sources. These techniques are described in Section 8. For the computation of the asymptotic parameters we use algorithms in [1], [2], [14].

We have also extended the numerical transform inversion algorithms to multidimensional transforms [15] and applied the multidimensional inversion to calculate *transient* performance measures in the MAP/G/1 queue [37]. Thus, we are in a position to study the *transient behavior* as well as the steady-state behavior. With the ability to calculate time-dependent tail probabilities, we can study whether or not transient analysis is needed in the admission control problem. Here, however, we only discuss steady-state behavior.

In our examples here, we consider only homogeneous sources, but in Section 7 we also indicate how to treat large systems with heterogeneous sources approximately. We primarily consider sources that are more bursty than a Poisson process, in particular, *Markov modulated Poisson processes* (MMPPs) with two-state Markov environment processes. (The MMPP is a Poisson process whose rate is itself a continuous-time Markov chain.) For the main example, we consider on-off sources in which the arrival rate in one environment state is zero, i.e., interrupted Poisson processes, which are known to be renewal processes having hyperexponential interarrival-time distributions [34]. These bursty source models are often used to model ATM traffic. They can represent quite bursty traffic. We have also obtained similar results for other (non-renewal) MMPP sources and multiple numbers of more than one kind (two or three) of MMPP sources.

We also consider sources that are *less bursty* than a Poisson process. In particular, we consider renewal processes with E_2 (Erlang of order 2) interarrival times. Such less bursty sources might appear in ATM networks as a consequence of traffic shaping at the edge of the network. Contrary to conventional wisdom, we show that the effective-bandwidth approximation *underestimates* the exact tail probabilities with these less bursty sources, so that *the effective bandwidth approximation need not be conservative.*

As a surrogate for the steady-state blocking probability with a finite buffer, we consider the tail probability of the steady-state waiting time in our infinite-capacity queue. When the service times are deterministic with mean 1, as is the case with ATM cells (with the appropriate time units), the least integer above the steady-state waiting time coincides with the steady-state queue length or buffer content seen by an arrival. (We can also directly compute the queue-length distribution in the MAP/G/1 queue.) We are thus approximating the steady-state blocking probability in the finite-capacity system by the steady-state probability that the buffer content exceeds that capacity level in the infinite-capacity system at an arrival epoch.

Hence, we let W be the steady-state waiting time until beginning service and we focus on the tail probabilities $P(W > x)$. It turns out that in considerable generality these waiting-time tail probabilities are *asymptotically exponential,* i.e.,

$$P(W > x) \sim \alpha e^{-\eta x} \text{ as } x \to \infty , \tag{1.1}$$

where η is a positive constant called the *asymptotic decay rate,* α is a positive constant called the *asymptotic constant,* and $f(x) \sim g(x)$ as $x \to \infty$ means that $f(x)/g(x) \to 1$ as $x \to \infty$; see Abate, Choudhury and Whitt [2]. In standard single-source examples the approximation provided by (1.1) is often remarkably accurate [2]–[4]. Moreover, numerical experience indicates that for bursty sources the approximation provided by (1.1) with α replaced by 1 often tends to be conservative, i.e.,

$$P(W > x) \le e^{-\eta x} \text{ for all } x . \tag{1.2}$$

This is a basis for the *simple one-parameter approximation*

$$P(W > x) \approx e^{-\eta x} \tag{1.3}$$

The asymptotic decay rate η in (1.1)–(1.3) is the basis for the concept of effective bandwidths. The simple one-parameter approximation (1.3) is appealing because the key parameter η in (1.1)–(1.3) is relatively easy to determine, exactly or approximately. Indeed, it is typically much easier to obtain than the asymptotic constant α. Moreover, the resulting admission-control algorithm using effective bandwidths based on (1.3) is remarkably simple. Hence, *we call* (1.3) *the effective-bandwidth approximation*.

It is important to note that much of the effective-bandwidth literature has not arrived at (1.3) via (1.1). The large-deviation result supporting (1.3) is the weaker limit

$$x^{-1} \log P(W > x) \to -\eta \text{ as } x \to \infty . \tag{1.4}$$

General sufficient conditions for (1.4) are given in Chang [10] and Glynn and Whitt [27]. Clearly (1.1) implies (1.4), but not conversely. Indeed, the weaker limit (1.4) yields no information about the asymptotic constant α in (1.1).

We have indicated that most of the effective-bandwidth literature has arrived at approximation (1.3) via the limit (1.4) or related bounds such as (1.2). However, several people have focused on (1.1) too, namely, Abate et al. [2], Baiocchi [7], Elwalid and Mitra [23,24] and Neuts [38]. Given (1.1), (1.3) is a convenient simple approximation because α is much harder to obtain than η, and because (1.3) is consistent with the notion of effective bandwidths.

As discussed in [3], for standard single-source models approximation (1.3) is often a reasonable substitute for approximation (1.1)–(1.4) when we are primarily interested in percentiles of the distribution, rather than the probabilities themselves. If the asymptotic constant α in (1.1) is not too different from 1, then it plays a relatively small role in higher percentiles.

It should be noted that the asymptotic decay rate η in (1.1) is identical to the asymptotic decay rate for the blocking probability in the corresponding finite-buffer model as the buffer size gets large; see Baiocchi [7]. Hence, our analysis of the infinite-capacity model is directly relevant to the finite-capacity model with large buffer sizes.

In order to do better than (1.1), without calculating the exact tail probabilities, we have also developed a *refined three-term approximation* of the form

$$P(W > x) \approx \alpha_1 e^{-\eta_1 x} + \alpha_2 e^{-\eta_2 x} + \alpha_3 e^{-\eta_3 x} , \tag{1.5}$$

where α_1 and η_1 are the asymptotic constant and asymptotic decay rate in (1.1), while α_2, α_3 η_2 and η_3 are chosen to match the probability of delay $P(W > 0)$ and the first three moments EW, $E(W^2)$ and $E(W^3)$, with η_2 and η_3 required to satisfy $\eta_1 \le \min\{\eta_2, \eta_3\}$; see Choudhury, Lucantoni and Whitt [17].

We thus have four ways to ''calculate'' the tail probability $P(W > x)$: (1.3), (1.1), (1.5) and the full numerical algorithm. Each successive method tends to be more accurate, but each successive method requires substantial more computational effort. The effective-bandwidth approximation (1.3) is by far the easiest to compute, with there being virtually no limit to the size of the model; e.g., see [45].

Using our exact numerical algorithm for the MAP/G/1 queue, we have investigated the three approximations in (1.3), (1.1) and (1.5). For many standard single-source models, the refinement in (1.1) is remarkably accurate even for x not too large, e.g., for the 90^{th} percentile of the distribution. For standard single-source models for which (1.1) is excellent at the 90^{th} percentile and beyond, (1.5) typically is excellent for the entire distribution.

However, we have found that *the story changes dramatically when the arrival process is the superposition of many component processes.* Then the effective bandwidth approximation (1.3) can perform very badly. It is even possible for (1.1) and (1.5) to perform badly, but there is a substantial region where (1.3) performs badly and (1.1) performs well. This is the basis for our new improved approximations beyond (1.3).

We emphasize that, for the model being considered, the limits in (1.1) and (1.4) do indeed hold, and we are indeed interested in very small tail probabilities, and relatively large x. However, there is a serious difficulty (evidently pointed out for the first time here): *The asymptotic constant* α *in (1.1) can be very different from 1 when the number of component arrival processes is large.*

We also explain *why* the asymptotic constant α can be very different from 1 with many sources. To do so, we consider the case of n identical sources with fixed total rate. As n increases, the total rate is kept fixed by properly scaling the individual streams. With this structure, it is known that the asymptotic decay rate η in (1.1) is actually independent of n. We give numerical examples showing that the asymptotic constant with n sources (and this scaling), α_n, is itself *asymptotically exponential* in n, i.e.,

$$\alpha_n \sim \beta e^{-n\gamma} \text{ as } n \to \infty , \tag{1.6}$$

where $\beta > 0$ and, for sources more bursty than Poisson, $\gamma \ge 0$, while for sources less bursty than Poisson, $\gamma \le 0$. Moreover, $|\gamma|$ in (1.6) tends to be larger when the burstiness gets further from Poisson and the traffic intensity decreases, which is a likely operating condition for ATM networks.

Combining (1.1) and (1.6), we obtain the refined asymptotic approximation (with the scaling for (1.6))

$$P(W > x) \approx \beta e^{-n\gamma} e^{-\eta x} . \tag{1.7}$$

Our numerical results supporting (1.6) and (1.7) naturally suggest a limit theorem, on which we were working but others have delivered. The idea is still to consider large-deviation asymptotics, but now letting both n and x go to infinity. Since this paper was originally submitted, large deviation limits supporting (1.6) and (1.7) have been reported by several authors, the first known to us being A. Dembo and O. Zeitouni (talk at National Meeting of the Operations Research Society of America, Boston, April 1994). Large-deviation papers supporting (1.6) and (1.7) have been written by Botvich and Duffield [9], Courcoubetis and Weber [21], Simonian [40] and Tse, Gallager and Tsitsiklis [44]. To state their result, let W^n be the steady-state waiting time with n sources. Paralleling (1.4), their result is

$$n^{-1} \log P(W^n > nx) = I(x) \text{ as } n \to \infty , \tag{1.8}$$

for an appropriate function $I(x)$, yielding the approximation

$$P(W^n > x) \approx e^{-nI(x/n)} . \tag{1.9}$$

Clearly (1.8) provides strong theoretical support for (1.6) and (1.7), but (1.8) is substantially weaker than (1.1) and does not yield the asymptotic constant β in (1.6) and (1.7). Nevertheless, the form of the limit function $I(x)$ provides useful insights, as can be seen from the references.

We propose (1.7) as a basis for developing useful approximations for the tail probabilities when there are large numbers of each of a few types of sources. In particular, we suggest using (1.7) to estimate β and γ by extrapolating using (1.6) and exact calculations for small n; see Section 7 for further discussion.

Approximation (1.7) can be regarded as an approximation to the true asymptote (1.1). Our numerical experience indicates that for sources more bursty than Poisson the true asymptote (1.1) is a good approximation at 10^{-9} if α is not too small, e.g., $\alpha \geq 10^{-4}$. For a wide range of problems, this is the region of interest. Our experience also indicates that the true asymptote (1.1) is a poor approximation at 10^{-9} if α is too small, e.g., $\alpha \leq 10^{-8}$. For estimated values of α in between, e.g., $10^{-8} < \alpha < 10^{-4}$, the true asymptote is only moderately good as an approximation. In this region, we propose using the true asymptote with α replaced by 10^{-4} as a rough conservative heuristic approximation. For sources less bursty than Poisson, (1.1) tends to be a good (bad) approximation at a target blocking probability 10^{-9} if $\alpha < 10^4$ ($\alpha > 10^8$).

Here is how the rest of the paper is organized. In Section 2 we consider a specific example with bursty sources to demonstrate that the effective-bandwidth approximation (1.3) can perform poorly. In Section 3 we consider the case of n identical sources scaled so that the total rate is independent of n, and give examples supporting (1.6). Next in Section 4 we present modifications of the example in Section 2 with different numbers of sources in which the two asymptotic approximations in (1.1) and (1.3) are, first, both good and, second, both bad. These cases seem identifiable by looking at the observed value of the asymptotic constant α. In Section 5 we investigate the impact of changing other variables. We show that the asymptotic approximations tend to get worse as the number of sources increases, the buffer size decreases, the channel utilization decreases, the target blocking probability increases, and the source gets further from Poisson, either more bursty or less bursty.

In Section 6 we consider sources less variable than Poisson. For these sources, we show that the effective-bandwidth approximation need not be conservative; i.e., (1.2) does not hold. In Section 7 we show how to do approximations in large systems with heterogeneous sources. In Section 8 we briefly describe the computational algorithm. Finally, in Section 9 we state our conclusions.

II. An Example where the Effective-Bandwidth Approximation Fails

In this section we consider an example with homogeneous on-off sources. (The on-off property is not essential for our model or for the conclusions we draw from it. A source can have multiple states, each characterized by a different arrival rate. We use such sources in later examples.) Our primary purpose is to show that the effective-bandwidth approximation (1.3) can be a bad approximation for a realistic example and can lead to seriously underestimating the number of sources the queue can accommodate.

We let the service times be i.i.d. with mean 1. Thus the unit of time is the time required to transmit one ATM cell. We let the service time have an Erlang distribution of order 16 (E_{16}), which has *squared coefficient of variation* (SCV, variance divided by the square of the mean) $c_s^2 = 1/16$. This distribution is similar to the deterministic (D) distribution of ATM cells, but better behaved for numerical calculations. (It is possible to do calculations for D distributions, but it requires more computational effort to achieve comparable precision.) In fact, we have computed using the E_{1024} distribution and have found for the examples in this paper with MMPP sources that there is little difference between E_{16} and E_{1024}, supporting our hypothesis that D is well approximated by E_{16}.

The on-off source has exponentially distributed on and off periods. During the on periods, cells arrive according to a Poisson process; during the off periods there are no arrivals. Each on-off source is characterized by three parameters, the mean on period ω, the mean off period ζ and the (peak) rate during the on period, p. Here we let the on and off periods have means $\omega = 436.36$ and $\zeta = 4363.63$, respectively. We let the peak rate be $p = 0.1375$. The overall average rate is thus $\lambda = p\omega/(\omega + \zeta) = 0.0125$. The ratio of peak to mean rates is thus $p/\lambda = 11.0$. The mean number of cells during an on period is $p\omega = 60.0$.

We feel that the level of source burstiness considered above can happen quite naturally in bursty data sources. Furthermore, the example is chosen mainly to illustrate that bad behavior can occur. Later we discuss how the basic behavior of superposed sources changes with change in burstiness, buffer size and other parameters.

The above on-off source (an interrupted Poisson process) can also be characterized as a renewal process with a hyperexponential (H_2, mixture of two exponentials) interarrival time distribution [34]. The mean interarrival time for one source is thus $1/\lambda = 80$ and the SCV is $c_a^2 = 100.2$. Scaled to have mean 1, the first four moments of an interarrival time are: $m_1 = 1$, $m_2 = 101.2$, $m_3 = 1.68 \times 10^4$ and $m_4 = 3.73 \times 10^6$. This is another way to see that the individual sources are quite bursty. (For a simple exponential random variable with mean 1, the n^{th} moment is $n!$)

Since the arrival process is a renewal process, The SCV coincides with the normalized asymptotic variance or *limiting index of dispersion for counts* for the counting process $N(t)$, i.e.,

$$I_c(\infty) \equiv \lim_{t\to\infty} \frac{\mathrm{Var}\, N(t)}{EN(t)} = c_a^2 = 100.2 , \tag{2.1}$$

see [25,43]. The value $I_c(\infty) = 100.2$ is large, so that each source is indeed quite bursty.

We let the buffer size be 600 cells, which falls in the range considered by switch manufacturers. The buffer size 600 is ten times the mean number of cells in an on period; i.e., the buffer contains 10 bursts. The quality of the approximations depends strongly on the buffer size. (We discuss this further in Section 5.) Currently buffer size is limited by the availability of fast memory. We regard 600 as a representative moderate buffer size.

As is customary in ATM, we let the target blocking probability be 10^{-9}. Hence, we choose the number of sources so that

$$P(W > 600) \approx 10^{-9} . \tag{2.2}$$

Our exact numerical results indicate that (2.2) is attained with $n = 24$ sources, yielding a utilization of 30%.

Figure 1 displays the exact tail probabilities $P(W > x)$ and the approximations (1.1), (1.3) and (1.5). In addition, Figure 1 displays the Poisson approximation, which is the exact tail probability in the $M/E_{16}/1$ queue, computed by numerical transform inversion. The tail probability $P(W > x)$ is our approximate probability of buffer overflow with x being the buffer size.

Insert Figure 1 about here.

The (exact) asymptotic parameters in (1.1) are $\eta = 0.01809$ and $\alpha = 1.453 \times 10^{-5}$. The three-term approximation is also fully determined, based on the exact probability of delay $P(W > 0) = 0.4242$ and first three moments $EW = 0.6924$, $E(W^2) = 8.573$ and $E(W^3) = 579.3$.

Of course, there is the question of numerical accuracy of the exact results reported in Figure 1. We verified them using a built-in accuracy check in our numerical procedures, as explained in [15, 37] and Section 8 here. In the related case of 64 sources with exponential service times, numerical results by Elwalid and Mitra [24] confirmed our results.

Figure 1 indicates that (1.1) and (1.5) are still in error by a factor of 3.9 at $x = 600$. This error is relatively small, though, compared to the error in (1.3), which is by a factor of 10^5. The Poisson approximation obviously is even worse than (1.3).

It is interesting to consider Figure 1 in relation to previous discussions of ‘‘cell-level congestion’’ and ‘‘burst-level congestion,’’ e.g., in [39]. To a large extent, Figure 1 indicates two nearly linear regions for the exact curve. As discussed on p. 19 of [39], there is an initial period of steep nearly linear decline corresponding to ‘‘cell-level congestion’’ and a later period of more gradual nearly linear decline corresponding to ‘‘burst-level congestion.’’ This burst level congestion corresponds to the true asymptote (1.1). However, the transition between these two regimes is smooth and fairly long. As in [39], much of the ATM literature considers *separate models* to represent cell-level and burst-level congestion. In contrast, we represent *both within a single model*. Indeed, we consider precisely the multiple-MMPP-source model described as difficult on p. 185 of [39]. In this context, a major contribution here is to point out that the intercept with the y-axis of the asymptote, which is the asymptotic constant α in (1.1), may well be very small.

Now suppose, instead, that we use approximation (1.3). Let η_n be η in (1.1) as a function of n (without any rescaling of individual sources). As indicated above, $\eta_{24} = 0.01809$, but η increases as n (and ρ) decreases. Figure 2 displays η_n as a function of n (without any rescaling of individual sources). It turns out that

$$e^{-600\eta_n} \leq 10^{-9} \text{ for all } n \leq 12 , \tag{2.3}$$

but not for $n \geq 13$; $\eta_{12} = 0.03749$ and $\eta_{13} = 0.03354$. Hence, using the effective-bandwidth approximation (1.3), we conclude that the queue can accommodate only 12 sources with criterion (2.2). Hence, (1.3) indeed significantly underestimates the capacity. *The actual capacity is two times that predicted by (1.3).*

Insert Figure 2 about here.

Table 1 compares seven different procedures for determining the number of sources that can be supported:

(i) The exact tail probabilities

(ii) the effective-bandwidth approximation (1.3)

(iii) the full asymptotic approximation (1.1)

(iv) the three-term approximation (1.5)

(v) the Poisson approximation

(vi) average-rate engineering

(vii) peak-rate engineering.

Insert Table 1 about here.

From Figure 1 and Table 1, we see that even though the true asymptote is not too accurate for the tail probability (being off by a factor of 4), it produces a good estimate for the number of sources (being off by only one).

Since the average rate is $\lambda = 0.0125$, if we could *size by average rate*, then the queue could accommodate 80 sources. Since the peak rate is $p = 0.1375$, if we *sized by peak rate*, then the queue can accommodate 7 sources. The numbers 7 and 80 help put the realized improvement from 12 to 24 going from approximation (1.3) to the exact numerical result in perspective.

There is a simple explanation for the poor performance of approximation (1.3). For $n = 24$, the asymptotic constant in (1.1) is $\alpha_{24} = 1.453 \times 10^{-5}$. As can be seen from Figure 1, in this case the one-term asymptotic approximation $\alpha e^{-\eta x}$ based on (1.1) is a good approximation, but replacing α by 1 introduces a large error.

In customary models with a single source, the asymptotic constant α is usually not too different from 1. Hence, it is interesting to see how α_n depends on n (again, without rescaling the individual sources). This is shown in Figure 3. There we see that α_n declines from $\alpha_1 = 0.340$ to a minimum value of $\alpha_{11} = 6 \times 10^{-8}$ and then increases toward 1 again as the traffic intensity approaches 1. From heavy-traffic theory, we anticipate that $\alpha_n \to 1$ as $\rho_n \to 1$. Assuming that α_n stays well above the target value 10^{-9} for all n, as it does in Figure 3, we can anticipate that the asymptote (1.1) will be a reasonably good approximation, but we clearly cannot simply replace the asymptotic constant α by 1.

Insert Figure 3 about here.

Instead of fixing the buffer size at 600 and asking how many sources the queue can accommodate, we could instead fix the number of sources at 24 and ask what size buffer is required. The exact numerical results indicate a buffer size of 600. The approximations (1.1) and (1.5) indicate that a buffer size of 530 is required, approximation (1.3) indicates 1146 and the Poisson approximation indicates only about 20. These results are also displayed in Table 1. This is another way to look at the weakness of the effective-bandwidth and Poisson approximations. This view also shows that approximations (1.1) and (1.5) are reasonably good.

As shown in [4], the asymptotics (1.1) for the steady-state waiting time are closely related to corresponding asymptotics for other steady-state quantities, such as the workload, sojourn time and queue length (at arrivals and at arbitrary times). For this example, the queue length decay rate parameter σ in [4] is $\sigma = 0.9821$. Although we consider only the waiting time here, our experience indicates that a detailed analysis of one of the other steady-state distributions tells essentially the same story. For the example considered in this section, if we look at the queue length at an arbitrary time (instead of at the arrival instant) then both the exact value and approximation (1.1) drop roughly by a factor of 3, but the effective-bandwidth approximation (1.3) remains unchanged (another weakness of (1.3)). However this further difference by a factor of 3 is small compared to the already-established difference by a factor of 10^5.

III. Asymptotics for the Asymptotic Constant in Scaled Superposition Processes

As shown in [10], [18], [23], [24], [27], [45], the asymptotic decay rate η_n in the $\sum_{i=1}^{n} G_i/G/1$ model with the superposition of n i.i.d. arrival processes is independent of n when we fix the total arrival rate by scaling the component arrival processes. In particular, let $\{N(t) : t \geq 0\}$ be the arrival counting process with only one source. The proper scaling is achieved with n sources by letting each component arrival process be distributed as $\{N(t/n) : t \geq 0\}$. If the arrival rate of $N(t)$ is λ, then the arrival rate of $N(t/n)$ is λ/n, so that the total arrival rate with n sources is λ for all n. (This scaling is also discussed previously in [43] and 25].)

Of course, when we add sources in a real network, we do not do any rescaling, so we did not do any rescaling in Section 2. However, since η_n is independent of n with this rescaling, the rescaling helps us understand what is happening with the various approximations. As discussed in [43], a key theoretical reference point is the fact that, with the scaling, the superposition process approaches a Poisson process as $n \to \infty$. Since η_n does not change with n, we see that the two limits $x \to \infty$ and $n \to \infty$ do not interchange. This is a source of our difficulties.

Our numerical experience indicates that with this rescaling as $n \to \infty$ the asymptotic constant α_n in (1.1) approaches 0 for sources more bursty than Poisson and approaches infinity for sources less bursty than Poisson. More precisely, α_n appears to decay or grow exponentially as in (1.6) as $n \to \infty$. Numerical evidence supporting (1.6) is given in Figures 4 and 5 for sources more bursty than Poisson. There we display α_n as a function of n in log scale for different examples. Figure 4 displays α_n as a function of n for the example in Section 2, while Figure 5 displays α_n as a function of n in four other examples. For Figure 4, the reference case is the case with $n = 24$ sources and $\rho = 0.30$. All other cases in Figure 4 are rescaled to have this same ρ. In each example we rescale the arrival processes as n changes, so that η_n is independent of n. The linearity in Figures 4 and 5 for n not too small provides strong support for (1.6). (Figure 12 in Section 6 provides similar support for sources less bursty than Poisson.)

Insert Figures 4 and 5 about here.

The four examples depicted in Figure 5 represent all combinations of two different traffic intensities, $\rho = 0.5$ and $\rho = 0.8$, and two MMPP arrival processes. The service-time distributions always are E_{16} with mean 1. As before, these represent nearly deterministic service times. The two arrival processes are two-state MMPPs with asymptotic variance constants in (2.1) of 2.8 and 28. These represent moderately bursty and substantially more bursty sources, respectively (but less bursty than the example in Section 2). These examples have positive arrival rates in both environment states, so that the component MMPPs are *not* renewal processes.

In particular, the MMPPs are characterized by four parameters, one of which can be taken to be the arrival rate. The ratio of the arrival rates in the two environment states is fixed at 4. The expected numbers of arrivals during visits to the two environment states are equal, 5 in the less bursty example with $I_c(\infty) = 2.8$ and 75 in the more bursty example with $I_c(\infty) = 28$.

From Figure 5, we see that the key decay rate parameter γ in (1.6) is decreasing in the traffic intensity ρ but increasing in the burstiness. This appears to be a general tendency. This means that the phenomenon in Section 2 is most likely to occur with low ρ and high burstiness, which is what we anticipate for ATM. The phenomenon might have been missed by others, because they focused on higher ρ and less bursty sources. For instance, the examples in Section 5 of Elwalid and Mitra [23] all have high ρ (above 0.75) and lower burstiness, so that $\alpha > 10^{-2}$.

IV. Three Regimes for the Asymptotic Approximations

The example in Section 2 was one for which the asymptotic approximation $\alpha e^{-\eta x}$ in (1.1) is pretty good, but the simple approximation $e^{-\eta x}$ in (1.3) is bad (because $\alpha \approx 10^{-5}$). In this section we present modifications of this example in which, first, *both* approximations (1.1) and (1.3) are *good* and, second, *both* approximations (1.1) and (1.3) are *bad.*

To obtain these alternative cases, we simply modify the number of sources in the example with $n = 24$ in Section 2, scaling the arrival processes as in Section 3 to keep the total arrival rate fixed at $\lambda = 0.3$ and η fixed at $\eta = 0.01809$. In particular, the two alternative regimes are obtained by letting $n = 2$ and $n = 60$. The results are displayed in Figures 6 and 7.

Since η is independent of *n*, approximation (1.3) is the same for the three cases $n = 2$, $n = 24$ and $n = 60$. However, $\alpha_2 = 0.505$, $\alpha_{24} = 1.45 \times 10^{-5}$ and $\alpha_{60} = 2.22 \times 10^{-12}$. (For $n = 1$, $\alpha_1 = 0.992$.) From Figures 1, 6 and 7, it is evident that we have the three regimes as claimed. In Figure 6 with $n = 2$, approximations (1.1), (1.3) and (1.5) are all very close, while in Figure 7 it is evident that approximations (1.1), (1.3) and (1.5) are very far apart.

Insert Figures 6 and 7 about here.

Since $\alpha_{60} \approx 10^{-12} < 10^{-9}$ when $n = 60$, it should come as little surprise that the asymptotics (1.1) have not kicked in by the time the tail probabilities reach 10^{-9} in this case. As a rough rule of thumb, it appears that approximation (1.1) tends to be good only when α_n is greater than the desired tail probability $P(W > x)$. To a large extent, it appears that we are able to understand when the various approximations will be sufficiently accurate by computing only the asymptotic constant α. (This computation is possible, even for very large *n*, by computing α_n for small values of *n* and then extrapolating using (1.6).)

From Figure 7, we also see that the three-term approximation (1.5) is not accurate at 10^{-9}. The performance of approximation (1.5) in Figure 7 with $n = 60$ is much worse than in Figure 1 with $n = 24$. It is reassuring that its poor performance is signalled by the fact that it is not possible to find parameters exactly matching the third moment in this case; see [17]. Moreover, the three-term approximation performs pretty well even in Figure 7 for tail probabilities above 10^{-4}, for which it was originally designed. The poor performance at 10^{-9} in Figure 7 suggests that alternative fitting procedures should be considered for extremely small probabilities such as 10^{-9}. (This is being investigated.)

V. Changing Other Parameters

In Section 4 we saw what happens as we changed the number of sources, scaling them in the manner of Section 3 so that the total arrival rate remains unchanged. We saw that the quality of the asymptotic approximations decline as *n* increases.

More generally, we conclude that the asymptotic approximations tend to get worse as the number of sources increases, the buffer size decreases, the channel utilization decreases, the target blocking probability increases, and the source gets further from Poisson. However, we would also like to point out that there is significant difference in the accuracy and the region of applicability of the effective bandwidth asymptotic (1.3) and the other asymptotic approximations (1.1) and (1.5). Roughly speaking, (1.3) is not bad when $10^{-2} < \alpha < 10^{2}$, whereas (1.1) and (1.5) are not bad when $10^{-8} < \alpha < 10^{8}$. We expect that in practical engineering situations with multiple sources (1.1) and (1.5) will typically be a reasonably good approximation, while (1.3) may be quite bad.

In this section we consider the effects of buffer size and burstiness. Figures 8 and 9 compare the approximations with exact values when the buffer size is 6000 and 60, respectively, instead of 600 as in Section 2. Here we keep the number of sources fixed at $n = 24$. We scale the sources in the manner of Section 3 until the blocking probability at the indicated capacity is 10^{-9}.

Insert Figures 8 and 9 about here.

So far, we have focused on the steady-state waiting time at arrival epoch. Similar results hold for the steady-state workload at an arbitrary time (the virtual waiting time). To demonstrate this, the remaining numerical results in this paper, including those in Figures 8 and 9, are for the steady-state workload. (The differences between the waiting time and workload are negligible compared to the main phenomena being discussed.)

When the buffer size is increased to 6000, the buffer holds 100 bursts instead of 10. Figure 8 shows that all the approximations perform very well with this larger buffer, just as in Figure 6. Now the utilization is 0.835. The effective-bandwidth approximation works very well; it predicts that the system can support 23 sources, which is within 1 of the exact value.

In contrast, Figure 9 shows that the approximations get even worse when we decrease the buffer size from 600 to 60, which corresponds to just one burst. Figure 9 parallels Figure 7. Now the utilization is only 0.18. The effective-bandwidth approximation would now only admit 10 sources.

Suppose now we keep the buffer size at the high value of 6000 and the number of sources at $n = 24$, but increase the burstiness. Suppose that we increase the burstiness by multiplying the mean number of bursts in an on period by 5. We keep the arrival rate fixed by multiplying the ratio of the off period to the on period by 5. This makes the mean number of arrivals in an on period 300, which means that the buffer now holds $6000/300 = 20$ bursts.

Figure 10 compares the approximations with exact values in this case. Figure 10 shows that there is once again a big difference between the effective-bandwidth approximation and the exact result, just as in Figure 1. For this example, the effective-bandwidth approximation would admit only 11 sources. Hence, the advantage of the larger buffer of size 6000 is offset by the larger burstiness. The actual performance of the approximations obviously depends on the combination of variables that actually prevails.

Insert Figure 10 about here.

VI. Sources Less Bursty than Poisson

In this section we consider sources that are less bursty than Poisson. In particular, we assume that each source is renewal with interarrival times that are E_2. As before, we assume that the overall arrival process is the superposition of independent versions of the single source process. It seems unlikely that the sources in an ATM network will actually be less bursty than Poisson, but this is a possibility, due to traffic shaping at the network edge.

In the earlier bursty model we approximated the deterministic cell-length distribution by an E_{16} distribution and commented that E_{16} and D give about the same result. However, this is no longer true with less bursty arrival processes. Therefore, in order to remain pretty close to D, in this section we assume an E_{1024} service time distribution.

Figure 11 displays the approximations and exact tail probabilities for this case assuming 24 sources and a buffer size of 8. In order to meet the 10^{-9} buffer overflow probability, the channel utilization has to be 29%. In this case the buffer overflow probability is greatly *underestimated* by the effective-bandwidth approximation (1.3). For this case of less bursty sources, the effective-bandwidth approximation (1.3) would admit 39 sources in order to meet the buffer overflow requirement with a buffer of size 8, instead of the proper number of 24.

Insert Figure 11 about here.

Figure 11 also shows the probability of buffer overflow for Poisson arrivals. For large tail probabilities, the exact result is close to the Poisson prediction and, for small tail probabilities, the exact result approaches the true asymptote. It is also interesting to observe that the true asymptote is about ten orders of magnitude higher than the effective bandwidth approximation. This is in sharp contrast with the earlier highly bursty examples where the true asymptote was always smaller than the effective bandwidth approximation. Also, note that there is a qualitative change in the shape of the exact curve. It is concave in Figure 11 as opposed to being convex for the more bursty sources.

The behavior of the tail probabilities in Figure 11 may be understood by plotting, in log scale, the asymptotic constant α_n as a function of n, the number of sources. This we do in Figure 12 for the E_2 sources at channel utilizations 0.3 and 0.7 respectively. This is similar to what we did in Figure 5, and indeed Figures 5 and 12 look similar in the sense that in both cases the logarithm of α_n is asymptotically linear with n, i.e., α_n changes exponentially with n. However, the striking difference is that with more bursty sources (Figure 5) α_n decays exponentially with n and approaches zero, while in the less bursty E_2 case, α_n grows exponentially with n and approaches infinity. The growth rate increases as the channel utilization decreases. We have also plotted how the α_n changes with n for Poisson sources with channel utilizations of 0.3 and 0.7 respectively. Of course, α_n does not change with n in the Poisson case. The Poisson case places in perspective the spectacular growth rate of α_n with n in the non-Poisson case.

Insert Figure 12 about here.

As before, the effective-bandwidth approximation (1.3) performs better with larger buffer sizes. This is shown in Figure 13 where the buffer size is 100 instead of 8. As before, the number of sources is $n = 24$. The sources have been scaled in the manner of Section 3 so that $P(W > 100) = 10^{-9}$. This drives the channel utilization up to 0.949. Figure 13 shows that all the approximations are close to the exact values in this case.

Insert Figure 13 about here.

VII. Approximations for Large Systems with Heterogeneous Sources

In this section we propose a method for obtaining useful approximations for tail probabilities in large systems. Approximations are needed, because our exact algorithm cannot handle large numbers of sources, since the number of phases of the superposed Markovian Arrival Process (MAP) grows rapidly with the number of sources. Specifically, it can be shown that if there are L types of sources, k_i sources of type i, and each source of type i has m_i phases, then the total number, P, of phases of the superposed MAP is given by

$$P = \prod_{i=1}^{L} \begin{bmatrix} k_i + m_i - 1 \\ m_i - 1 \end{bmatrix} . \tag{7.1}$$

Note that for Poisson sources $m_i = 1$ and $\begin{bmatrix} k_i + m_i - 1 \\ m_i - 1 \end{bmatrix} = 1$, so that we can add any number of them without increasing P. To run the exact model in reasonable time, we need P to be at most about 100. In all the numerical examples in this paper we assumed $L = 1$ and $m_i = 2$ so that we could treat up to about $K_1 = 99$ sources. (We actually considered up to 60 sources). However, with $L = 2$ and 3 (with m_i still fixed at 2) we can treat only up to $K_i = 9$ and $K_i = 3$ sources of each type.

Now we specify our approximation procedure in terms of two examples. First, suppose that there are three classes, each with 100 homogeneous sources. We can calculate the asymptotic decay rate η exactly by considering the three-class system with one source in each class, with each source scaled appropriately, as in Section 3, so that the arrival rate of each single source equals the total arrival rate for all 100 sources, for each class, in the original system.

In order to approximately determine the asymptotic constant α in the original system, we calculate the exact asymptotic constants in the systems with three classes and k sources in each class, again scaled to be consistent with the original system according to Section 3, for several feasible k, e.g., $k = 1$, 2 and 3. (Note that here $L = 3$ and $m_i = 2$, so that $k_i = 3$ for all i is feasible.) Then, assuming (1.6), we obtain an approximate asymptotic constant for the original system by fitting β and γ in (1.6) to the data. If the resulting estimated asymptotic constant α is substantially greater than the target tail probability, then we can apply approximation (1.1) with confidence. If the estimated asymptotic constant α is less than the target tail probability, then we note that (1.1) is probably not appropriate. If the estimated asymptotic constant is greater than the target tail probability, but not much greater, then we might use the heuristic suggested in Section 1; i.e., we might use (1.1) with a higher value of α as a rough conservative estimate.

The approach we have just described is satisfactory if all classes have many sources, as when there are 100 sources from each of the three classes. However, in actual applications, e.g., with video sources, there may be only a few sources from some classes. To illustrate, suppose that we have four classes, with 1 source in the first class, 2 sources in the second class and 100 sources each in the last two classes. Let all sources have two states. The method we have just applied does not work for this example, but a modification does.

We have found that the asymptotic relation (1.6) still holds if we divide the sources into two groups, and hold one group of sources fixed, while we multiply the number of sources in the second group by n. As before, we scale the sources in the second group, so that the total arrival rate for each class remains fixed, independent of n.

As before, (1.1) holds for each n and the asymptotic decay rate η is independent of n. Moreover, numerical experience indicates that the asymptotic relation (1.6) still holds, but now β in (1.6) is a function of the fixed sources. Now to estimate α in the original system we at first estimate β and γ in (1.6) using $n = 2$ and 3 and then estimate α from those using $n = 100$. Note that here $L = 4$, $m_i = 2$ for each i, $k_1 = 1$, $k_2 = 2$, $k_3 = k_4 = n$ and hence $n = 2$ and 3 are feasible for the exact model.

More generally, the asymptotic result just described suggests an approximation for m classes with multiplicities $n_1, n_2, \ldots, n_m$ of the form

$$\alpha_{n_1, \ldots, n_m} \approx \beta e^{-(\gamma_1 n_1 + \ldots + \gamma_m n_m)} . \tag{7.2}$$

For multiple classes, approximation (7.2) is convenient because we can determine the asymptotic decay rates γ_i by changing one class at a time. However, more work is needed on this.

VIII. The Algorithm

In this section we describe our algorithm for numerically computing the exact tail probabilities. To compute the tail probabilities, we draw heavily on Lucantoni [35]; see [36] for a review. We model each source as a *Markovian Arrival Process* (MAP), which is a two-dimensional Markov process $\{N(t), J(t)\}$ on the state space $\{(i,j) : i \geq 0, 1 \leq j \leq m\}$ with an infinitesimal generator having the structure

$$Q = \begin{bmatrix} D_0 & D_1 & 0 & 0 & \ldots \\ 0 & D_0 & D_1 & 0 & \ldots \\ 0 & 0 & D_0 & D_1 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \end{bmatrix} \qquad (8.1)$$

where 0, D_0 and D_1 in (8.1) are $m \times m$ matrices; D_0 has negative diagonal elements and nonnegative off-diagonal elements; and $D \equiv D_0 + D_1$ is an irreducible infinitesimal generator. We also assume that $D \neq D_0$, which ensures that arrivals will occur. The variable $N(t)$ counts the number of arrivals in the interval $[0,t]$, while $J(t)$ gives the arrival ''phase'' at time t.

In this paper we consider the superposition of many sources. If the individual sources are characterized by matrices D_{0i} and D_{1i} for $i = 1,2, \ldots, K$, where K is the number of sources, then it can be shown [35] that the superposed system is also a MAP characterized by matrices

$$D_0 = D_{01} \oplus D_{02} \oplus \ldots \oplus D_{0K} \qquad (8.2)$$

$$D_1 = D_{11} \oplus D_{12} \oplus \ldots \oplus D_{1K} \qquad (8.3)$$

where $\oplus$ is the Kronecker sum. We do all computations on the superposed MAP characterized by D_0 and D_1.

As explained in [35], the *Laplace-Stieltjes transform* (LST) of the steady-state waiting-time cumulative distribution function (CDF) is given by

$$\hat{W}(s) = (\pi d)^{-1} s(1-\rho) g [sI + D(\hat{h}(s))]^{-1} e \;, \qquad (8.4)$$

where e is a vector of all 1's, $\hat{h}(s)$ is service time LST, ρ is the traffic intensity (the mean arrival rate times the mean service time), $d = D_1 e$, π is the stationary probability vector of the Markov process with generator D (i.e., π satisfies $\pi D = 0$ and $\pi e = 1$), g satisfies the equations $gG = g$ and $ge = 1$, and G satisfies the equation

$$G = \int_0^\infty e^{(D_0 + D_1 G)x} dH(x) \;, \qquad (8.5)$$

where $H(x)$ is the service-time CDF and

$$D(\hat{h}(s)) = D_0 + D_1 \hat{h}(s) \;. \qquad (8.6)$$

Our computation is based on, first, fast and accurate computation of $\hat{W}(s)$ using (8.4) and, second, fast and accurate transform inversion to get the waiting time CDF from $\hat{W}(s)$. To do the inversion, we use the Fourier-series method in [5] along with the round-off error control procedure in [15]. The procedure in [15] also provides a self-contained accuracy check by doing the calculation twice with different round-off control parameters. This amounts to using different contours for complex inversion integration. With this procedure, we achieve high accuracy even at the tail probability of 10^{-9}.

Next we describe two techniques for greatly speeding up the computation of the LST $\hat{W}(s)$. First, note that all individual sources are two-phase sources. The Kronecker sum operations in (8.2) and (8.3) makes the superposed source a 2^K-phase source. This would prevent us from using large K, but we can take advantage of the fact that the sources are homogeneous to greatly reduce the dimensionality of the superposed source. We can define the phase of the superposed source as one plus the number of component sources in phase 1. Therefore, the total number of phases is $K + 1$ instead of 2^K. The D_0 and D_1 matrices of the new superposed sources are given in terms of the component D_{01} and D_{11} matrices of the (identical) component sources as follows:

$$(D_1)_{i,i} = (i-1)(D_{11})_{1,1} + (K-i+1)(D_{11})_{2,2} \quad \text{for} \quad i = 1,2, \ldots, K+1$$

$$(D_1)_{i,i+1} = (K-i+1)(D_{11})_{2,1} \quad \text{for} \quad i = 1,2, \ldots, K \qquad (8.7)$$

$$(D_1)_{i,i-1} = (i-1)(D_{11})_{1,2} \quad \text{for} \quad i = 2,3, \ldots, K+1$$

with $(D_1)_{i,j} = 0$ for all other pairs (i,j),

$$(D_0)_{i,i+1} = (K-i+1)(D_{01})_{2,1} \quad \text{for} \quad i = 1,2, \ldots, K$$

$$(D_0)_{i,i-1} = (i-1)(D_{01})_{1,2} \quad \text{for} \quad i = 2,3, \ldots, K+1 \qquad (8.8)$$

with $(D_0)_{i,j} = 0$ for all other pairs (i,j) with $i \neq j$, and

$$(D_0)_{i,i} = -[(D_0)_{i,i-1} + (D_0)_{i,i+1} + (D_1)_{i,i} + (D_1)_{i,i-1} + (D_1)_{i,i+1}] \;. \qquad (8.9)$$

For the service-time distribution, we approximate the deterministic distribution by an Erlangian distribution of order k (E_k). A significant computational burden is the computation of G from the matrix integral equation (8.5). The uniformization procedure recommended in [35] for general service-time distributions can be significantly improved for Erlang distributions by noting that for the E_k distribution with mean 1 (8.5) reduces to

$$G = [I - k^{-1}(D_0 + D_1 G)]^{-k} \;. \qquad (8.10)$$

Equation (8.10) can easily be solved by successive substitution, i.e.,

$$G_{n+1} = [I - k^{-1}(D_0 + D_1 G_n)]^{-k} \;, \qquad (8.11)$$

where G_0 is chosen to be a stochastic matrix. Note that since we choose $k = 2^m$ (specifically 2^4 for bursty sources and 2^{10} for smooth sources), each iteration in (8.11) involves only a single matrix inversion and m matrix multiplications.

In order to compute the asymptotic parameters α and η in (1.1), we use two independent algorithms (that cross-check each other): the algorithm in [2] and the moment-based procedure in [14] and [1].

IX. Conclusions

Our first main conclusion here is that the effective-bandwidth approximation (1.3) can break down when there is a large number of independent sources. Approximation (1.3) tends to get worse as the number of sources increases, the channel utilization decreases, the buffer size decreases and the source gets further from Poisson, either more bursty or less bursty.

If the sources are more bursty than Poisson, as is anticipated for ATM networks, then the effective-bandwidth approximation (1.3) is conservative. When the approximation is bad, there may be substantially more statistical multiplexing gain than approximation (1.3) predicts. On the other hand, if the sources are less bursty than Poisson, then the effective-bandwidth approximation is no longer conservative, and may also be bad. In general, contrary to many statements, the effective-bandwidth approximation need not be conservative.

Typically, the exact tail probabilities lie between the effective-bandwidth approximation (1.3) and the true asymptote (1.1). If these two curves are very close to each other (differ by less than a factor of 10), then (1.3) is usually reasonably accurate. If the two curves are far from each other, but not too far (differ by less than 10^8), then (1.3) is bad, but (1.1) and (1.5) are reasonably accurate. Finally, if the two curves are extremely far apart (differ by more than 10^8), then even (1.1) and (1.5) are not accurate. Our second main conclusion is that, even though (1.3) may not be a good approximation, the true asymptote (1.1) often is a good approximation. Moreover, from calculations of the asymptotic parameters α and η, we have a way to estimate whether or not the approximations will be good.

A reason for the degradation of the effective-bandwidth approximation as the number n of sources increases is that the asymptotic constant α_n in (1.1) is itself asymptotically exponential in n. For sources more bursty than Poisson, α_n decrease to 0 exponentially fast, as shown in Figures 4 and 5. For sources less bursty than Poisson, α_n increases to infinity exponentially in n, as shown in Figure 12. The Poisson case is the reference case, because α_n there does not change with n. The asymptotically exponential form for α_n in (1.6) allows us to compute it approximately for arbitrary n by extrapolating based on computed values for small n. For heterogeneous sources with different multiplicities, we exploit (7.1). Having a way to approximate the asymptotic constant α is important, not only because we can use it in approximations (1.1) and (1.5), but also because the value of α indicates whether or not the approximations will be good.

Finally, in addition to gaining a better understanding of the effective-bandwidth approximation (1.3), we have provided bases for more refined analysis tools via our exact MAP/G/1 numerical algorithm, the refined approximation (1.5) [17] and the exponential relation for α_n in (1.6). We have indicated how (1.6) and (7.1) can be combined with the true asymptote (1.1) to get an approximation that is almost as simple as the effective-bandwidth approximation (1.3), but is applicable in a substantially wider region.

Acknowledgments. We thank our colleagues Bharat Doshi, Ted Eckberg, Dave Houck and Pat Wirth for helpful comments.

References

[1] J. Abate, G. L. Choudhury, D. M. Lucantoni and W. Whitt. Asymptotic analysis of tail probabilities based on the computation of moments, AT&T Bell Laboratories, 1994, submitted.

[2] J. Abate, G. L. Choudhury and W. Whitt. Asymptotics for steady-state tail probabilities in structured Markov queueing models. *Stochastic Models*, 10 (1994) 99-143.

[3] J. Abate, G. L. Choudhury and W. Whitt. Exponential approximations for tail probabilities in queues, I: waiting times. *Opns. Res.*, to appear.

[4] J. Abate, G. L. Choudhury and W. Whitt. Exponential approximations for tail probabilities in queues, II: sojourn time and workload. *Opns. Res.*, to appear.

[5] J. Abate and W. Whitt. The Fourier series method for inverting transforms of probability distributions. *Queueing Systems* 10 (1992) 5-88.

[6] D. Anick, D. Mitra and M. M. Sondhi. Stochastic theory of a data handling system with multiple sources. *Bell System Tech. J.*, 61 (1982) 1871-1894.

[7] A. Baiocchi. Asymptotic behavior of the loss probability of the MAP/GI/1/K queue, Part I: theory. INFOCOM Dept., University of Rome ‘‘La Sapienza,’’ 1992.

[8] A. Baiocchi, N. Blefari, M. Listanti, A. Roveri and R. Winkler. Loss performance analysis of an ATM multiplexer loaded with high-speed ON-OFF sources. *IEEE J. Sel. Areas Commun.* 9 (1991) 388-393.

[9] D. D. Botvich and N. G. Duffield, Large deviations, the shape of the loss curve, and economies of scale in large multiplexers, Dublin City University, Dublin, Ireland.

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

[11] C.-S. Chang, P. Heidelberger, S. Juneja and P. Shahabuddin. Effective bandwidth and fast simulation of ATM intree networks, IBM RC 18586, IBM T. J. Watson Research Center, Yorktown Heights, NY, 1992.

[12] G. L. Choudhury, K. K. Leung and W. Whitt. An algorithm for product-form loss networks based on numerical inversion of generating functions. *Proc. IEEE Globecom ‘94.* (1994) 1123-1128.

[13] G. L. Choudhury, K. K. Leung and W. Whitt. An inversion algorithm for loss networks with state-dependent rates. *Proc. IEEE Infocom ‘95,* to appear.

[14] G. L. Choudhury and D. M. Lucantoni. Numerical computation of moments with application to asymptotic analysis. *Opns. Res.*, to appear.

[15] G. L. Choudhury, D. M. Lucantoni and W. Whitt. Multidimensional transform inversion with applications to the transient M/G/1 queue. *Ann. Appl. Prob.* 4 (1994) 719-740.

[16] G. L. Choudhury, D. M. Lucantoni and W. Whitt. On the effectiveness of effective bandwidths for admission control in ATM networks. *The Fundamental Role of Teletraffic in the Evolution of Telecommunication Networks, Proceedings of ITC14,* J. Labetoulle and J. Roberts, eds., 1a (1994) 411-420.

[17] G. L. Choudhury, D. M. Lucantoni and W. Whitt. Refined approximations for G/G/1 queues, in preparation.

[18] G. L. Choudhury and W. Whitt. Heavy-traffic approximations for the asymptotic decay rate in BMAP/G/1 queues. *Stochastic Models* 10 (1994) 453-498.

[19] E. Çinlar. Superposition of point processes, pp. 549-606 in *Stochastic Point Processes: Statistical Analysis, Theory and Applications,* ed. P. A. W. Lewis, Wiley, New York, 1972.

[20] C. Courcoubetis, G. Kesidis, A. Ridder, J. Walrand and R. R. Weber. Admission control and routing in ATM networks using inferences from measured buffer occupancy. *IEEE Trans. Commun.*, to appear.

[21] C. Courcoubetis and R. R. Weber. Buffer overflow asymptotics for a switch handling many traffic sources, University of Cambridge, 1994.

[22] A. Dembo and O. Zeitouni. presentation at the national meeting of the Operations Research Society of America, Boston, April 1994.

[23] A. I. Elwalid and D. Mitra. Effective bandwidth of general Markovian traffic sources and admission control of high speed networks, *IEEE/ACM Trans. Networking* 1 (1993) 329-343.

[24] A. I. Elwalid and D. Mitra. Markovian arrival and service communication systems: spectral expansions, separability and Kronecker-product forms, 1993.

[25] K. W. Fendick, V. R. Saksena and W. Whitt. Investigating dependence in packet queues with the index of dispersion for work. *IEEE Trans. Commun.*, 39 (1991) 1231-1244.

[26] R. J. Gibbens and P. J. Hunt. Effective bandwidths for the multi-type UAS channel. *Queueing Systems* 9 (1991) 17-28.

[27] P. W. Glynn and W. Whitt. Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *Studies in Applied Probability, Papers in Honour of Lajos Takács*, J. Galambos and J. Gani (eds.), Applied Probability Trust, Sheffield, England, 1994, 131-156.

[28] R. Guerin, H. Ahmadi and M. Naghshineh. Equivalent capacity and its application to bandwidth application in high-speed networks. *IEEE J. Sel. Areas Commun.* 9 (1991) 968-981.

[29] R. Guerin and L. Gun. A unified approach to bandwidth allocation and access control in fast packet-switched networks. INFOCOM '92, Florence, Italy.

[30] H. Heffes and D. M. Lucantoni. A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance. *IEEE J. Sel. Areas Commun.* SAC4 (1986) 856-868.

[31] J. Y. Hui. Resource allocation for broadband networks. *IEEE J. Select. Areas Commun.* SAC-6 (1988) 1598-1608.

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

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

[34] A. Kuczura. The interrupted Poisson process as an overflow process. *Bell System Tech. J.*, 52 (1973) 437-448.

[35] D. M. Lucantoni. New results on the single server queue with a batch Markovian arrival process. *Stochastic Models* 7 (1991) 1-46.

[36] D. M. Lucantoni. The BMAP/G/1 queue: a tutorial. *Models and Techniques for Performance Evaluation of Computer and Communication Systems,* L. Donatiello and R. Nelson, eds., Springer-Verlag, 1993, 330-358.

[37] D. M. Lucantoni, G. L. Choudhury and W. Whitt. The transient BMAP/G/1 queue. *Stochastic Models* 10 (1994) 145-182.

[38] M. F. Neuts. The caudal characteristic curve of queues. *Adv. Appl. Prob.* 18 (1986) 221-254.

[39] J. W. Roberts. *Performance Evaluation and Design of Multiservice Networks, COST 224 Final Report,* Commission of the European Communities, Luxembourg, 1992.

[40] A. Simonian and J. Guilbert. Large deviations approximation for fluid queues fed by a large number of on-off sources, preprint, 1994.

[41] K. Sohraby. On the asymptotic behavior of heterogeneous statistical multiplexer with applications. *IEEE INFOCOM '92,* Florence, Italy.

[42] K. Sohraby. On the theory of general on-off sources with applications in high speed networks. *IEEE INFOCOM '93,* San Francisco, CA.

[43] K. Sriram and W. Whitt. Characterizing superposition arrival processes in packet multiplexers for voice and data. *IEEE J. Sel. Areas. Commun.* SAC-4 (1986) 833-846.

[44] D. N. C. Tse, R. G. Gallager and J. N. Tsitsiklis. Statistical multiplexing of multiple time-scale Markov streams. Laboratory for Information and Decision Sciences, M.I.T. 1994.

[45] W. Whitt. Tail probabilities with statistical multiplexing and effective bandwidths for multi-class queues. *Telecommunication Systems*, 2 (1993) 71-107.

Figure 1. A comparison of approximations and exact values of the probability of buffer overflow (approximated by the tail probability $P(W > x)$) as a function of the buffer size x for the example in Section 2.

Figure 2. The decay rate η_n in (1.1) as a function of the number n of sources (without scaling the sources) for the example in Section 2.

Figure 3. The asymptotic constant α_n in (1.1) as a function of the number n of sources (without scaling the sources) for the example in Section 2.

Figure 4. The asymptotics constant α_n as a function of the number n of sources (with scaling) for the example in Section 2.

Figure 5. The asymptotic constant α_n as a function of the number n of sources (with scaling) for four examples in Section 3.

Figure 6. A comparison of approximations and exact values of the probability of buffer overflow (approximated by the tail probability $P(W > x)$ as a function of the buffer size x for the example with $n = 2$ in Section 4.

Figure 7. A comparison of approximations and exact values of the probability of buffer overflow (approximated by the tail probability $P(W > x)$ as a function of the buffer size x for the example with $n = 60$ in Section 4.

Figure 8. A comparison of approximations and exact values of the probability of buffer overflow (approximated by the workload-tail probability) as a function of the buffer size x for the example with $n = 24$ in Section 2 and higher buffer capacity 6000 in Section 5.

Figure 9. A comparison of approximations and exact values of the probability of buffer overflow (approximated by the workload-tail probability) $n = 24$ in Section 2 and lower buffer capacity 60 in Section 5.

Figure 10. A comparison of approximations and exact values of the probability of buffer overflow (approximated by the workload-tail probability) $n = 24$ in Section 2, higher buffer capacity 6000 and higher burstiness in Section 5.

Figure 11. A comparison of approximations with exact values of the probability of buffer overflow (approximated by the workload tail probability) as a function of the buffer size x for the example in Section 6 with E_2-renewal-process sources, $n = 24$ and buffer size 8.

Figure 12. The asymptotic constant α_n as a function of the number n of sources (with scaling for the E_2 examples in Section 6.

Method of Computation	Number of Sources Allowed for Buffer Size $x = 600$	Buffer Size Required to Support $n = 24$ Sources
exact	24	600
effective-bandwidth approximation (1.3)	12	1146
(1.1) and (1.5)	25	530
Poisson	78	20
Average-Rate engineering	80	not applicable
peak-rate engineering	7	not applicable

Table 1. A comparison of different methods for determining (i) the number of sources for the fixed buffer size $x = 600$ and (ii) the buffer size required to support $n = 24$ sources for the example in Section 2.

Figure 13. A comparison of approximations with exact values of the probability of buffer overflow (approximated by the workload tail probability) as a function of the buffer size x for the example in Section 6 with E_2-renewal-process sources, $n = 24$ and buffer size 100.