A Fast Performance Model for Real-Time Multimedia Communications

Shiouming Stanley Wang and John A. Silvester

CENG Technical Report 92-21

Department of Electrical Engineering - Systems
University of Southern California
Los Angeles, California 90089-2562
(213)740-4579

A Fast Performance Model for Real-Time Multimedia Communications

Shiouming Stanley Wang and John A. Silvester

Department of Electrical Engineering-Systems
University of Southern California
Los Angeles, California 90089

Abstract

The Markov Modulated Poisson Process (MMPP), a doubly stochastic Poisson process has received a lot of attention due to its ability to model a highly correlated arrival stream while retaining analytical tractability. In this paper, cell arrival process from integrated voice and data sources is approximated by a *two-state MMPP with batch arrivals* ($MMPP^{[X]}$) making use of a new matching technique which leads to a very accurate delay performance prediction. We then use a similar approach for the approximation of integrated video, voice and data sources. The approximations are verified by simulation and comparisons to another model recently proposed by Baiocchi *et al.* [1].

1 Introduction

The communication industry is currently in the process of migrating its infrastructure to the new era of *Asynchronous Transfer Mode* (ATM) based *Broadband Integrated Services Digital Networks* (B-ISDN). The goals of B-ISDN, i.e., a wide range of new services and effective utilization of network resources, present a great challenge in the design of traffic and congestion control. The ability of the communication providers to meet this challenge will have a great impact on the viability of B-ISDN.

While the effectiveness of ATM derives largely from the statistical multiplexing gains; this presents a potential threat that network resources may be over-utilized especially when the traffic sources are bursty. Thus, it is necessary to study the queueing behavior of the traffic sources in order to design a good traffic and congestion control algorithm that can take the advantage of the statistical multiplexing.

In modeling B-ISDN traffic sources, we need different models for the different types of traffic sources, e.g., video, voice and data, since they present totally different statistical characteristics. Doing so, however, increases the complexity of the traffic model. For real-time traffic control, where processing time is critical, a simple approximate traffic model allowing fast yet accurate prediction is needed.

In this paper, superposition of voice sources and superposition of video sources are modeled as *birth-death processes* separately (these models are discussed in the next section). Cell arrival process from aggregated data sources is modeled as a *Poisson process with batch arrivals*. We propose an approximation for the integrated video, voice and data arrival process by a two-state *Markov Modulated Poisson Process*es (MMPP) with batch arrivals. An MMPP, a special case of *Neuts' versatile Markov process* [17], is a *doubly stochastic Poisson process* where the arrival rate is determined by the state of the underlying finite-state Markov chain. MMPP's are suitable since they allow representation of highly correlated arrivals yet retain analytical tractability of their queueing behavior [12], [17], [18], [20].

The paper is organized as follows: more discussion of the traffic models for voice and video sources is found in section 2. The approximation technique is discussed in section 3. Simulation and numerical results are presented in section 4. Finally, section 5 provides some conclusions.

2 Models for Voice and Video Traffic

Numerous studies that characterize packetized voice sources, e.g., [1], [4], [5], [8], [9], [11], [14], [21], [22] and [24], have been made. Among these studies, Brady's *ON-OFF process* [5] has been widely adopted as the model for the arrival process corresponding to a single voice source. In the ON-OFF process model, the voice source alternates between exponentially distributed (or geometrically distributed if we use a discrete time scale) *ON* periods and exponentially distributed *OFF* periods. The *ON* periods correspond to talkspurts while the *OFF* periods represent silence periods. Packets are created with a constant interarrival time during the *ON* periods and no packets are generated during the *OFF* periods. The transition rates (or *transition probabilities* for the discrete time case), from ON-to-OFF (α) and from OFF-to-ON (β), are determined by the expected durations of the talkspurts and the silence periods (see Fig. 1).

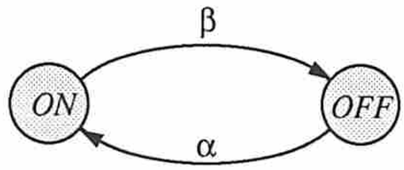

Fig. 1. An ON-OFF process.

We can represent the superposition of N such ON-OFF processes by a *finite-state birth-death process* with the states representing the number of voice sources in talkspurt (in the ON-state) as shown in Fig. 2.

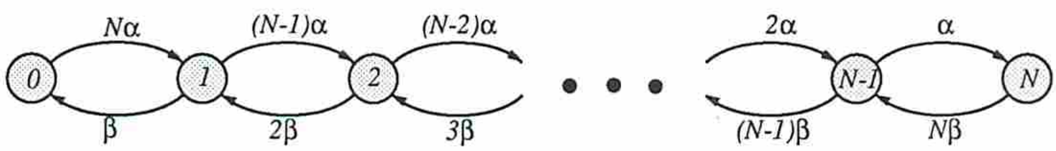

Fig. 2. Model for superposition of N voice sources.

We assume that the video sources are of the video-telephony type (showing a person talking in front of the camera without sudden movements) and use a *continuous-time Markov chain* to model the superposition of several video sources, as suggested by Maglaris *et al.* [13]. In this model, the output bit rates from codecs are assumed to fall into a set of discrete bit rate levels evenly spaced between zero and the peak rate. Transitions are assumed to occur between adjacent levels only, according to a *birth-death Markov chain*. The transition rates are determined by matching the model's parameters with statistical characteristics of the traffic, as described in [13]. Fig. 3 shows the model with $M + 1$ bit rate levels.

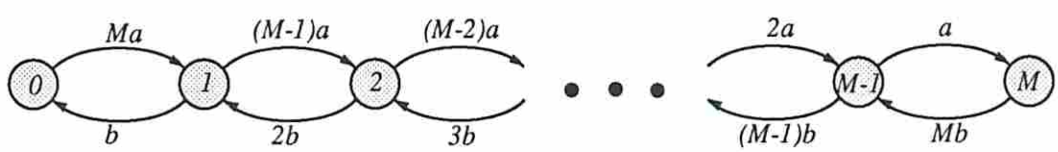

Fig. 3 Model for video sources with a total of M+1 bit rate levels.

The difficulty in generating useful analytical results from a queueing system with an arrival process based on the birth-death processes defined above has forced researchers to look for approximations. Two main approaches have been proposed. The first approach uses a fluid flow approximation [6], [15], [22], which overlooks the randomness (short-term variations) of the arrivals and has a limited applicable range. The second approach carefully matches the parameters of the process to a simpler one, e.g., the two-state MMPPs used in [1], [8] and [15], and the *renewal process* used in [15]. The later approach suffers from low accuracy and high computational overhead thus has little value in modeling real-time traffic.

3 The Approximation Technique

The cell arrival process corresponding to superposed voice sources is approximated by means of a two-state MMPP as in [1] but a different procedure for parameter matching is proposed to resolve the problems that were noted in the previous section. We use a similar approach to include video sources into our model by observing that integration of the models for voice and video sources results in a *two-dimensional Markov chain.* Data packets are assumed to form a Poisson process, i.e., data cells form a Poisson process with batch arrivals. The arrivals from the integrated video, voice and data sources are then modeled as a *two-state MMPP with batch arrivals* (denoted $\mathrm{MMPP}^{[X]}$ in this paper).

3.1 Using two-state MMPP as an approximation for voice sources

Heffes and Lucantoni, in their work [8], use the following parameters of the two-state MMPP to match with the underlying Markov chain model of the superimposed voice sources:

i. the expected arrival rate;

ii. the variance-to-mean ratio of the number of arrivals in some time interval;

iii. the long-term variance-to-mean ratio of the number of arrivals;

iv. the third moment of the number of arrivals in some time interval.

Upon observing the strong role of the overload† period in determining the performance of the multiplexer, Nagarajan *et al.*, in their matching procedure [15], replace the step (*iv*) in [8] by "*the variance of the number of arrivals in some time interval giving that the system is in overload state,*" and achieve an improvement in predicting packet loss for a finite-buffered system. In [1], Baiocchi *et al.* attack the problem from a different angle. They use a theorem proved in [16] (theorem 2.3.1, p. 62) and present an "asymptotic matching procedure" which leads to more accurate results than those of [8] and [22] (see the reference for details). As was pointed out in the previous section, these approaches have their drawbacks. We thus propose a new matching technique to overcome these problems.

3.2 The matching technique

Consider a multiplexer loaded with N independent voice sources each of which is modeled by

† Defined to be when the number of active sources exceeds the capacity of the system.

an ON-OFF process with the following parameters:

i. ω, the constant arrival rate in the ON-state;

ii. α, the transition rate from the OFF-state to the ON-state;

iii. β, the transition rate from the ON-state to the OFF-state.

Let C denote the capacity of the multiplexer; $L = \lfloor C/\omega \rfloor$ be the maximum number of active voice sources which the multiplexer can support; T_i be the expected time until the superposition process of the N ON-OFF processes visits state i-1 for the first time starting from state i; and π_k defined by:

$$\pi_k = \binom{N}{k} \left(\frac{\alpha}{\alpha+\beta}\right)^k \left(\frac{\beta}{\alpha+\beta}\right)^{N-k}, \quad 0 \le k \le N \tag{1}$$

be the steady state probability that k out of the N voice sources are in the ON-state.

To find T_i, let us first observe that it satisfies the following recurrence relation:

$$T_i = \frac{1}{(N-i)\alpha + i\beta} + \frac{(N-i)\alpha}{(N-i)\alpha + i\beta} (T_{i+1} + T_i), \quad 1 \le i \le N-1 \tag{2}$$

with the boundary condition, $T_N = 1/(N\beta)$. Note that, in (2), the first term on the right-hand side gives the expected sojourn time in state i. The first part of the second term on the right-hand side specifies the probability that the process will make a transition to state $i+1$ given that it is currently in state i. It can be easily proved, by successive substitution for example, that T_{N-i} satisfies the following equation:

$$T_{N-i} = \sum_{k=0}^{i} \frac{\frac{i!}{k!}\alpha^{i-k}}{\frac{(N-k)!}{(N-i-1)!}\beta^{i+1-k}}$$

Thus, T_{N-i} has the following explicit closed-form expression:

$$T_{N-i} = \frac{1}{N\beta\binom{N-1}{i}} \sum_{k=0}^{i} \binom{N}{k} \left(\frac{\alpha}{\beta}\right)^{i-k}, \quad 0 \le i \le N-1 \tag{3}$$

Using (3), we can find the expected time until the system load drops below the system capacity, T_{L+1}, once the load exceeds the system limit, L.

Our matching technique works as follows: referring to Fig. 4, the states of the superposition arrival process are divided into two disjoint subsets, the overload states, $\{L+1, L+2, \ldots, N\}$, and the underload states, $\{0, 1, 2, \ldots, L\}$. These two sets of states are mapped into the state-II (the overload-state) and the state-I (the underload-state) of a two-state MMPP respectively.

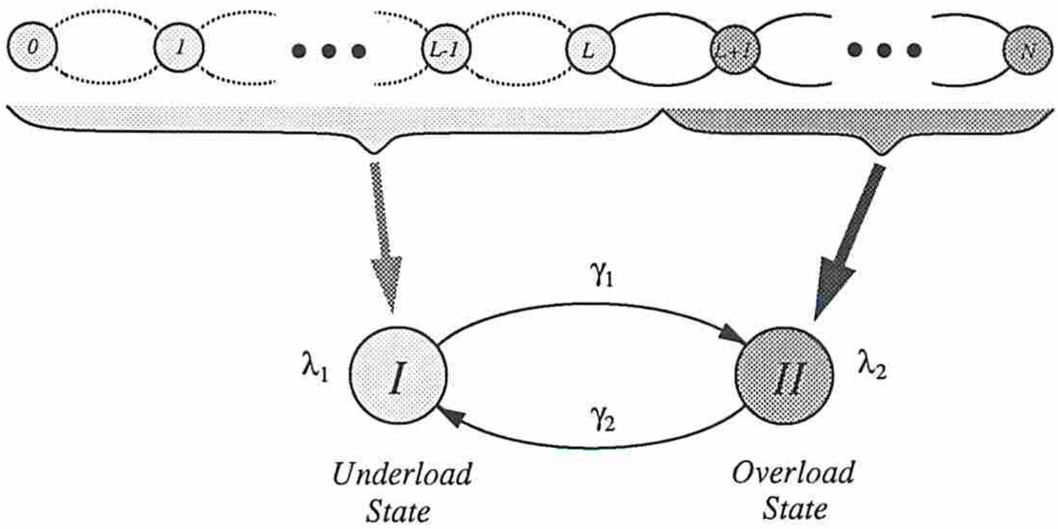

Fig. 4. State mapping between the arrival process and the two-state MMPP for integrated voice and data traffic.

Let γ_1, γ_2 be the transition rates and λ_1, λ_2 be the cell arrival rates for state-I and state-II correspondingly. We then propose the following approach to match the parameters (note that, *ii-iv* are effectively the same as those used in [1]):

i. $$\gamma_2 = \frac{1}{T_{L+1}} = \left(\frac{1}{N\beta \binom{N-1}{N-L-1}} \sum_{k=0}^{N-L-1} \binom{N}{k} \left(\frac{\alpha}{\beta}\right)^{N-L-1-k} \right)^{-1}$$

ii. $$\lambda_1 = \omega \sum_{k=0}^{L} k \left(\frac{\pi_k}{\Pi_u} \right), \text{ where } \Pi_u = \sum_{k=0}^{L} \pi_k$$

iii. $$\lambda_2 = \omega \sum_{k=L+1}^{N} k \left(\frac{\pi_k}{\Pi_o} \right), \text{ where } \Pi_o = \sum_{k=L+1}^{N} \pi_k$$

iv. $\gamma_1 = \gamma_2 \frac{\omega\phi - \lambda_1}{\lambda_2 - \omega\phi}$, where $\phi = N(\frac{\alpha}{\alpha+\beta})$ is the expected number of active calls.

As one might expect, the arrival stream from the superposed process is burstier than that from the corresponding two-state MMPP especially when the two-state MMPP is in its overload-state (when there are a relatively larger number of voice sources in talkspurt). This can be understood by observing the "smoothing" effect introduced by the reduced number of states in the approximate model, because when the two-state MMPP is in either one of its states the arrival process is just an ordinary Poisson process rather than each active source emitting cells at the peak rate while it is "*ON*" and none while it is "*OFF*." Since, on the average, a longer queue is expected by a new arrival for a burstier arrival process due to a shorter interarrival time within the same burst, the approximation is expected to perform worse and worse as the system load increases, i.e, the number of active sources and the burstiness of the arrival process increase. The same problem is also experienced in the asymptotic matching procedure proposed in [1].

3.3 Improve the accuracy of the approximation

In order to improve the accuracy of our model, let us observe that if we overestimate T_{L+1}, we may, in fact, improve the model's accuracy. This is because the arrival rate in the overload-state is higher than that in the underload-state (see step *ii* and *iii* of the matching procedure). Also, increasing the average overload time causes a decrease on the average underload time (see step *iv* of the matching procedure). Therefore, overestimating the average overload time, T_{L+1}, will increase the burstiness of the two-state MMPP. Note also that (for the same number of voice calls) the number of states which are mapped into the overload-state of the two-state MMPP increases, i.e., L decreases, as the channel capacity decreases. This corresponds to the two-state MMPP staying a longer time in the overload-state on the average. In addition, the longer the average overload time is, the more severe the underestimation of the burstiness is. Thus, a higher overestimation for T_{L+1} is needed to offset the increasingly severe underestimation of the burstiness.

One approach that overestimates T_{L+1} is to replace the expression in (3) by an upper-bound. After several trials, we found that the following upper-bound for (3) leads to very good results:

$$\frac{i+1}{N\beta\binom{N-1}{i}} \max_{0 \le k \le i} \binom{N}{k} \left(\frac{\alpha}{\beta}\right)^{i-k} \tag{4}$$

By applying *Stirling's formula* [7] to (4), we have:

$$\frac{i+1}{N\beta\binom{N-1}{i}} \max_{0 \le k \le i} O(\sqrt{N})\left(\frac{N^N}{k^k (N-k)^{N-k}}\right)\left(\frac{\alpha}{\beta}\right)^{i-k}$$

$$= \frac{i+1}{N\beta\binom{N-1}{i}} \max_{0 \le k \le i} O(\sqrt{N}) e^{N\log N - (N-k)\log(N-k) - k\log k + (i-k)\log(\frac{\alpha}{\beta})}$$

$$= \frac{i+1}{N\beta\binom{N-1}{i}} \max_{0 \le k \le i} O(\sqrt{N}) e^{N(-\frac{N-k}{N}\log\frac{N-k}{N} - \frac{k}{N}\log\frac{k}{N}) + (i-k)\log(\frac{\alpha}{\beta})} \tag{5}$$

The problem now reduces to determining which value of k provides the largest exponent in (5). Let us replace k/N by x and rewrite the exponent of (5) as:

$$N\left[(x-1)\log(1-x) - x\log x + \left(\frac{i}{N} - x\right)\log\left(\frac{\alpha}{\beta}\right)\right] \tag{6}$$

After taking the derivative of (6) and equating to zero, we get:

$$\log(1-x) - \log x - \log\left(\frac{\alpha}{\beta}\right) = 0 \tag{7}$$

From (7), it can be easily verified that (6) is maximized at $k = \frac{N\beta}{\alpha+\beta}$. Thus, we conclude that if

$$i < \frac{N\beta}{\alpha+\beta} \tag{8}$$

$k = i$ should be chosen to find the upper-bound for (3); otherwise, $k = \left\lfloor \frac{N\beta}{\alpha+\beta} \right\rfloor$ should be used.

Theorem 1. For any stable system, i.e., utilization $\rho = (N\alpha\omega)/[C(\alpha+\beta)] < 1$, $N-L-1 < N[\beta/(\alpha+\beta)]$ always holds.

Proof. Assume the theorem is not true, i.e., for some α, β, and N:

$$N-L-1 \geq \frac{N\beta}{\alpha+\beta}$$

$$\Rightarrow L+1 \leq N - \frac{N\beta}{\alpha+\beta}$$

$$\Rightarrow \frac{C}{\omega} < L+1 \leq N - \left(1-\frac{\alpha}{\alpha+\beta}\right)N$$

$$\Rightarrow \frac{N\alpha\omega}{(\alpha+\beta)C} > \frac{\alpha+\beta}{\alpha}\left[1-\left(1-\frac{\alpha}{\alpha+\beta}\right)\right]$$

$$\Rightarrow \rho > 1$$

which contradicts the assumption. Therefore, the theorem is always true. □

Theorem 1 implies that for the systems of interest, (8) always holds for $i = N-L-1$. The upper-bound that we found for $T_{N-(N-L-1)}$, denoted $\tilde{T}_{L+1}$, is then given by:

$$\tilde{T}_{L+1} = \frac{N-L}{N\beta\binom{N-1}{N-L-1}}\binom{N}{N-L-1} = \frac{N-L}{(L+1)}\cdot\frac{1}{\beta} \tag{9}$$

Note that, in the extreme case, i.e., when $L+1 = N$, $\tilde{T}_N$ is equal to T_N. Furthermore, for the same number of voice calls, as the value of $L+1$ decreases, i.e., the channel capacity decreases, the value of $(\tilde{T}_{L+1} - T_{L+1})$ increases. This is precisely how we want $\tilde{T}_{L+1}$ to behave, as pointed out before. We can now refine our matching procedure using the result of (9) to be the following:

i. $\gamma_2 = \dfrac{1}{\tilde{T}_{L+1}} = \dfrac{(L+1)\beta}{N-L}$

ii-iv. (*as before*).

3.4 Include video traffic to the model

To motivate the discussion of our extension to include video sources into the model, let us observe that the expected time that the system will stay in the overload state depends on two major factors: for the same number of traffic sources, the heavier the system load is, the longer each visit to the overload state is; and, the quicker the system can reduce its current load, the faster it will leave the overload state. Note that, in (9), $(N-L)/(L-1)$ is the ratio of the number of overload states to the number of underload states, which gives a measure of how heavy the system load is. The other term, $1/\beta$, is the expected time until an active source becomes idle, which is an indication of how soon the current system load will drop.

To include video traffic into the model, let us combine the models for voice and video sources to a *two-dimensional Markov chain*. As demonstrated by Fig. 5, where $M+1$ bit rate levels for video sources and N voice sources have been assumed, the states represent the bit rate level for video sources and the number of active voice sources. The steady-state distribution of this two-dimensional Markov chain, which can be easily derived from its *local balance equations* and *global balance equations*, is given by:

$$\pi_{ij} = \binom{N}{j}\binom{M}{i}\left(\frac{a}{a+b}\right)^{i}\left(\frac{b}{a+b}\right)^{M-i}\left(\frac{\alpha}{\alpha+\beta}\right)^{j}\left(\frac{\beta}{\alpha+\beta}\right)^{N-j}, \quad 0 \le i \le M, 0 \le j \le N \quad (10)$$

We say a state (i, j) is overloaded if $(\eta i + \omega j) > C$, where η is the step size of the bit rate levels for video sources; otherwise, we say that it is underloaded. Let S_u, S_o be the sets and θ_u, θ_o be the numbers of the underload states and overload states in Fig. 5 respectively; and σ_i be the number of voice sources which the system can support given that the video sources are currently in bit rate level i. Note that σ_i's must satisfy the following conditions: $\sigma_i = (C-\eta i)/\omega$ and $0 \le \sigma_i \le N$. By adopting the similar idea for voice sources, as shown in Fig. 5, we propose the following matching procedure as an approximation of the traffic due to integrated voice and video sources:

i. $$\gamma_2 = \frac{\theta_o}{\theta_u} \cdot \frac{1}{\beta+b} = \left(\sum_{i=0}^{M} (N-\sigma_i) / \sum_{i=0}^{M} (\sigma_i + 1)\right)\frac{1}{\beta+b}$$

ii. $$\lambda_1 = \sum_{(i,j) \in S_u} (\eta i + \omega j)\left(\frac{\pi_{ij}}{\Pi_u}\right), \text{ where } \Pi_u = \sum_{(x,y) \in S_u} \pi_{xy}$$

iii. $$\lambda_2 = \sum_{(i,j) \in S_o} (\eta i + \omega j)\left(\frac{\pi_{ij}}{\Pi_o}\right), \text{ where } \Pi_o = \sum_{(x,y) \in S_o} \pi_{xy}$$

iv. $\gamma_1 = \gamma_2 \dfrac{(\omega\phi_1 + \eta\phi_2) - \lambda_1}{\lambda_2 - (\omega\phi_1 + \eta\phi_2)}$, where $\phi_1 = N\dfrac{\alpha}{\alpha+\beta}$ and $\phi_2 = M\dfrac{a}{a+b}$

As before, the first term of γ_2 provides an indication as how heavy the system load is and the second term measures how quickly the current system load will drop (note that, it is the mean of a random variable that is the minimum of two exponential distributed random variables with mean of $1/\beta$ and $1/b$). Note also that step (*ii*) calculates the average cell arrival rate when the process is in an underload situation. Similarly, the third step finds the expected cell arrival rate when the process is in an overload period. The last step matches the average two-state MMPP arrival rate with that of the underlying cell arrival process.

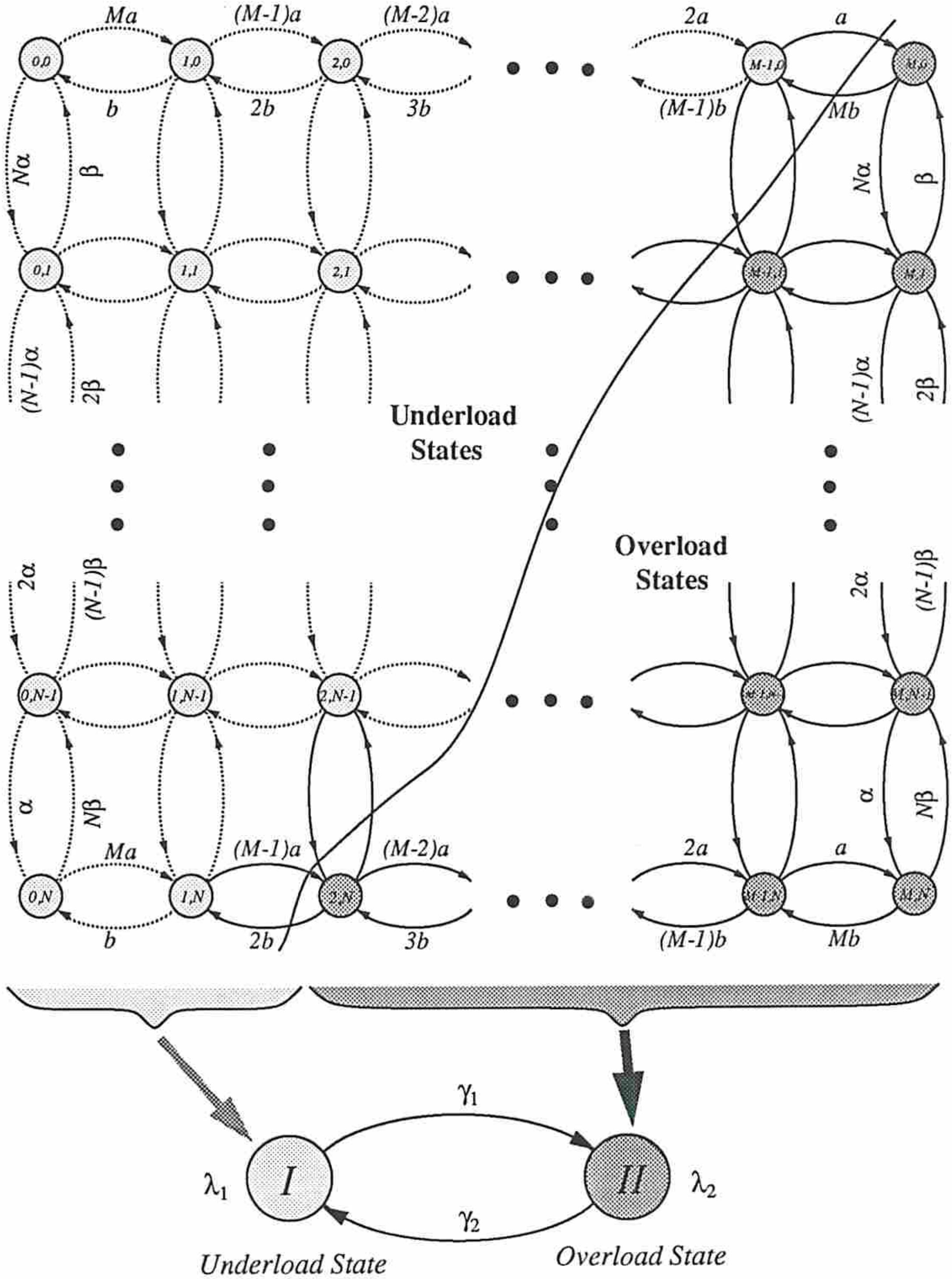

Fig. 5. Transition rate diagram and state mapping for integrated video and voice sources.

3.5 Add data traffic to the model

We now have a two-state MMPP representing the arrival process from a number of video and voice sources. To include data sources into the model, let us assume that data sources generate data packets according to a Poisson process with rate μ. The packet length has an arbitrary probability mass function, $P_k = Pr\{\text{data packet is comprised of } k \text{ cells}\}$. If we combine the data sources with the two-state MMPP, we have a two-state MMPP with batch arrivals. The transition rates, γ_1 and γ_2, are determined by the matching procedure. The arrival rates are $\lambda_1 + \mu$ and $\lambda_2 + \mu$ for the underload-state and the overload-state respectively. The probability that a batch has a size of k cells is:

$$P_k \frac{\mu}{\mu + \lambda_i} + \delta_{k-1} \frac{\lambda_i}{\mu + \lambda_i}, \quad i = 1, 2; \quad k = 1, 2, 3\ldots \tag{11}$$

where:

$$\delta_{k-1} = \begin{cases} 1, & \text{if } k = 1 \\ 0, & \text{otherwise} \end{cases}$$

4 Numerical Results and Discussions

In this section we present some numerical results followed by some discussion.

4.1 Numerical results for voice and data integration

We model a voice/data ATM multiplexer as a two-state MMPP$^{[X]}$/D/1 queue and the procedure described in [23] is applied to find the expected system time. Each voice call is characterized by $\omega = 1/6$ cells/msec (assume 64 Kbps PCM coding with speech activity detector and standard 48-octet payload), $\alpha = 1.538$ and $\beta = 2.778$ (according to the conclusions drawn by [4]). Aggregated data traffic is assumed to have an arrival rate of $20N_d/3$ packets per second, where N_d is the number of data calls. We assume that the packet size is geometrically distributed with an average of 5 cells per packet. Fig. 6 and Fig. 7 correspond to a fixed number of voice and data calls (20 voice calls plus 20 data calls for Fig. 6 and 200 voice calls plus 200 data calls for Fig. 7), i.e., fixed system load, and we plot average system time versus channel utilization. Fig. 8 shows the relationship between average system time and channel capacity with the system utilization kept unchanged at 0.9 and a fixed ratio of the number of voice and data calls at 3:5. In these figures, MMPP-1 is the model suggested by [1] extended to include data sources and MMPP-2 is the model presented in this study.

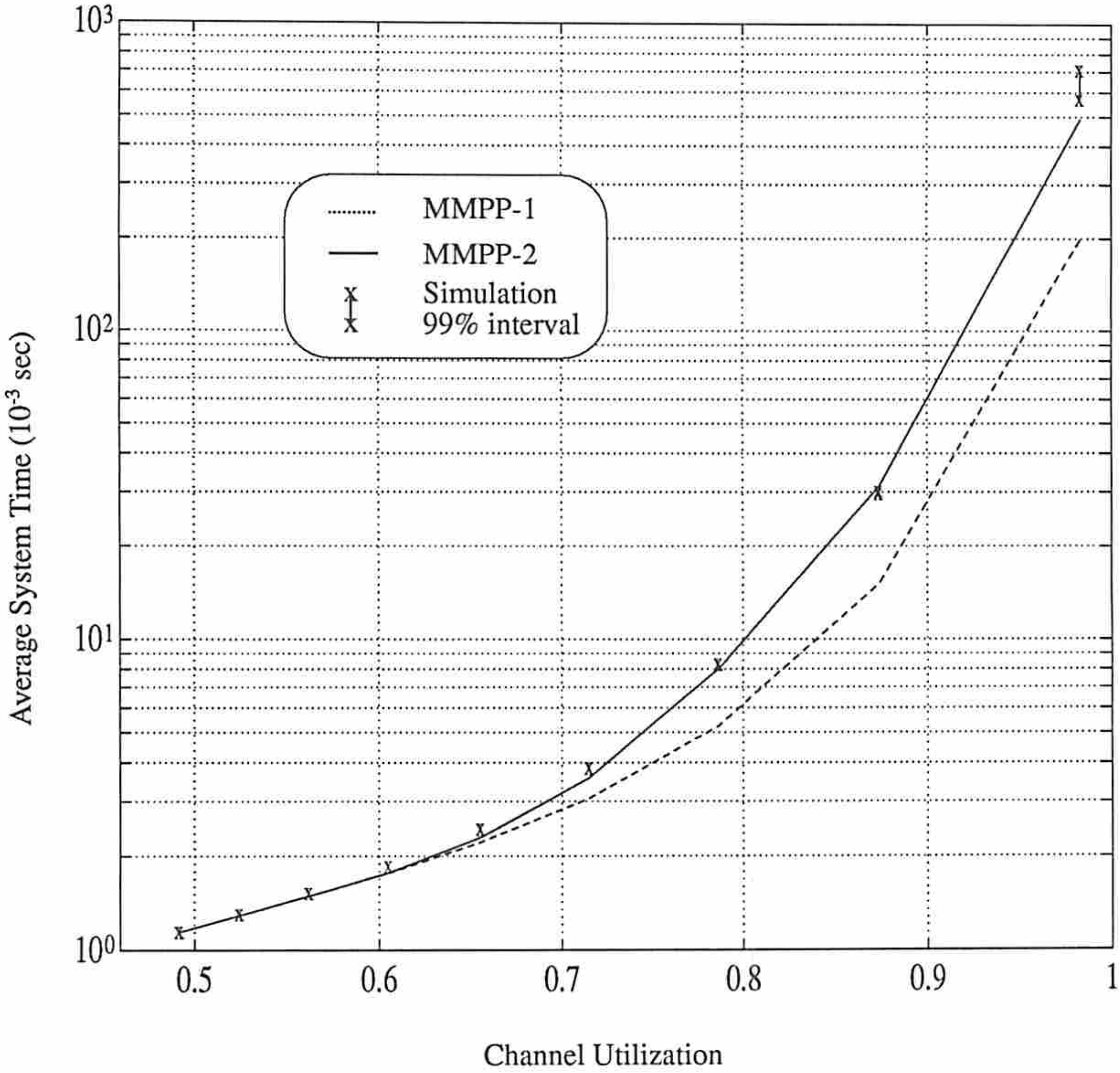

Fig. 6. Expected system time versus channel utilization for 20 voice calls and 20 data calls.

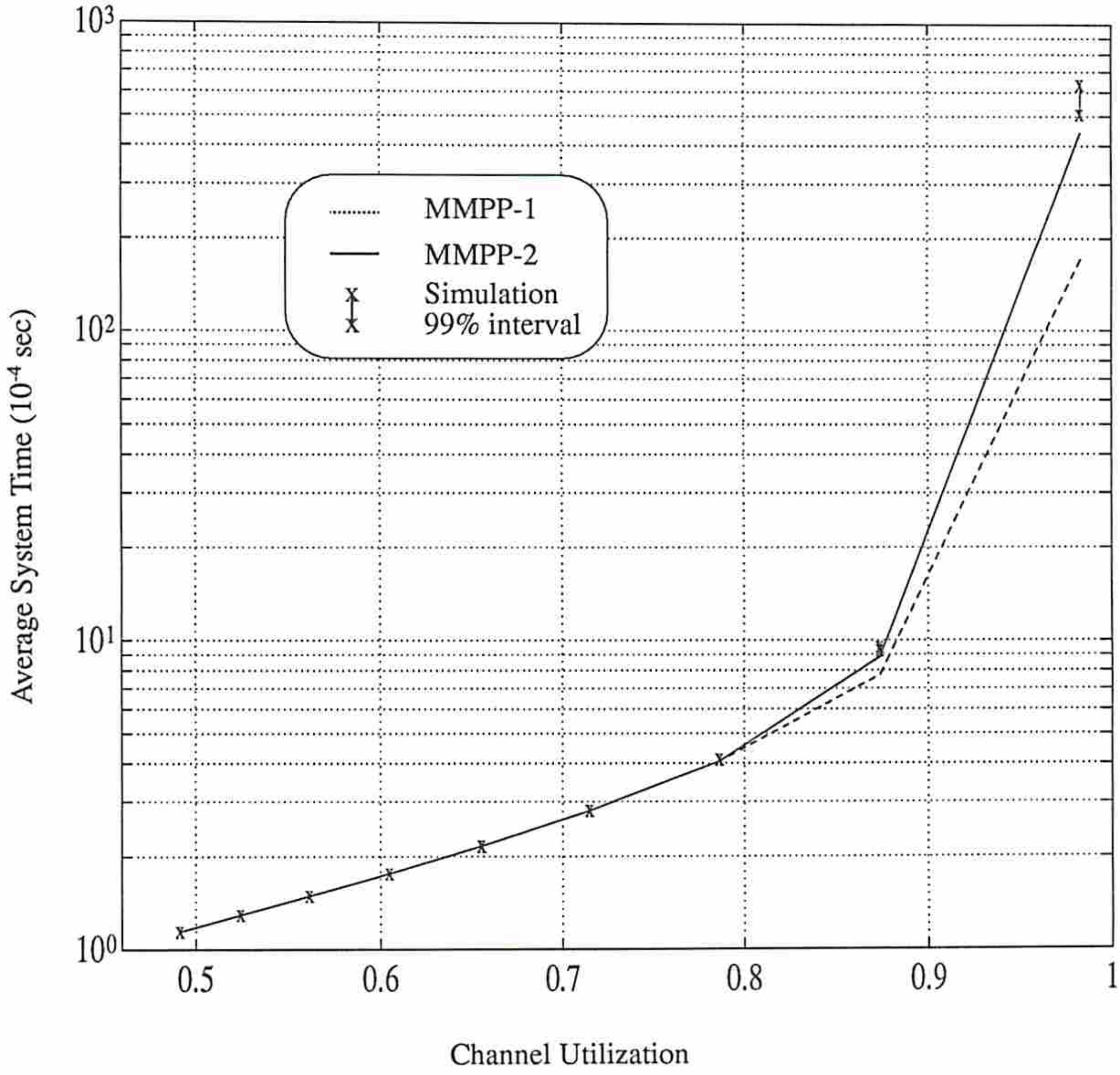

Fig. 7. Expected system time versus channel utilization for 200 voice calls and 200 data calls.

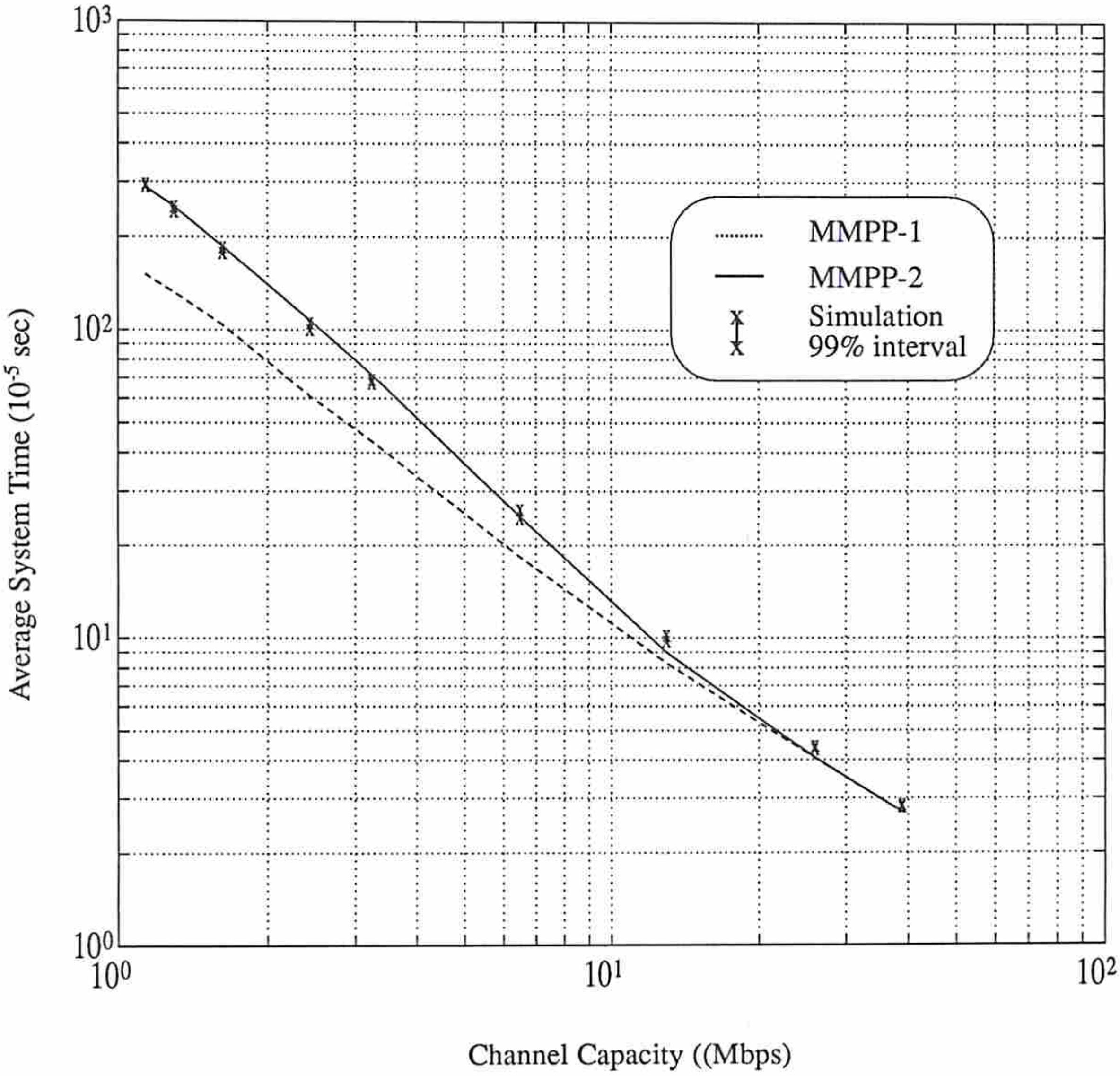

Fig. 8. Expected system time versus channel capacity for $\rho = 0.9$.

4.2 Numerical results for video, voice and data integration

We use the same voice and data sources as in the previous examples and use the same set of parameters for the video sources as used by [3] and [13], i.e., video sources are characterized by: an average bit rate of 3.9 Mbps, a peak bit rate of 10.58 Mbps, a standard deviation of the bit rate of 1.73 Mbps and a parameter for the autocorrelation function of 3.9. The total number of discrete bit rate levels for video sources is assumed to be 16 times the number of video sources. In Fig. 9, we assume 100 voice sources and 100 data sources in the background and plot the relation between average system time and the number of video sources under a system utilization of 0.8 (varying the channel capacity). Our model is shown as curve MMPP-2 and for comparison purposes we plot a curve MMPP-1 using the maximum of $1/\beta$ and $1/b$ rather then $1/(\beta+b)$ as the measure for the time until of the current system load will drop. In Fig. 10, we keep the ratio of the number of video, voice and data calls fixed at 1:20:20 and show average system time versus channel capacity with a constant system utilization of 0.9 (i.e., by varying the total load).

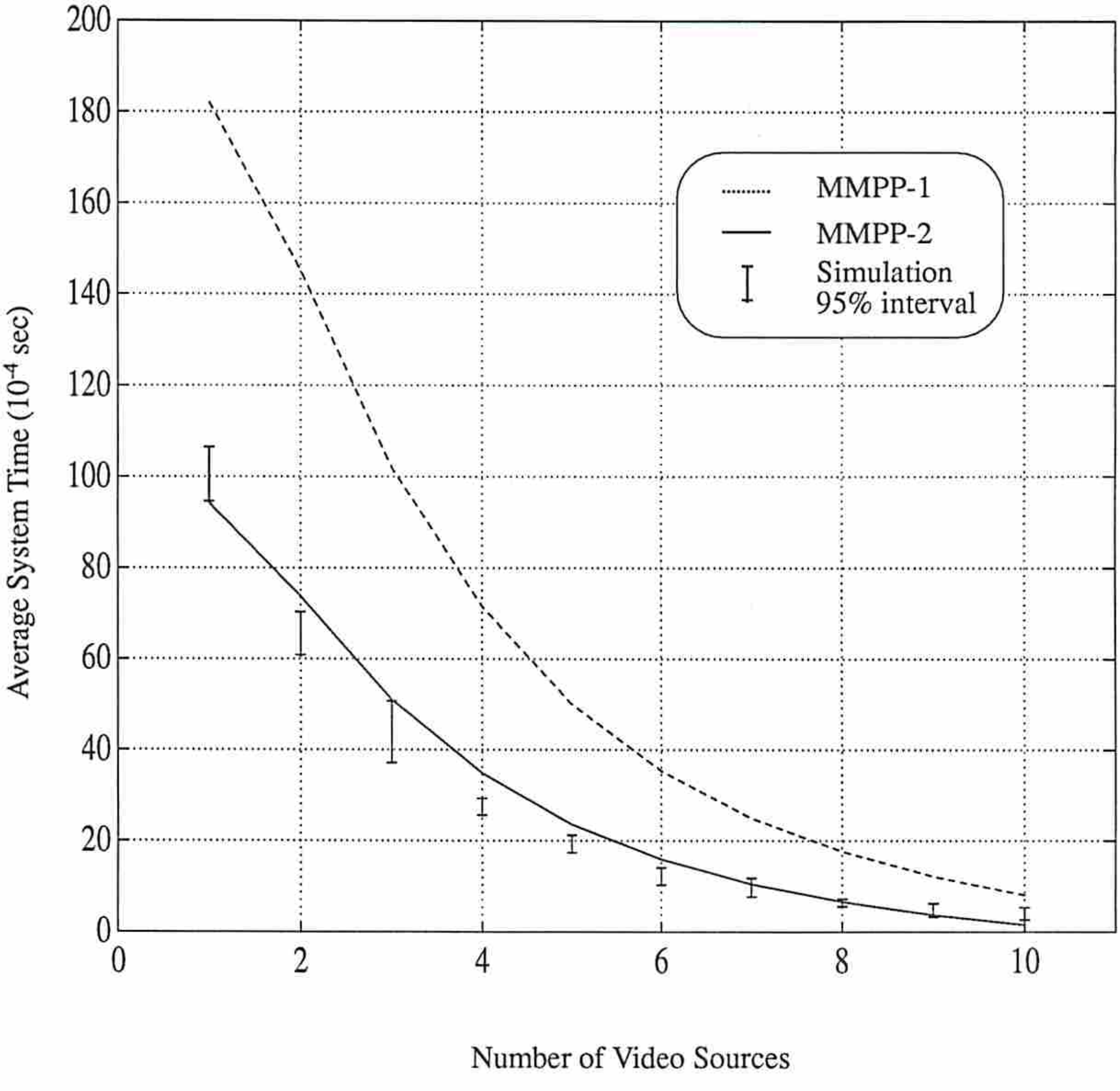

Fig. 9. Expected system time versus number of video sources for $\rho = 0.8$.

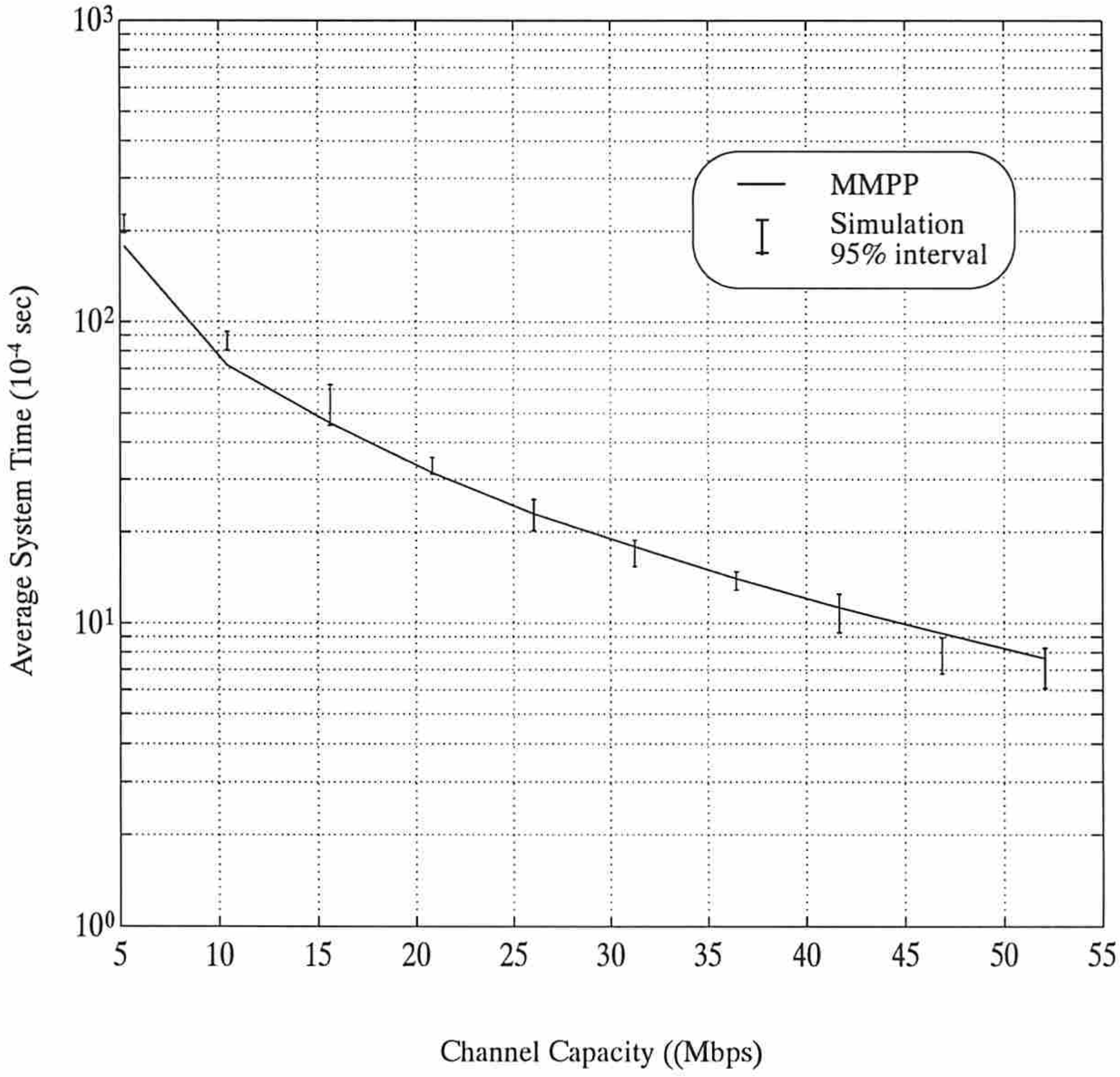

Fig. 10. Expected system time versus channel capacity for $\rho = 0.9$.

4.3 Discussions

Several advantages of our model can be identified. *First, a lower computation overhead is involved.* In [1], solving an (L+1)-dimensional eigen-system is required in order to find the parameters for the MMPP, a significant effort for a large L. Whereas in our model, the computational requirement is negligible. *Second, a real-time traffic control algorithm can be developed.* Note that if the system is in an overload situation, the queue size is expected to build-up until the system returns to an underload status. Hence the expected length of the overload periods, $1/\gamma_2$, can be used as an indicator for the seriousness of an overload period, once the system is overloaded. Using our results, $1/\gamma_2$ can be computed very quickly for any given number of video, voice and data calls and used as a criterion for traffic control. *Third and more importantly, it provides better performance predictions*. As it was pointed out in section 2 and 3, with increased system load the model used by [1] severely underestimates the average system time, whereas our approach agrees with the simulation very well (see Fig. 6 through Fig. 8). For video, voice and data integration, although we provide no comparison to other models (since no similar models exist), simulation results show that the approximation is very good even under heavy system load (see Fig. 9 and Fig. 10).

5 Conclusions

In this paper, we study the performance of an ATM multiplexer loaded with video, voice and data traffic. The actual arrival process is approximated by a two-state Markov modulated Poisson process with batch arrivals; and the multiplexer is modeled as a two-state $\mathrm{MMPP}^{[X]}/D/1$ queue.

The modeling technique developed here leads to very accurate results and has very low computational requirements; hence it is useful for modeling real-time traffic modeling.

We note that our results are for mean performance statistics. Additional work is required to see if the accuracy holds for higher moments of these statistics. More studies are also necessary to study the effectiveness of this approach for loss models (limited buffer space).

References

[1] A. Baiocchi et al., "Loss Performance Analysis of an ATM Multiplexer Loaded with High-Speed On-Off Processes," *IEEE J. Select. Areas Commun.*, vol. 9, pp. 388-393, Apr. 1991.

[2] C. Blondia, "The N/G/1 Finite Capacity Queue," *Stochastic Models*, vol. 5, no. 2, pp. 273-294, 1989.

[3] C. Blondia and O. Casales, "Performance Analysis of Statistical Multiplexing of VBR Sources," in *Proc. INFOCOM'92*, pp. 6C.2.1-6C.2.11.

[4] P. T. Brady, "A Statistical Analysis of On-Off pattern in 16 Conversations," *Bell Systems Technical Journal*, pp. 73-91, Jan. 1968.

[5] P. T. Brady, "A Model for Generating On-Off Speech Patterns in Two-Way Conversation," *Bell Systems Technical Journal*, pp. 2445-2472, Sep. 1969.

[6] J. N. Daigle and J. D. Langford, "Models for Analysis of Packet Voice Communications Systems," *IEEE J. Select. Areas Commun.*, vol. 4, pp. 847-855, Sep. 1986.

[7] R. L. Graham, *Concrete Mathematics: A Foundation for Computer Science*, Menlo Park, CA: Addison-Wesley, 1989.

[8] H. Heffes and D. M. Lucantoni, "A Markov Modulated Characterization of Packetized Voice and Data Traffic and Related Statistical Multiplexer Performance," *IEEE J. Select. Areas Commun.*, vol. 4, pp. 856-868, Sep. 1986.

[9] I. Ide, "Superposition of Interrupted Poisson Processes and Its Application to Packetized Voice Multiplexers," in *Proc. ITC '88*, pp. 3.1B2.

[10] C. Knessl and B. J. Matkowsky, "A Markov Modulated M/G/1 Queue I: Stationary Distribution," *Queueing Systems*, vol. 1, pp. 355-374, 1987.

[11] S.-Q. Li, "A General Solution Technique for Discrete Queueing Analysis of Multimedia Traffic on ATM," *IEEE Trans. Commun.*, vol. 39, pp. 1115-1132, Jul. 1991.

[12] D. Lucantoni, "New Results on the Single Server Queue with a Batch Markov Arrival Process," *Stochastic Models*, vol. 7, no. 1, pp. 1-46, 1991.

[13] B. Maglaris et al., "Performance Models of Statistical Multiplexing in Packet Video Communications," *IEEE Trans. Commun.*, vol. 36, pp. 834-844, Jul. 1988.

[14] B. Mukherjee, "Integrated Voice-Data Communications Over High-Speed Fiber Optic Networks," *Computer*, pp. 49-58, Feb. 1991.

[15] R. Nagarajan et al., "Approximation Techniques for Computing Packet Loss in Finite-Buff-

ered Voice Multiplexers," *IEEE J. Select. Areas Commun.*, vol. 9, pp. 368-377, Apr. 1991.

[16] M. F. Neuts, *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*, Baltimore, MD: Johns Hopkins Univ. Press, 1981.

[17] M. F. Neuts, "A Versatile Markovian Point Process," *J. Applied Prob.*, vol. 16, pp.764-779, Dec. 1979.

[18] N. U. Prabhu and Y. Zhu, "Markov Modulated Queueing Systems," *Queueing Systems*, vol. 5, pp. 215-246, 1989.

[19] G. Ramamurthy and B. Sengupta, "Delay Analysis of a Packet Voice Multiplexer by the $\Sigma D_i/D/1$ Queue," *IEEE Trans. Commun.*, vol. 39, pp. 1107-1114, Jul. 1991.

[20] V. Ramaswami, "The N/G/1 Queue and Its Detailed Analysis," *Adv. Appl. Prob.*, vol. 12, pp. 222-261, Mar. 1980.

[21] K. Sriram and W. Whitt, "Characterizing Superposition Arrival Processes in Packet Multiplexers for Voice and Data," *IEEE J. Select. Areas Commun.*, vol. 4, pp. 833-846, Sep. 1986.

[22] R. C. F. Tucker, "Accurate Method for Analysis of a Packet-Speech Multiplexer with Limited Delay," *IEEE Trans. Commun.*, vol. 36, pp. 479-483, Apr. 1988.

[23] S. S. Wang and J. A. Silvester, "A Model for the Performance Analysis of Voice/Data ATM Multiplexers," *Technical Report*, CENG 92-07, University of Southern California, Jun. 1992.

[24] C. Yuan and J. A. Silvester, "Queueing Analysis of Delay Constrained Voice Traffic in a Packet Switching System," *IEEE J. Select. Areas Commun.*, vol. 7, pp. 729-738, Jun. 1989.

[25] Y. Zhu, "A Markov Modulated M/M/1 Queue with Group Arrival," *Queueing Systems*, vol. 8, pp. 255-264, Apr. 1991.