UNIVERSITY
OF TRENTO

DEPARTMENT OF INFORMATION AND COMMUNICATION TECHNOLOGY

38050 Povo – Trento (Italy), Via Sommarive 14
http://www.dit.unitn.it

A PERFORMANCE MODEL FOR MULTIMEDIA SERVICE PROVISIONING ON NETWORK INTERFACES.

Damiano Carra

Paola Laface

Renato Lo Cigno

August 2004

Technical Report # DIT-04-092

A Performance Model for Multimedia Services Provisioning on Network Interfaces*

D. Carra[1], P. Laface[2], R. Lo Cigno[1]

[1]Dipartimento di Informatica e Telecomunicazioni – Università di Trento

[2]Dipartimento di Elettronica – Politecnico di Torino

29th September 2004

Keywords: Multimedia Traffic, QoS Network Planning, Markov Modeling, Approximate Solutions

Abstract

This paper presents a method for the performance evaluation of multimedia streaming applications on IP network interfaces with differentiated scheduling. Streaming applications are characterized by the emission of data packets at constant intervals, changing during the connection lifetime, hence a single source can be effectively modeled by an MMDP.

We propose an MMDP/D/1/K model to represent the aggregate arrival process and the network interface, assuming that multimedia packet dimensions are approximately constant. A method for solving the above queuing system providing upper and lower bounds to the packet loss rate is presented and results for realistic VoIP applications are discussed and validated against accurate event-driven simulations showing the efficiency and accuracy of the method.

1 Introduction

Multimedia and Quality of Service (QoS) are probably the most repeated words in networking research during the past fifteen years or so. Spawned by research on ATM (Asynchronous Transfer Mode) in the late '80s and early '90s, topics related to offering the appropriate QoS in integrated packet networks received even more attention when the application context moved to IP-based networking.

One of the key aspects of heterogeneous service provisioning is the guarantee of the QoS the service will receive during its lifetime. Enforcing QoS encompasses a number of

*This work was supported in Torino by the Italian Ministry for University and Research (MIUR) under the FIRB project TANGO

different aspects, ranging from service architecture, to network dimensioning, to protocol design and many others. All the design aspects pivot around the performance evaluation of the provisioned service: without a means to evaluate the performance, it is not possible to select an appropriate Service Level Agreement (SLA) between the network and the user.

In this paper we explore an analytical approach based on the solution of DTMCs (Discrete Time Markov Chains) embedded in a more general CTC (Continuous Time Chain) to evaluate the performance of several classes of multimedia services, namely those, as voice and video, that are characterized by intermittent or variable bit rate, and piecewise constant inter-packet emission times. After describing the general framework, we focus our attention on voice services comparing the analytical solution with simulations. We consider IP Telephony, or VoIP for short, which is by far the most diffused (and widespreading), multimedia application on IP networks.

The main contribution of this work is providing a simple and efficient analytical framework to predict the performance of multimedia services in several possible scenarios, like for instance an access link to the Internet or DiffServ [1] interfaces with *Expedited Forwarding Per Hop Behavior* [2]. Though the mathematical modeling is not entirely novel (as discussed in Sect. 2), since similar problems were tackled studying ATM networks (see [3, 4] and the references therein, or [5] for a review), the solution technique we propose yields the derivation of exact solutions for upper and lower bounds on the packet loss performance, which enables service planning and SLA definition. The bounds are shown to be tight. We generalize the solution technique to the case of n-state sources, transmitting at different rates in each of the n states, while known solutions only address On-Off sources.

To conclude with, we note that the approximations we introduce for analytical tractability are equivalent, in many aspects, to fluid flow approximations, but they ensure exact bounding of the solution (instead of a generic approximation), and do not suffer from numerical instabilities in the solution.

2 Problem Formulation

Multimedia services are related to audio and video, i.e., to connection oriented, streaming applications, whose emission characteristics are studied fairly well. Indeed, although voice and video can potentially generate packets at variable intervals, all existing applications encode blocks of information at fixed lengths intervals. Moreover, the packet size is often fixed, either by the encoder, the protocol or the application program. In other words, multimedia streaming or conversational applications can be efficiently modeled as sources with a piecewise constant emission rate of fixed size packets.

In queuing theory notation, these are Modulated Deterministic Processes (MDP). If the time between state transitions is exponentially distributed and the source can transmit with n different speeds, then the source is an n-state Markov Modulated Deterministic Process (MMDP), whose states s_k are characterized by the emission of packets with rate

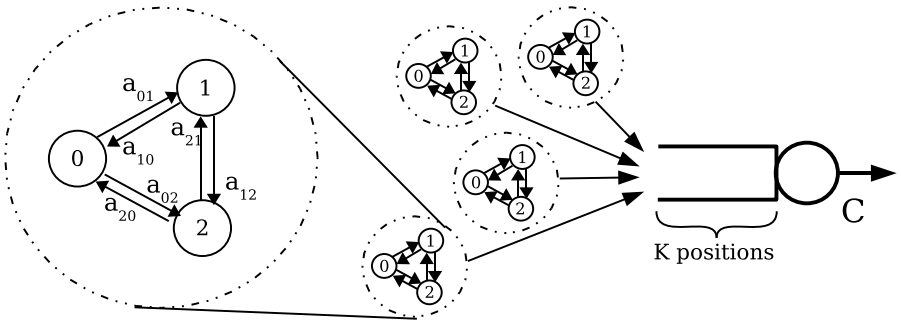

Figure 1: Arrival process: superposition of three-state sources

δ_k. On-Off sources are a subset with two states and $\delta_0 = 0$ and $\delta_1 = c \neq 0$.

A superposition of MMDP is not an MMDP, unless the sources are appropriately synchronized. Source synchronization is not an unrealistic scenario (VoIP calls generated by the same media conversion gateway are either synchronized or can be synchronized easily). Approximating the superposition of MMDPs with a single MMDP is equivalent to neglect short term congestion, due to the arrival process higher variability. The impact of this approximation is discussed in Sect. 5, where event driven simulations are used to validate both the modeling assumptions and the bounding approximations.

We consider a single network interface (can be at the access or in any section of the network) assuming it is the only point of potential congestion, and model the system as an MMDP/D/1/K queuing system. Fig. 1 shows an example of the system with four three-state sources. The states of each source are labeled as 0,1,2, with transmission rates δ_0^i, δ_1^i, δ_2^i, not necessarily equal one another.

The number of states m of the arrival modulating process is, in the most general case, $m = \prod_{i=1}^{M} n_i$, where M is the number of considered sources and n_i is the number of possible transmission rates of source i. No constraints are posed on δ_k^i, but all packets are of the same length for deterministic service times.

This same modeling approach was taken in [3] limiting the analysis to homogeneous On-Off sources. When possible we use the same notation used in that work, to help readers familiar with it.

For homogeneous sources, the arrival modulating process is an $(n-1)$-dimensional quasi-birth-death CTMC. Fig. 2 shows the modulating CTMC for $n = 3$, where the the first index i is the number of sources in state 1 (say high transmission rate), the second index j is the number of sources in state 2 (say low transmission rate) and $M - i - j$ sources are in state 0 (for example silent). Evolution along rows and columns follows a simple birth-and-death process; evolution along diagonals means that a source can switch from high to low and vice versa.

Although the arrival process is modulated by a CTMC, the overall queuing system is not Markovian, since phase transitions (i.e., transitions of the above CTMC) can occur at any time between arrivals (deterministic) and departures (again deterministic), which do not

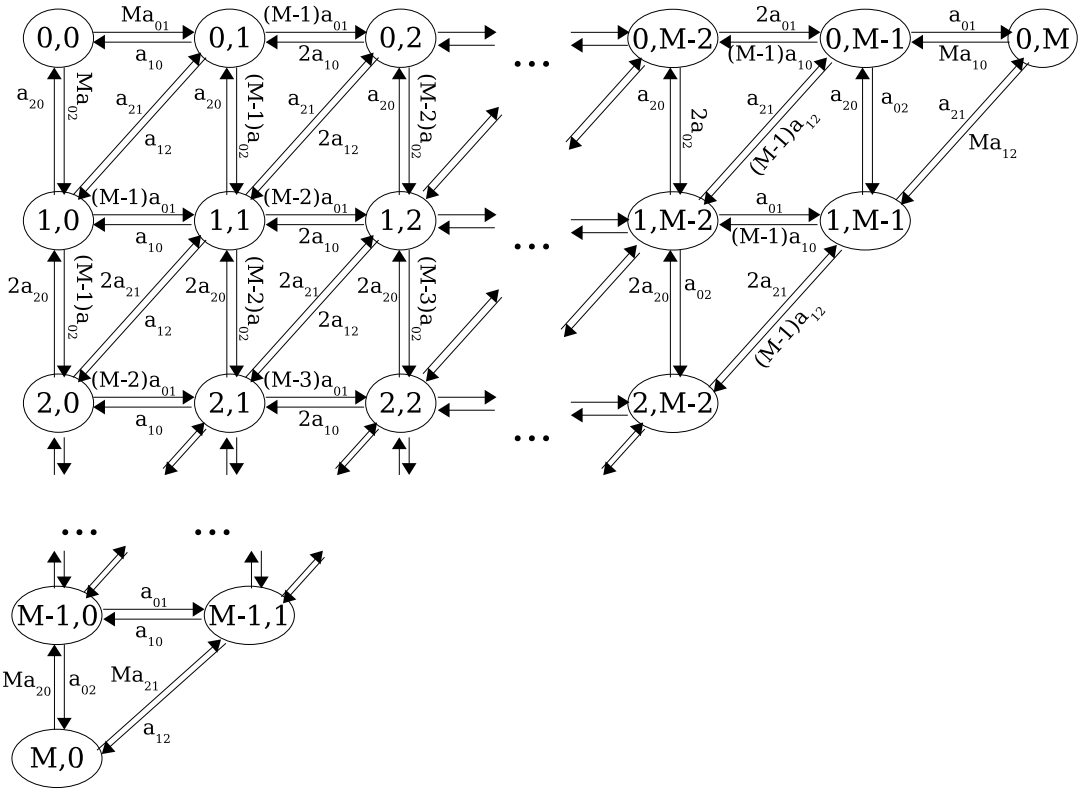

Figure 2: Markov chain describing the evolution of the MMDP deriving from the superposition of M sources with three different emission rates

enjoy memoryless properties.

3 Model Analysis

Define:

- $\{X(t), t \geq 0\}$: the finite, irreducible CTMC representing the arrival process;
- $\boldsymbol{S} = \{\underline{f}\}$: the state space of $X(t)$; The vector $\underline{f} = [f_0, f_1, \ldots, f_F]$ represents the arrival process phase; $f_k = 0, 1, \ldots, M_k$ is the number of sources sending at rate δ_k; if sources are all homogeneous $F = n - 1$ and $M_k = M,\ \forall k$; if sources are all different $F = \prod_{i=1}^{M} n_i$ and $M_k = 1,\ \forall k$;
- $\boldsymbol{R} = [r_{\underline{f},\underline{f}'}]$: the $m \times m$ infinitesimal generator of $X(t)$, $m = ||\boldsymbol{S}||$;
- $1/\gamma_{\underline{f}}$ ($\underline{f} \in \boldsymbol{S}$): the mean sojourn time in state $\underline{f}$

Two examples help understanding the system.

Example 1: heterogeneous On-Off sources. We have F types of On-Off sources, M_0 with emission rate δ_0, M_1 with emission rate δ_1, $\cdots$, M_F with emission rate δ_F, $M = \sum_{i=0}^{F} M_i$; the components f_k of vector $\underline{f}$ represent the number of active sources with rate δ_k. The cardinality of $\boldsymbol{S}$ is $m = (M_0 + 1) \cdot (M_1 + 1) \cdot \ldots \cdot (M_F + 1)$, and the CTMC is a combination of birth-death processes, i.e., only transitions of the type

$$\begin{array}{rcl} [f_0, \ldots, f_i, \ldots, f_F] & \rightarrow & [f_0, \ldots, f_i + 1, \ldots, f_F]; \\ [f_0, \ldots, f_i, \ldots, f_F] & \rightarrow & [f_0, \ldots, f_i - 1, \ldots, f_F] \end{array}$$

are allowed.

Example 2: homogeneous multirate sources. We have M sources that can transmit n different rates δ_k, $k = 0, 1, \ldots, n$; the components f_k of vector $\underline{f}$ still represent the number of active sources with rate δ_k; we have $F = n+1$ and $m = \binom{M + n - 2}{M}$, since we don't have to explicitly represent the number f_0 of sources with rate δ_0 because the relation $M = \sum_{k=0}^{F} f_k$ holds. The CTMC is no more the product of birth-death chains and 'diagonal' transitions (as shown in Fig. 1) are admitted, thus we have three possible transition types:

$$\begin{array}{rcl} [f_1, \ldots, f_i, \ldots, f_F] & \rightarrow & [f_1, \ldots, f_i + 1, \ldots, f_F]; \\ [f_1, \ldots, f_i, \ldots, f_F] & \rightarrow & [f_1, \ldots, f_i - 1, \ldots, f_F]; \\ [f_1, \ldots, f_i - 1, \ldots, f_F] & \rightarrow & [f_1, \ldots, f_j + 1, \ldots, f_F]. \end{array}$$

Let C be the service rate in packets per second and $Y(t) = k \leq K$ the number of packets in the buffer at the time t. The process $\{(X(t), Y(t)), t > 0\}$ represents exactly the MMDP/D/1/K queue we consider.
Let ξ_n (n = 1,2, ... with $\xi_0 = 0$), be the transition epochs of $X(t)$. Sampling the process $\{(X(t), Y(t)), t > 0\}$ in the instants defined by the sequence ξ_n, we obtain an embedded DT process $\{(X_n, Y_n), n = 0, 1, ...\}$, where

- $X_n = X({\xi_n}^+)$ is the state of the modulating Markov process at time ${\xi_n}^+$;
- $Y_n = Y({\xi_n}^+)$ is the number of packets in the buffer (including the one being served) at time ${\xi_n}^+$.

We define

- $V_{\underline{f}}$: the arrival packet rate when $X_n = \underline{f}$ ($\underline{f} \in \boldsymbol{S}; n = 0, 1, \ldots$);
- $U_{\underline{f}} = \xi_{n+1} - \xi_n$: the time interval during which $X(t)$ is in state $\underline{f}$. It is (by construction) a random variable exponentially distributed with parameter $\gamma_{\underline{f}}$.

$\{(X_n, Y_n), n = 0, 1, ...\}$ is non-Markovian, since sampling ξ_n can happen during a packet service and between packet arrivals which are not exponentially distributed. To the best of our knowledge, an exact analysis is impossible, but neglecting either the residual or

elapsed service and interarrival time we obtain four different approximated DTMCs. Formally this is equivalent to make a *service* and *arrival* renewal assumptions.

Service renewal assumption. The elapsed service time T_n^s after the n-th transition is equal either to 0 or to $1/C$. $T_n^s = 0$ (the packet is not yet served) overestimates the system load; $T_n^s = 1/C$ (the packet is served entirely) underestimates the system load.

Arrival renewal assumption. The elapsed inter-arrival time T_n^a after the n-th transition is equal either to 0 or to $1/V_{\underline{f}}$. $T_n^a = 0$ (the first packet of the new phase arrives following the new rate and the last packet of the previous phase is neglected) underestimates the system load; $T_n^a = 1/V_{\underline{f}}$ (the last packet of the old phase arrives in any case on phase transition, no matter what) overestimates the system load.

Under these assumptions we can get four cases:

LL:	if $T_n^s = 1/C$ and $T_n^a = 0$	UU:	if $T_n^s = 0$ and $T_n^a = 1/V_{\underline{f}}$
LU:	if $T_n^s = 1/C$ and $T_n^a = 1/V_{\underline{f}}$	UL:	if $T_n^s = 0$ and $T_n^a = 0$

The interesting cases are LL and UU, that yield a lower and upper bound to the system load and hence on the loss probability, while the UL and LU cases are approximations, but it is not easy to tell whether they are upper or lower bounds. It must be noted at this point that the authors in [3], besides considering only homogeneous On-Off sources ($F = 1$ and the state identifier $\underline{f}$ is a scalar and not a vector), approximate the $\{(X_n, Y_n), n = 0, 1, ...\}$ process with a limiting passage, assuming that $E[U_{\underline{f}}] \gg E[1/V_{\underline{f}}], 1/C$, which is, for $E[1/V_{\underline{f}}] \to 0$ and $1/C \to 0$, equivalent to the UL approximation.
The renewal assumptions introduced refer only the buffer evolution (the Y_n component of the (X_n, Y_n) process). They approximate the real evolution of the buffer with incremental geometric random variables. In formulas:

$$Y_{n+1} = \begin{cases} \min\{K, Y_n + I_{\underline{f}}\}, & \text{if } V_{\underline{f}} > C \\ \max\{0, Y_n - O_{\underline{f}}\}, & \text{if } V_{\underline{f}} < C \end{cases} \tag{1}$$

where

$$I_{\underline{f}} = \begin{cases} \lfloor (V_{\underline{f}} - C)U_{\underline{f}} \rfloor & \text{in UL and LL} \\ \lceil (V_{\underline{f}} - C)U_{\underline{f}} \rceil & \text{in LU and UU} \end{cases} \tag{2}$$

and

$$O_{\underline{f}} = \begin{cases} \lfloor (C - V_{\underline{f}})U_{\underline{f}} \rfloor & in\ UL\ and\ UU \\ \lceil (C - V_{\underline{f}})U_{\underline{f}} \rceil & in\ LL\ and\ LU \end{cases} \tag{3}$$

are the geometric increments with parameter $\rho_{\underline{f}} = exp\{-\gamma_{\underline{f}}/|V_{\underline{f}} - C|\}$, where $\gamma_{\underline{f}}$ is the average holding time of state $\underline{f}$.

3.1 Solution of the $\{(X_n, Y_n), n = 0, 1, ...\}$ embedded Markov chain

Let

$$a_{k,h}^{\underline{f}} = P\{Y_{n+1} = h | X_n = f, Y_n = k\} \tag{4}$$

be the probability that the number of packets in the buffer passes from k to h while the arrival process is in phase $\underline{f}$, and $\mathbf{A}_{\underline{f}} = [a^{\underline{f}}_{k,h}]$ of dimension $(K+1) \times (K+1)$ the transition probability matrix (notice that it is a stochastic matrix by construction). Then the transition probability from state $(\underline{f}, k)$ to state $(\underline{f}', h)$ of (X_n, Y_n) is:

$$q_{(\underline{f},k),(\underline{f}',h)} = a^{\underline{f}}_{k,h} p_{\underline{f},\underline{f}'}$$

and the transition probability matrix of (X_n, Y_n) is

$$\mathbf{Q} = [q_{(\underline{f},k),(\underline{f}',h)}].$$

The renewal assumptions defined in Sect. 3 affects only the $a^{\underline{f}}_{k,h}$ distribution, while the structure, and the solution of the chain are unaffected. The computation of this distribution is straightforward, though it can be a bit cumbersome. Appendix A reports the detailed computation for the UU and LL assumptions we use in this paper. The resulting DTMC is finite and ergodic by construction.

Q is highly structured and can be recursively partitioned into blocks changing one of the $\underline{f}$ components at a time. Recall that in general we have at most M_k sources that can transmit with rate δ_k which corresponds to the state component f_k, then we have

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Q}_{0,0} & \mathbf{Q}_{0,1} & \mathbf{Q}_{0,2} & \cdots & \mathbf{Q}_{0,M_0} \\ \mathbf{Q}_{1,0} & \mathbf{Q}_{1,1} & \mathbf{Q}_{1,2} & \cdots & \mathbf{Q}_{1,M_0} \\ \mathbf{Q}_{2,0} & \mathbf{Q}_{2,1} & \mathbf{Q}_{2,2} & \cdots & \mathbf{Q}_{2,M_0} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \mathbf{Q}_{M_0-1,0} & \mathbf{Q}_{M_0-1,1} & \mathbf{Q}_{M_0-1,2} & \cdots & \mathbf{Q}_{M_0-1,D_0} \\ \mathbf{Q}_{M_0,0} & \mathbf{Q}_{M_0,1} & \mathbf{Q}_{M_0,2} & \cdots & \mathbf{Q}_{M_0,D_0} \end{bmatrix} = [\mathbf{Q}_{f_0,f_0'}] \tag{5}$$

The block $\mathbf{Q}_{f_0,f_0'}$ refers to transitions from the states where the first component of $\underline{f}$ is equal to f_0 to the states where the first component amounts to f_0'.

The block decomposition can be iterated and the general form of a block is

$$\mathbf{Q}^{f_0,f_1,\ldots f_k-2}_{f_{k-1},f'_{k-1}} = [\mathbf{Q}^{f_0,f_1,\ldots f_{k-1}}_{f_k,f'_k}] \qquad k = 1,2,\ldots F \tag{6}$$

that refers to transitions from the states where the number of sources transmitting at rate $\delta_0, \delta_1, \ldots, \delta_{k-1}$ is fixed to $f_0, f_1, \ldots f_{k-1}$ and the number of sources transmitting at rate δ_k passes from f_k to f_k'.

Finally, the last block partitioning is

$$\mathbf{Q}^{f_0,f_1,\ldots f_{F-1}}_{f_F,f'_F} = [p_{\underline{f},\underline{f}'}\mathbf{A}_{\underline{f}}] \tag{7}$$

The main performance index we're interested in is the packet loss probability P_l. Given the system structure, losses can occur only in states $\underline{f}$ for which $V_{\underline{f}} \geq C$, with the equality holding for the UU approximation and not for the LL one. Since services and arrivals are

deterministic within a single phase of the arrival process, P_l can be computed starting from the excess arrivals within phases:

$$P_l = \frac{\sum_{\underline{f}:V_{\underline{f}} \geq C} \sum_{j=0}^{K} E[R_{\underline{f},j}]\pi_{\underline{f},j}}{\sum_{\underline{f}=0}^{m-1} \sum_{j=0}^{K} E[N_{\underline{f},j}]\pi_{\underline{f},j}} \tag{8}$$

where

- $\underline{\pi}_{\underline{f}} = (\pi_{\underline{f},0}, \pi_{\underline{f},1}, \ldots, \pi_{\underline{f},K})$ is the steady state distribution of the embedded DTMC defined by $\mathbf{Q}$;
- $E[N_{\underline{f},j}]$ is the average number of packets arriving in phase $\underline{f}$ given that the number of packets in the buffer at the beginning of the phase is j;
- $E[R_{\underline{f},j}]$ is the average number of packets rejected in the above conditions.

$E[N_{\underline{f},j}]$ and $E[R_{\underline{f},j}]$ depend on the distribution of $a^{\underline{f}}_{k,h}$. Their computation is reported in Appendix A.

3.1.1 Solution method: Ad-Hoc Block Reduction

The numerical solution of the system may pose problems as the dimension of the matrix $\mathbf{Q}$ increases. Recall that $\mathbf{Q}$ has dimension $[m \cdot (K+1)] \times [m \cdot (K+1)]$, so that as soon as the number of sources and the buffer dimension increase above a few tens the dimension of $\mathbf{Q}$ grows to thousands.
If we restrict the analysis to homogeneous sources or to a limited number of source classes (which is the problem we're interested in), $\mathbf{Q}$ has a banded structure, so that efficient Block Reduction techniques [6] can be used to solve the linear system. Unfortunately the block and band structure depends on the transition structure of arrival modulating process, so that a general description is cumbersome, and a case-by-case analysis is required to obtain the best solution. Appendix B reports the structure for 3-state sources, while in the following we concentrate on 2-state sources for VoIP applications.

3.2 Application to packetized voice

Packetized voice applications are characterized by the presence of VAD (Voice Activity Detector) devices, that suppress the voice encoding when the speaker is silent and transmit *silence descriptors* for comfort noise instead of voice, at a much lower transmission rate. Voice packets have a constant dimension that depends on the encoder and silence descriptors have a constant dimension that in the general case can be different from the one of voice packets. Our model however dictates constant service times regardless of the source state, so we assume that voice and silence packets are equal in size.
A voice source can be described as a two-state source that in the On state (voice encoding) generates packets equally spaced at fixed rate δ_1; when the source is Off, it generates

silence description packets equally spaced at fixed rate δ_0. On and Off holding times are $1/\lambda$ and $1/\mu$ respectively. Given M of these sources we obtain an (M+1)-state MMDP modulating the arrivals. The state $\underline{f}$ is monodimensional with one component $f_1 = i$ to simplify the notation. The embedded chain transition probabilities are:

$$p_{\underline{f},\underline{f}'} = p_{i,j} = \begin{cases} (M-i)\lambda/\gamma_i & i = 0, ..., M-1;\ j = i+1 \\ i\mu/\gamma_i & i = 1, ...M;\ j = i-1 \\ 0 & otherwise \end{cases} \tag{9}$$

where $\gamma_i = (M-i)\lambda + i\mu$.
When the number of active voice sources is i, we have an aggregate arrival rate $V_i = i\delta_1 + (M-i)\delta_0$.
The probability transition matrix $\mathbf{Q}$ has the following banded structure,

$$\mathbf{Q} = \begin{bmatrix} \mathbf{0} & \alpha_0\mathbf{A}_0 & & ... & \\ \beta_1\mathbf{A}_1 & \mathbf{0} & \alpha_1\mathbf{A}_1 & ... & \\ & \beta_2\mathbf{A}_2 & \mathbf{0} & ... & \\ & & & ... & \alpha_{M-1}\mathbf{A}_{M-1} \\ & & & \beta_M\mathbf{A}_M & \mathbf{0} \end{bmatrix} \tag{10}$$

where $\mathbf{A}_i$ and $\mathbf{0}$ are $(K+1)\times(K+1)$ matrices;
$\mathbf{A}_i = [a^i_{k,h}]$; $a^i_{k,h} = P\{Y_{n+1} = h | X_n = i, Y_n = k\}$;
$\alpha_i = (M-i)\lambda/\gamma_i$ for $i = 0, 1,\ ...\ M-1$, and $\beta_i = i\mu/\gamma_i$ for $i = 1,\ ...\ M$.
The main diagonal is zero because sources can only move between the On and Off states.
In this particular case the Block Reduction algorithm used is the following:

$$\underline{\pi}_i = (\pi_{i,0}, \pi_{i,1}, ..., \pi_{i,K})$$

$$\underline{\pi}_i = \underline{\pi}_{i+1}\mathbf{U}_i \quad i = 0, ...\ M-1$$

where

$$\begin{array}{ll} \mathbf{U}_0 = -\beta_1\mathbf{A}_1 & i = 0 \\ \mathbf{U}_i = -\beta_{i+1}\mathbf{A}_{i+1}(\mathbf{I} - \alpha_{i-1}\mathbf{U}_{i-1}\mathbf{A}_{i-1})^{-1} & i = 1,\ ...\ M-1 \end{array} \tag{11}$$

and

$$\underline{\pi}_M(\mathbf{I} - \alpha_{M-1}\mathbf{U}_{M-1}\mathbf{A}_{M-1}) = \mathbf{0};$$

$\underline{\pi}_i$ are normalized during the iteration.
The complexity is $O((K+1)M^3)$ and we solved the system with the Open Source application Octave [7] on standard PC hardware for any value of M and K of practical interest.

4 System simulation

As a numerical example we consider IP Telephony. We consider standard applications like NetMeeting [8], Open H323 [9], or any other application using either H.323 [10] or

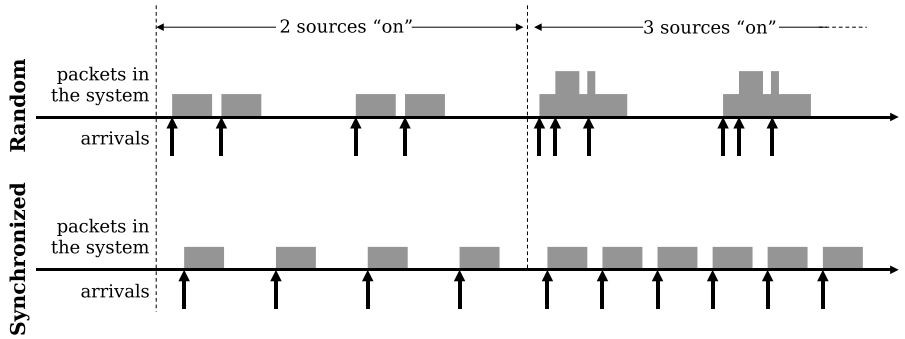

Figure 3: Comparison between actual synchronized and random arrivals

SIP [11] standards for signaling. All applications use RTP [12, 13] upon UDP/IP as transport protocol.

We implemented an ad-hoc simulator [15] because the system is simple enough to discourage the use of a general purpose network simulator as *ns-2* [16], and, most of all, because we want to control all details of the implementation and its efficiency, so as to be able to estimate accurately loss probabilities as low as 10^{-7}.

Among the different encoders we consider G.729 [14] encoder with VAD[1]. Voice packets contain 40 bytes of data, that, with RTP/UDP/IP headers make 80 bytes packets. In case of header compression, the total packet dimension becomes 44 bytes. We assume that also silence packets contain 40 bytes of data to preserve deterministic service.

Sources are homogeneous corresponding to the case of Sect. 3.2; $\delta_{On} = \delta_1 = (1/20)\,\text{ms}^{-1}$, $\delta_{Off} = \delta_0 = (1/160)\,\text{ms}^{-1}$. The mean On and Off periods are equal, and we consider two different situations: intra-word silence detection, with $T_o = 1/\mu = 1/\lambda = 0.5\,\text{s}$, and macro silence detection, with $T_o = 1/\mu = 1/\lambda = 5\,\text{s}$.

Voice is a delay (and delay jitter) sensitive application: a single interface must not introduce excessive delay, thus limiting the buffer requirements. We consider two cases of maximum allowed delay: $d_{\max} = 5\,\text{ms}$ and $d_{\max} = 10\,\text{ms}$. This last constraint defines the dimension of the buffer dedicated to VoIP applications: $B = \left\lfloor \frac{d_{\max} C}{S_P} \right\rfloor$ where S_P is the packet size in bits. For instance, dedicating 1 Mbit/s to VoIP services with $d_{\max} = 10\,\text{ms}$ yields a 15 packets buffer.

Fig. 3 shows the two possibilities we're faced with, when multiplexing sources with deterministic arrivals. The lower part of the figure refer to the case when sources can be synchronized, as, for instance, when all sources belong to a same packetizing gateway. This case maps exactly to the MMDP/D/1/K queuing system. The upper part refers to a case where sources cannot be synchronized and within a single phase packets are not equally spaced and can overlap, leading to short term congestion and queue buildup. Our simulator handle both cases and in Sect. 5 we discuss the impact on performance.

[1] Any other standard, like GSM, G.723, G.711, etc. would only change the packet size or packet interarrival time.

5 Numerical Examples

We focus our attention on three different link capacities, assuming that all the capacity is reserved for VoIP services: 512 kbit/s, 2 Mbit/s and 10 Mbit/s. In the case of 512 Kbit/s we assume that header compression is present. For each capacity, we evaluate the performances with the two different buffer sizes and the two different mean on/off periods. All simulations are run till a 99% confidence level is reached on a $\pm 5\%$ interval of the point estimate. The load is varied changing the number of sources M.

Fig. 4 reports the model upper and lower bound for P_l and the simulation results assuming or not synchronization. Left hand plots refer to $T_o = 5\,\mathrm{s}$, right hand ones to $T_o = 0.5\,\mathrm{s}$. Different rows refer to different capacity Cs, and the buffer is for the case $d_{\max} = 10\,\mathrm{ms}$. Simulation results always fall between the model estimated upper and lower bounds, also when non-synchronized sources are simulated and the model is thus an approximation. As expected, the upper and lower bounds are tighter for long On-Off periods, since the renewal assumptions have a smaller relative impact on the performance. Also the short term congestion induced by non-synchronized sources is more evident if the On-Off periods are very short, and the relative simulation curve approaches the upper bound.

The role of short term congestion is greater reducing the buffer size. Fig. 5 reports the results for $d_{\max} = 5\,\mathrm{ms}$. Again left hand plots refer to $T_o = 5\,\mathrm{s}$, right hand ones to $T_o = 0.5\,\mathrm{s}$. We only report results for $C = 2\,\mathrm{Mbit/s}$ and $C = 512\,\mathrm{kbit/s}$ since those for $C = 10\,\mathrm{Mbit/s}$ are qualitatively equal to those with $C = 2\,\mathrm{Mbit/s}$ (a complete set of results can be found in [17]). As expected reducing the buffer size can dramatically change the quality of the approximation, but, most interestingly, it is only the absolute value of the buffer size and not the maximum delay introduced by the buffer or the average On-Off period that predominates, as it is clear comparing the four plots. If sources are synchronized, the model upper and lower bounds hold also for very low buffer sizes. We can conclude that the model fails to catch the system behavior only if sources are not synchronized (e.g., in backbone routers) and the buffer dedicated to VoIP services is extremely small.

Concluding it is interesting to compare the results yielded by our model with those obtained with simple On-Off approximations. Fig. 6 reports two possible comparison scenarios. In the left hand plot the silence descriptors are simply ignored. As expected, ignoring them provides an underestimation of P_l, which is quite large. A more interesting perspective is offered by the right hand side plot, where results are compared for an equal average offered load and not for an equal number of sources. In this case the simpler On-Off model overestimates P_l. For loads of practical interest (< 0.85) the gap can be larger than an order of magnitude. The reason is that for the same offered load On-Off sources are burstier than sources with a high and a low (but not zero) state.

6 Conclusions

This paper describes a novel analytical framework to evaluate the performance of a class of multimedia (namely those that can be described with a piecewise constant emission rate) services on a single network interface, with applications for VoIP and video services. The modeling technique is based on an MMDP/D/1/K queuing station we solve introducing renewal approximations. The solution technique we propose enables to obtain both upper and lower bound on performance.
The analytical results for realistic VoIP scenarios, based on two state sources, were validated against detailed event-driven simulations, showing the approach is correct and accurate. A comparison with simpler, On-Off models show that our approach can estimate the loss probability with greater accuracy. The numerical solution for three-state sources is sketched and we are currently deriving numerical results for video sources.

References

[1] S. Blake et al. *An Architecture for Differentiated Services*, RFC 2475, IETF, Dec. 1998.

[2] B. Davie et al., *An Expedited Forwarding PHB (Per-Hop Behavior)*, RFC3246, IETF, Mar. 2002.

[3] T. Yang, D.H.K. Tsang, "A Novel Approach to Estimating the Cell Loss probability in an ATM Multiplexer Loaded with Homogeneous On-Off Sources," *IEEE Trans. on Communications*, 43(1):117–126, Jan. 1995.

[4] Sang H. Kang, Yong Han Kim, Dan K. Sung, Bong D. Choi, "An Application of Markovian Arrival Process (MAP) to Modeling Superposed ATM Cell Streams," *IEEE Trans. on Communications*, 50(4):633–642, Apr. 2002.

[5] H. Saito, *Teletraffic Technologies in ATM Network*, Boston, MA, USA, Artech House, 1994.

[6] M. F. Neuts, *Structured Stochastic Matrices of M/G/1 Type and Their Applications*, M. Dekker, New York, 1989.

[7] The Octave Web Page. `http://www.octave.org/`

[8] Microsoft NetMeeting. `http://www.microsoft.com/windows/netmeeting/`

[9] The OpenH323 Project Homepage. `http://www.openh323.org/`

[10] ITU Standard H.323 Version 5, *Packet-based multimedia communications systems*, ITU, Geneva, CH, July 2003.

[11] J. Rosenberg et al., *SIP: Session Initiation Protocol*," RFC 3261, IETF, June 2002,

[12] H. Schulzrinne et al., *RTP: A Transport Protocol for Real-Time Applications*, RFC 3550, IETF, July 2003.

[13] H. Schulzrinne et al., "RTP Profile for Audio and Video Conferences with Minimal Control," RFC 3551, IETF, July 2003.

[14] ITU Standard G.729, *Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear-prediction (CS-ACELP)*, and subsequent modifications, ITU, Geneva, CH, March 2003 – Oct. 2002.

[15] D. Carra, MMDP Multimedia simulator: The home page. `http://netmob.unitn.it/mmdpms/`

[16] ns, network simulator (ver.2), Lawrence Berkeley Laboratory, `http://www-mash.cs.berkeley.edu/ns`

[17] D. Carra, P. Laface, R. Lo Cigno, "A Performance Model for Multimedia Services Provisioning on Network Interfaces (Extended Version)," Tecnical Report DIT-04, Università di Trento. `http://dit.unitn.it/locigno/preprints/CaLaLo04-92.pdf`.

A UU and LL Buffer transition distributions

We derive here the distributions describing the transitions between buffer levels for the UU and LL approximations.

A.1 Upper-Upper bound

The number of packets going in the buffer during the interval $U_{\underline{f}}$ is $\lceil V_{\underline{f}} U_{\underline{f}} \rceil$ and the number of packet served is $\lfloor C U_{\underline{f}} \rfloor$. We have three different cases depending on the relative value of $V_{\underline{f}}$ and C.
If $V_{\underline{f}} = C$ (In this case the event $\lceil V_{\underline{f}} U_{\underline{f}} \rceil - \lfloor C U_{\underline{f}} \rfloor > 0$ is always true)

$$a^{\underline{f}}_{k,h} = \begin{cases} P\{\lceil V_{\underline{f}} U_{\underline{f}} \rceil - \lfloor C U_{\underline{f}} \rfloor > 0\} = 1 & h = k+1 \\ 1 & h = k = K \\ 0 & otherwise \end{cases}$$

If $V_{\underline{f}} > C$

$$a^{\underline{f}}_{k,h} = \begin{cases} P\{\lceil (V_{\underline{f}} - C) U_{\underline{f}} \rceil = h-k\} = \rho_{\underline{f}}^{h-k-1}(1-\rho_{\underline{f}}) & k < h < K \\ P\{\lceil (V_{\underline{f}} - C) U_{\underline{f}} \rceil \geq K-k\} = \rho_{\underline{f}}^{K-k-1} & k < h = K \\ 1 & h = k = K \\ 0 & otherwise \end{cases}$$

where $\rho_{\underline{f}} = exp(-\gamma_{\underline{f}}/(V_{\underline{f}} - C))$.

If $V_{\underline{f}} < C$

$$a^{\underline{f}}_{k,h} = \begin{cases} P\{\lfloor (C - V_{\underline{f}})U_{\underline{f}} \rfloor = k - h\} = & \rho_{\underline{f}}^{k-h}(1 - \rho_{\underline{f}}) & k > h \geq 1 \\ P\{\lfloor (C - V_{\underline{f}})U_{\underline{f}} \rfloor \geq k\} = & \rho_{\underline{f}}^{k} & k > h = 0 \\ P\{(C - V_{\underline{f}})U_{\underline{f}} < 1\} = & 1 - \rho_{\underline{f}} & h = k + 1 \;\; k = h = K \\ P\{(C - V_{\underline{f}})U_{\underline{f}} > 1\} = & \rho_{\underline{f}} & h = k = 0 \\ 0 & & otherwise \end{cases}$$

where $\rho_{\underline{f}} = exp(-\gamma_{\underline{f}}/(C - V_{\underline{f}}))$.
The random variable $N_{\underline{f},j} = \lceil V_{\underline{f}} U_{\underline{f}} \rceil$ has geometric distribution over non-negative integers ≥ 1 with parameter $\rho_{\underline{f}} = exp\{-\gamma_{\underline{f}}/V_{\underline{f}}\};\ V_{\underline{f}} > 0$, and

$$E[N_{\underline{f},j}] = \frac{1}{1 - \rho_{\underline{f}}}$$

Computing $E[R_{\underline{f},j}]$ we have: $P\{R_{\underline{f},j} = 0\} = 1$, if $V_{\underline{f}} \leq C$, and $R_{\underline{f},j} = max(0, \lceil (V_{\underline{f}} - C)U_{\underline{f}} \rceil - K + j)$, if $V_{\underline{f}} > C$, which yields

$$P\{R_{\underline{f},j} = k\} = P\{\lceil (V_{\underline{f}} - C)U_{\underline{f}} \rceil = K - j - k\} = \rho_{\underline{f}}^{K-j+k-1}(1 - \rho_{\underline{f}}), \quad k = 1, 2, ...$$

and finally

$$E[R_{\underline{f},j}] = \frac{\rho_{\underline{f}}^{K-j}}{1 - \rho_{\underline{f}}}, \qquad V_{\underline{f}} > C$$

where $\rho_{\underline{f}} = exp\{-\gamma_{\underline{f}}/(V_{\underline{f}} - C)\}$.

A.2 Lower-Lower bound

With reasoning similar to the UU approximation we have. If $V_{\underline{f}} = C$

$$a^{\underline{f}}_{k,h} = \begin{cases} 1 & k = h = 0 \; and \; h = k - 1 \\ 0 & otherwise \end{cases}$$

If $V_{\underline{f}} < C$

$$a^{\underline{f}}_{k,h} = \begin{cases} 1 & & h = k = 0 \\ P\{\lceil (C - V_{\underline{f}})U_{\underline{f}} \rceil = k - h\} = & \rho_{\underline{f}}^{k-h-1}(1 - \rho_{\underline{f}}) & k > h > 0 \\ P\{\lceil (C - V_{\underline{f}})U_{\underline{f}} \rceil > k - 1\} = & \rho_{\underline{f}}^{k-1} & k \geq h = 0 \end{cases}$$

If $V_{\underline{f}} > C$

$$a^{\underline{f}}_{k,h} = \begin{cases} P\{\lfloor (V_{\underline{f}} - C)U_{\underline{f}} \rfloor \geq K - h\} = & \rho_{\underline{f}}^{K-h} & k \leq h = K \\ P\{\lfloor (V_{\underline{f}} - C)U_{\underline{f}} \rfloor = h - k\} = & \rho_{\underline{f}}^{h-k}(1 - \rho_{\underline{f}}) & k \leq h < K \\ 0 & & h < k \end{cases}$$

where $\rho_{\underline{f}} = exp\{-\gamma_{\underline{f}}/(V_{\underline{f}} - C)\} = P\{(V_{\underline{f}} - C)U_{\underline{f}} > 1\}$.

$$E[N_{\underline{f},j}] = \frac{\rho_{\underline{f}}}{1-\rho_{\underline{f}}}, \quad \rho_{\underline{f}} = exp\{-\gamma_{\underline{f}}/V_{\underline{f}}\}, \quad V_{\underline{f}} > 0.$$

$$E[R_{\underline{f},j}] = \frac{\rho_{\underline{f}}^{K-j+1}}{1-\rho_{\underline{f}}}, \quad V_{\underline{f}} > C, \quad \rho_{\underline{f}} = exp\{-\gamma_{\underline{f}}/(V_{\underline{f}} - C)\}.$$

B Solution for three-state sources

When the sources can assume three states the modulating CTMC assumes the structure depicted in Fig. 2. The corresponding transition matrix **Q** has the following structure

$$\mathbf{Q} = \begin{bmatrix} \mathbf{A}_0 & \mathbf{B}_0 & & \dots & \\ \mathbf{C}_1 & \mathbf{A}_1 & \mathbf{B}_1 & \dots & \\ & \mathbf{C}_2 & \mathbf{A}_2 & \mathbf{B}_2 & \\ & & & \dots & \mathbf{B}_{M-1} \\ & & & \mathbf{C}_M & \mathbf{A}_M \end{bmatrix}$$

The matrix blocks have the following form:

$$\mathbf{A}_i = \begin{bmatrix} \mathbf{0} & a^i_{0,1}\mathbf{A}_{i,0} & & \dots & \\ a^i_{1,0}\mathbf{A}_{i,1} & \mathbf{0} & a^i_{1,2}\mathbf{A}_{i,1} & \dots & \\ & a^i_{2,1}\mathbf{A}_{i,2} & \mathbf{0} & a^i_{2,3}\mathbf{A}_{i,2} & \\ & & & \dots & a^i_{D_i-1,D_i}\mathbf{A}_{i,D_i-1} \\ & & & a^i_{D_i,D_i-1}\mathbf{A}_{i,D_i} & \mathbf{0} \end{bmatrix}$$

with

$$a^i_{j,l} = \begin{cases} (M-j)\lambda_2/\gamma_{i,j} & l = j+1 \;\; j = 0,1,\dots D_i - 1 \\ j\mu_2/\gamma_{i,j} & l = j-1 \;\; j = 1,\dots D_i \end{cases}$$

$$\mathbf{B}_i = \begin{bmatrix} b^i_{0,0}\mathbf{A}_{i,0} & \mathbf{0} & & \dots & \\ b^i_{1,0}\mathbf{A}_{i,1} & b^i_{1,1}\mathbf{A}_{i,1} & \mathbf{0} & \dots & \\ & b^i_{2,1}\mathbf{A}_{i,2} & b^i_{2,2}\mathbf{A}_{i,2} & \mathbf{0} & \\ & & & \dots & b^i_{D_{i-1},D_{i+1}}\mathbf{A}_{i,D_i-1} \\ & & & \mathbf{0} & b^i_{D_i,D_{i+1}}\mathbf{A}_{i,D_i} \end{bmatrix}$$

with

$$b^i_{j,l} = \begin{cases} (M-i)\lambda_1/\gamma_{i,j} & l = j \;\; j = 0,1,\dots D_i \;\; l \le D_{i-1} \\ j\alpha/\gamma_{i,j} & l = j-1 \;\; j = 1,\dots D_i \;\; l \le D_{i+1} \end{cases}$$

$$\mathbf{C}_i = \begin{bmatrix} c^i_{0,0}\mathbf{A}_{i,0} & c^i_{0,1}\mathbf{A}_{i,0} & & \dots & \\ \mathbf{0} & c^i_{1,1}\mathbf{A}_{i,1} & c^i_{1,2}\mathbf{A}_{i,1} & \dots & \\ & \mathbf{0} & c^i_{2,2}\mathbf{A}_{i,2} & c^i_{2,3}\mathbf{A}_{i,2} & \\ & & & \dots & c^i_{D_i-1,D_{i-1}}\mathbf{A}_{i,D_i-1} \\ & & & c^i_{D_i,D_{i-1}-1}\mathbf{A}_{i,D_i} & c^i_{D_i,D_{i-1}}\mathbf{A}_{i,D_i} \end{bmatrix}$$

with

$$c^i_{j,l} = \begin{cases} i\mu_1/\gamma_{i,j} & l = j \;\; j = 0, 1, \ldots D_i \;\; l \leq D_{i-1} \\ i\beta/\gamma_{i,j} & l = j+1 \;\; j = 0, \ldots D_i - 1 \;\; l \leq D_{i-1} \end{cases}$$

The elements $a^i_{j,l}$, $b^i_{j,l}$ and $c^i_{j,l}$ are the transition probabilities from state (i, j) to state (i, l) of the modulating embedded Markov chain. The parameter $\gamma_{i,j}$ is the time spent by the modulating process in the state (i, j). The system can be solved with techniques similar to those used for two-state sources.

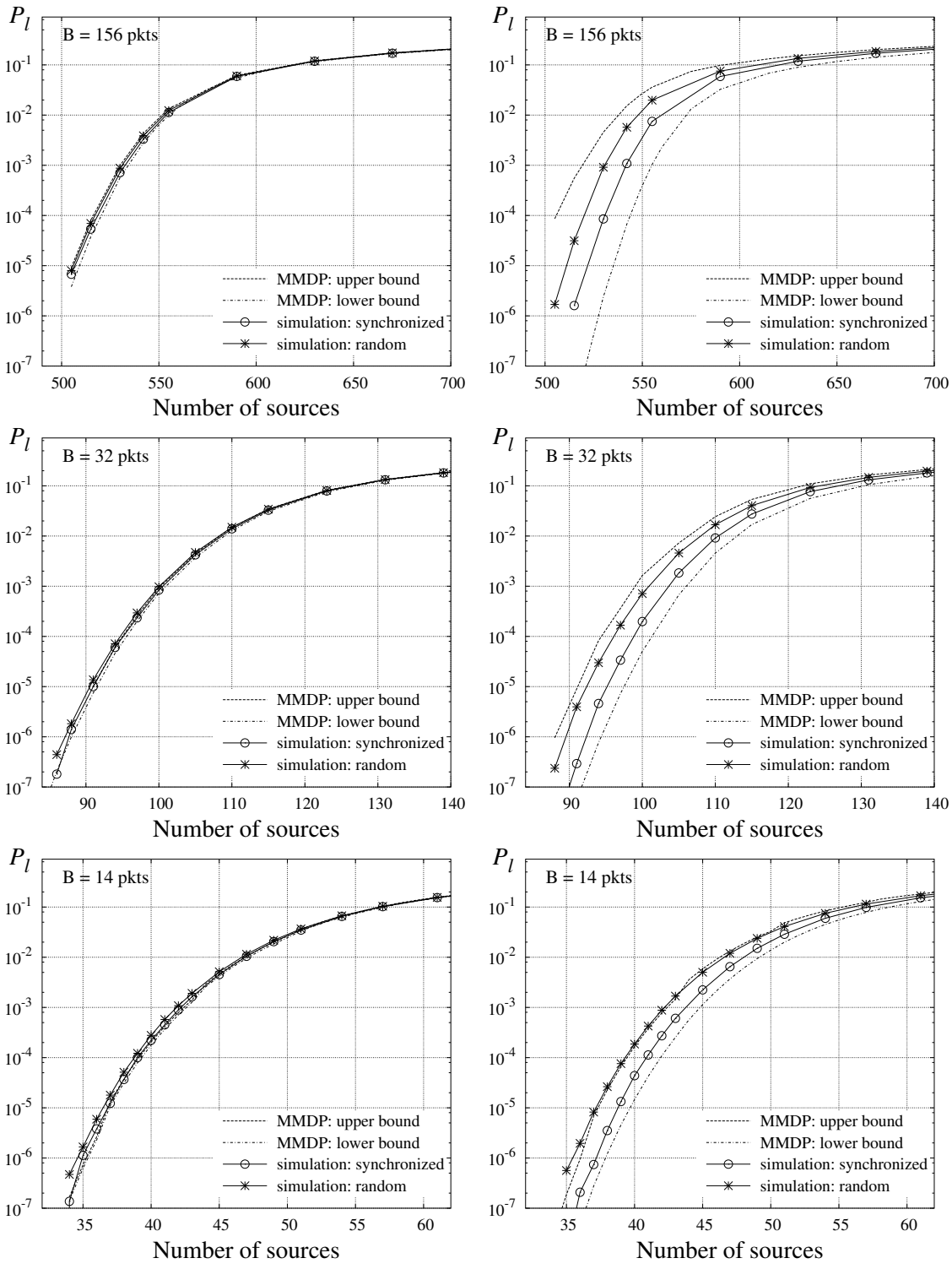

Figure 4: Packet loss: Model vs. simulations, $d_{\max} = 10\,\text{ms}$; top row $C = 10\,\text{Mbit/s}$, middle row $C = 2\,\text{Mbit/s}$, bottom row $C = 512\,\text{kbit/s}$; left column $T_o = 5\,\text{s}$, right column $T_o = 0.5\,\text{s}$,

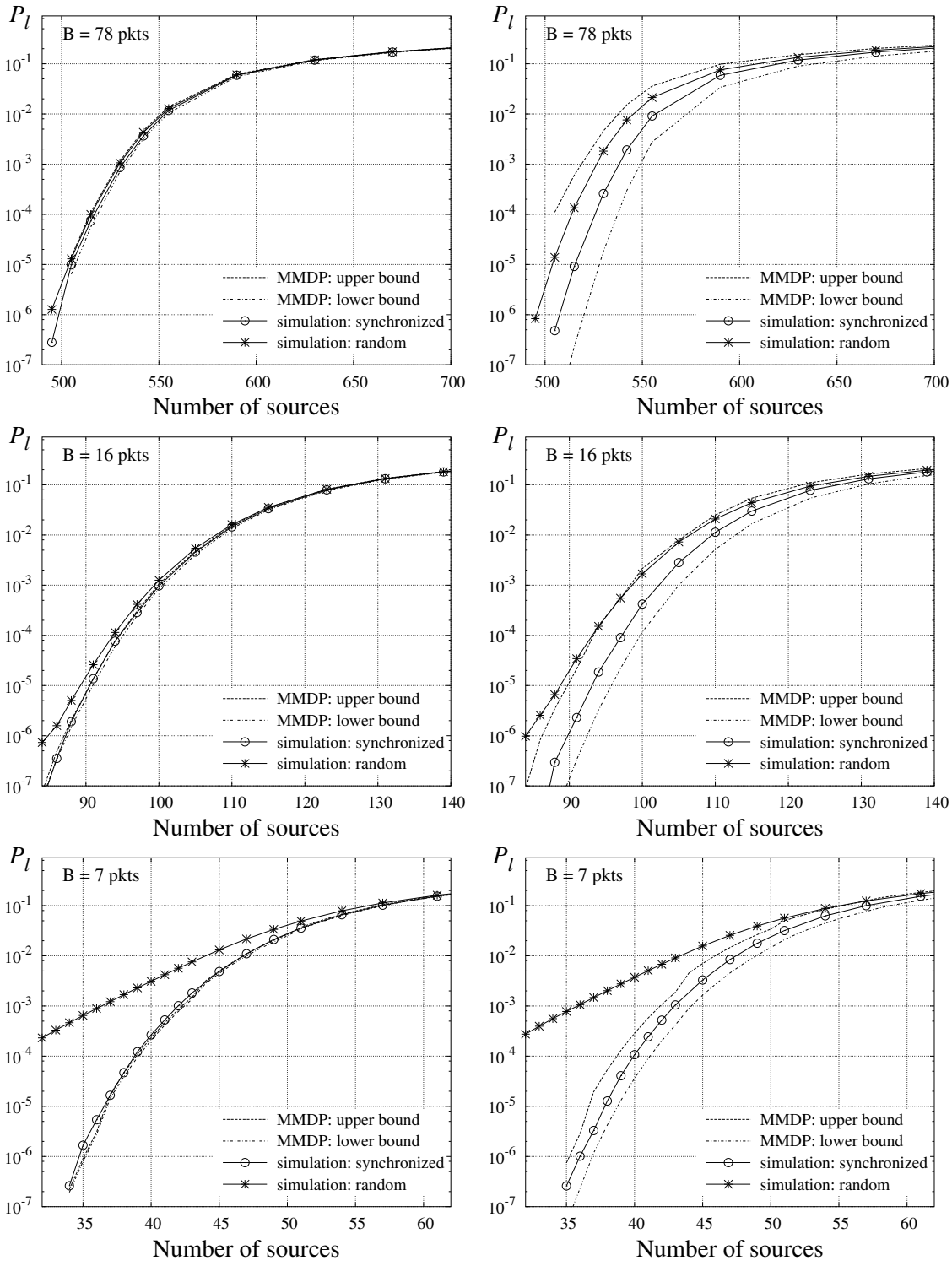

Figure 5: Packet loss: Model vs. simulations, $d_{\max} = 5\,\text{ms}$; top row $C = 10\,\text{Mbit/s}$, middle row $C = 2\,\text{Mbit/s}$, bottom row $C = 512\,\text{kbit/s}$; left column On and Off average time 5 s, right column On and Off average time 0.5 s,

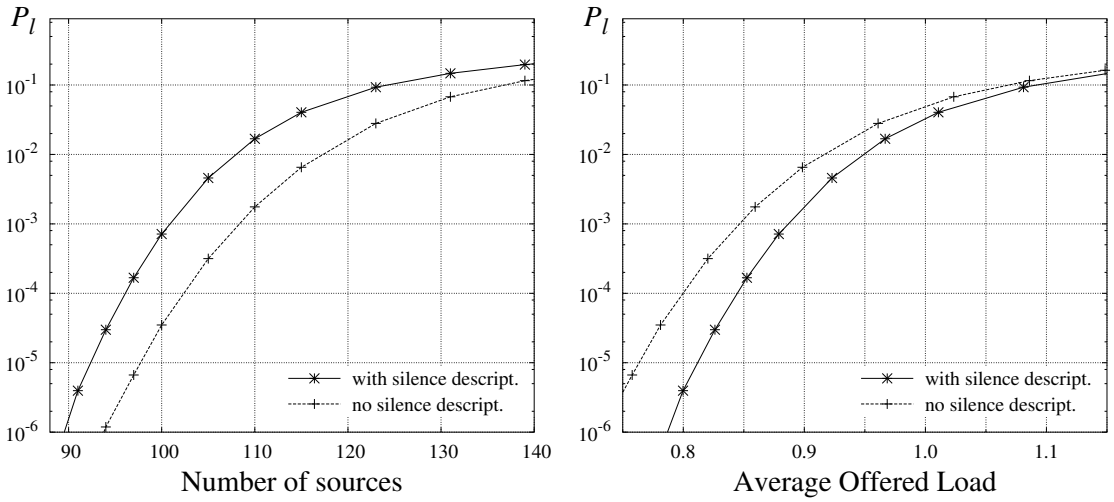

Figure 6: Comparison between our model and simpler On-Off models; C=2 Mb/s, t_{on}=500 ms, buff=32 pkts; ignoring the silence descriptors (left plot) or equalizing the offered load (right plot)