Proceedings of the 2005 Winter Simulation Conference
M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, eds.

APPLICATION OF RARE EVENT TECHNIQUES TO TRACE DRIVEN SIMULATION

Poul E. Heegaard

Department of Telematics,
Norwegian University of Science and Technology
O.S. Bragstads plass 2B
N-7491 Trondheim, NORWAY

Bjarne E. Helvik

Centre for Quantifiable Quality of
Service in Communication Systems, NTNU
O.S. Bragstads plass 2E
N-7491 Trondheim, NORWAY

Ragnar Ø. Andreassen

Telenor R&D,
Snarøyveien 30,
N-1331 Fornebu, NORWAY

ABSTRACT

This paper describes a trace driven, fast simulation approach applicable to deal with the performance evaluation of a multiplex of heterogeneous traffic streams with variable bit rate and long lived serial correlation offered routers in the Internet. A challenge with simulations of the Internet is the huge number of events that are needed for each event of interest, e.g. the loss or excessive delay of a packet. The simulation efficiency of the trace driven approach in this paper is improved by use of importance sampling to provoke constellation of traces where the loss and long delays are more likely. The approach is successfully applied to speed-up the simulation of multiplexing of heterogeneous MPEG encoded video streams.

1 INTRODUCTION

The fraction of traffic in the Internet from real-time applications is increasing. Some of the real-time applications will be provided with service guarantees like maximum delay or loss ratio. Hence, it is of great importance to be able to evaluate the traffic handling with respect to these guarantees. In general, evaluation of performance in IP based networks by means of simulation tends to be rather demanding with respect to computational effort when much traffic and long observation periods are required. This and other problems related to simulating Internet performance are discussed in (Floyd and Paxson 2001). Traffic from real-time applications has characteristics which are difficult to describe by generation models. Furhermore, the traffic is usually not loss aware. This actualizes the use of trace driven simulation, see for instance (Jain 1991), as a means for evaluation.

The objective in this paper is to provide an efficient method to determine the loss when a multiplex of a large number of traced traffic streams, e.g. MPEG encoded video streams, are offered to a single buffer or a network. The method aims at determining the low loss rate required by high QoS provision. For this, a trace driven approach is formulated, where traces are typically heterogeneous sequences of video frames with long-lived serial-correlations. For efficiency, trace driven simulation is combined with importance sampling, see for instance (Heidelberger 1995) for an introduction to the topic. The basic idea is to position traces relative to each other with respect to arrivals times, so temporal “overloads” occur and cause overloads. This corresponds to a shift of measure relative to the natural uniformly distributed relative positioning.

Applying importance sampling to speed up trace driven simulations of heterogeneous traffic streams offered to a common finite buffer, was first addressed by the authors in (Andreassen, Heegaard, and Helvik 1996). Later, the same problem has been addressed by (Chang, Chiu, and Song 2001) and (Paschalidis and Vassilaras 2004). There is however a significant difference in how the change of measure is obtained. The authors have adopted an approach where the entire input space that may lead to buffer overflows is systematically sampled, while in the referred papers, large deviations theory is used. The latter approach, if successful, may give results with a smaller computational effort. The method presented in this paper is expected to be more robust, since in this problem domain, there are likely to be several constellations of offered traffic which will contribute similarly to buffer overflows; a situation which is poorly handled by large deviations based importance sampling (Glasserman and Wang 1997).

The system model, described in Section 2, is not limited to video streams alone, but will generally be applicable to traffic sources that can be described by a sequence of packets or bits that may be chopped into frames with uniform traffic. In Section 3, an importance sampling based approach is proposed for reducing the simulation time by provoking the occurrence of the rare constellations that leads to loss or significant delay of packets in the trace. Generalizations of the proposed approach is described in Section 4, followed by experiments on a multiplex of heterogeneous MPEG encoded video streams in Section 5. Some closing remarks are given in Section 6.

2 SYSTEM MODEL

The objective is to estimate the packet loss ratio for loss-sensitive traffic, e.g. real-time video, in the Internet. The traffic source is modelled by a sequence of *frames*, $\mathbf{X} = X_1, X_2, \cdots$, where each frame represents a number of information *packets* sent in a frame period. A *packet* can for instance be bits, bytes, MAC frames, IP packets, or segments. The sequence of frames, denoted a *trace*, might have a strong, long-lived, serial correlation, typically observed in e.g. MPEG encoded video streams due to the inherent correlation in the video content and the *Group Of Picture (GOP)* structure.

The system is a network of nodes offered a large number of traces, $\mathbf{X}_m$, multiplexed with each other. All frames have the same duration, i.e., the frame rate are the same in all traces, and all traces are assumed to have the same length of $\mathbf{N}$ frames.

To simplify the notation and the description of the importance sampling (IS) approach in the next section, the multiplexing of traffic streams over an outgoing interface i in the node k is considered. However, the IS approach is generally applicable to networks of nodes.

The node k is illustrated in Figure 1. The index k is ignored for notation simplicity in the following. The traces in the set Ω_n are new traces offered to the node, while the set Ω_o contains the multiplex of traces that are routed from other nodes. The server process has capacity S [packets/sec], while $S_{t,i} \leq S$ are the current available capacity on outgoing interface i at time t.

It is assumed that each trace follows a deterministic intra-frame arrival process, with a large number of packets arriving each frame interval. The number of packets served during a frame period (excl. buffer-length) is, $d_t = S_{t,i} - \sum_{j\in\Omega_o} X_{t,j}$ which is the available server capacity for new trace traffic at time t. In the following this is deterministic $d_t = d$. Hence, the capacity d can be considered to be either the reserved capacity for trace modelled traffic sources (e.g. MPEG video streams), or that the node is only offered trace modelled traffic sources.

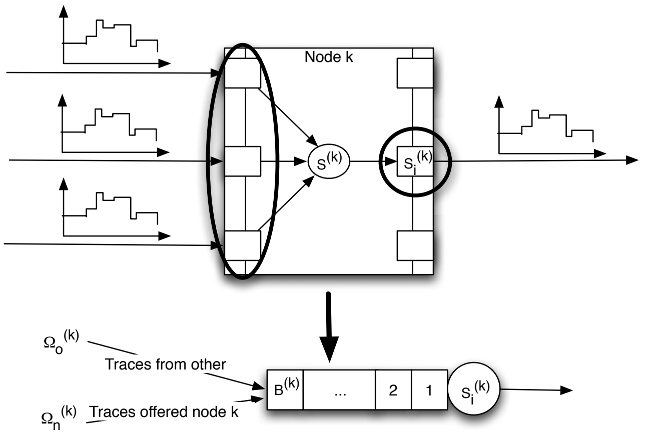

Figure 1: System Model of Node k

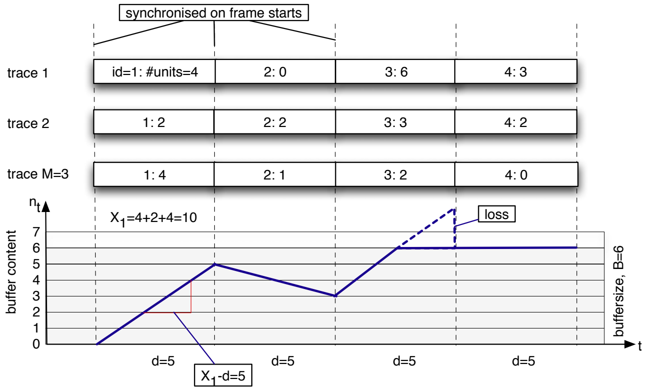

Figure 2: Traces Synchronized At Frame Interval Positions

Assuming packet arrival before service, and letting n_t be the number of packets at the selected interface in the node at time t, then the following recursion relation applies (the index i with reference to interface i is suppressed in the following)

$$n_{t+1} = \max(0, \min(n_t + \mathbf{X}_{\mathbf{k}_t} \cdot \mathbf{1} - d, B)) \qquad (1)$$

where B is the buffer capacity in packets and $\mathbf{k}_t$ is the *frame constellation* at time t. In Section 3.2.2 the frame constellation concept is introduced. In (1) the frames in the traces are synchronised at t. Figure 2 shows an example with 3 traces where the number of packets in the system varies over time and is truncated at the buffer size B.

A generalization to traces not synchronised at frame start can be done as follows. Let the system response still be determined by the system state at frame arrival instants, but now frame arrivals are spread throughout a frame interval according to the frame phases of the sources. Hence, the following generalization of the recursion in (1) applies

$$n_{j+1} = \max(0, \min(n_j + \frac{d_j}{d}\mathbf{X}_{\mathbf{k}_j} \cdot \mathbf{1} - d_j, B)) \qquad (2)$$

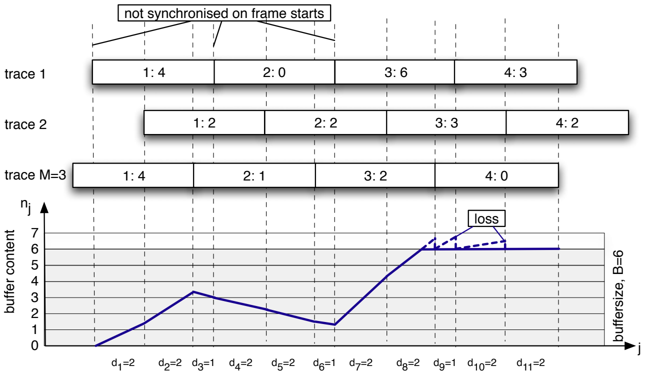

Figure 3: Traces Not Synchronized At Frame Interval Positions

where d_j is the number of packets served during the j'th fixed rate interval. Figure 3 shows an example with 3 traces where the number of packets in the system varies over time and is truncated at the buffer size B. Observe that the net increase and reduction of the number n_j might change for each frame interval of each individual trace, in contrast to the synchronized case where this changes only on synchronized frame interval common for all traces.

The simulations assumes source traces that are offered a node for a random period and starting from a random position. The traces are assumed to be cyclic, i.e. if you reach the end you will start all over again from position 1. The trace driven approach is well suited for modelling sources that have strong and long lived correlation where other processes like various Markov modulated processes fails.

3 IMPORTANCE SAMPLING IN TRACE DRIVEN SIMULATION

This section presents the general idea of the application of importance sampling to provoke packet loss in trace driven simulations. The basics are given in Section 3.2 followed by a brief description of importance sampling in Section 3.3 and details on how to change the underlying sampling distribution in Section 3.4.

3.1 General Idea

Trace simulation of the packet loss ratio of a multiplex of traces, see Figure 4, becomes very inefficient when the loss ratio is small or a large number of multiplexed traces and background traffic exists. The idea in this paper is to use importance sampling to increase the simulation efficiency by increasing the number of packet losses through provoking multiplex-patterns where overloads are more likely.

A straight forward heuristics is to sample a starting frame position for each trace in Ω_n according to the relative *load* (number of packets) in each frame of the sequence. In addition, a *load selection* of the starting frame positions will force the total number of packets generated at the starting frame position to exceed the server capacity, if feasible.

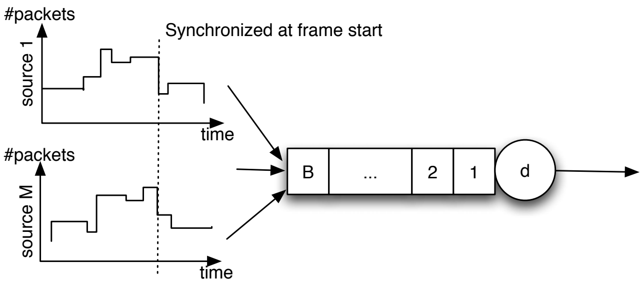

Figure 4: Multiplexing Of Synchronized Traces

3.2 Basics and Notation

Before the importance sampling strategy is introduced, the basic trace simulation of multiplexing of frame sequences is described, see Figure 4. Necessary notation and basic concepts are also included. Note that for notation simplicity, only the special case of homogeneous and synchronised trace sequences are described om this section; removal of these constraints are described in Section 4. This means that the frames arrive simultaneously, and that they all originate from the same sequence, although not in the same frame position, of course.

3.2.1 General Notations

The simulation strategy is presented assuming that all traces are equal and of length N, i.e. the number of frames in a sequence. The number of sources is M, i.e. traces offered the node.

In the description the modulo addition is applied, defined as $a \oplus b \equiv (a+b-1)mod(N)+1$. In addition, an identity vector of size N, equal to $\mathbf{1} = \{1, \cdots 1\}$, and an indicator function $I(x) = \begin{cases} 1 & x > 0 \\ 0 & x \leq 0 \end{cases}$ are needed.

3.2.2 Trace Driven Simulation

The simulation of multiplexed traces is done by sampling a starting position in the trace sequence for each source i, in sequence from 1 to N. In total there are N^M possible options. This ordered set of starting positions, and generally a synchronised or aligned frame-positions, are denoted a *frame constellation*, $\mathbf{k} = \{k(1), \cdots, k(M)\}$. The frame constellation at *frame position identifier* j is $\mathbf{k}_j = \mathbf{k} \oplus (j \cdot \mathbf{1})$, see Figure 5. The number of packets in the M frames given by constellation $\mathbf{k}$ is $\mathbf{X_k} = \{X_{k(1)}, \cdots, X_{k(M)}\}$ where $X_{k(i)}$ is the number of packets in sources i of constellation $k(i)$. The maximum number of packets in a frame is expressed as $X_{\max} = \max_{k \in 1,N} X_k$. To select the high load frames in

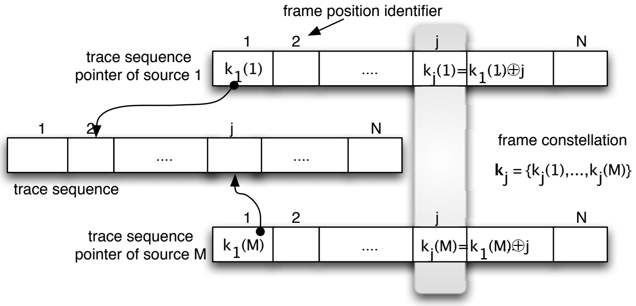

Figure 5: Basic Concepts Related to Trace Driven Simulation

a trace it is defined a set of frame positions having more than x packets, $J(x) = \{i \mid X_i > x\}$.

The system response of A, $Y(A)$, is the number of packet losses which are deterministically given by tracing through the N frames in the multiplex of frame sequences where each source starts at the position given by $\mathbf{k}$.

3.2.3 Concept of Alignments

The sampling of a starting frame constellation can be viewed as a sampling of a frame sequence alignment, because the relative position between the M sources is constant throughout the entire sequence. This means, that the same alignment can be sampled as a result of sampling any of the frame constellations this alignment consists of. However, the system response will, due to an initial transient caused by buffers, for high loads be slightly dependent on the starting frame constellation.

An alignment constituted by the specific frame constellation is then

$$A_l^{(\mathbf{k})} = \{\mathbf{k}, \mathbf{k} \oplus \mathbf{1}, \cdots, \mathbf{k} \oplus ((N-1) \cdot \mathbf{1})\} \tag{3}$$

and $A_l = A_l^{(\mathbf{k})}$, for all $i = 1$ to N, is the alignment number l with either (non-specified) frame constellation in A_l as the starting constellation. Observe that N different rotations of the alignment depicted in Figure 5 will result in the same alignment. Whenever a reference to the starting frame constellation is needed, the index refers to the first source, called *frame position identifier* as indicated in Figure 5.

Note that there exists several *permutations* of the order of the M sources (a source corresponds to a trace vector of pointers in Figure 5), and that each permutation will give identical response. It is, however, essential that each permutation constitutes a unique alignment. This assumption is necessary to make a simple sampling algorithm, see Section 3.4.

3.2.4 Alignment Probabilities

The probability of sampling the alignment A_l with an offset position j relative to the *frame position identifier* is denoted $P(l, j)$. Hence, the probability of an alignment A_l is $P(l) = \sum_{j=1}^{N} P(l, j)$. The original sampling distribution is uniform, i.e.

$$P(l) = \sum_{j=1}^{N} P(l, j) = \sum_{j=1}^{N} 1/N^M = 1/N^{M-1}. \tag{4}$$

System response is obtained under the assumption of a deterministic intra-frame packet arrival process, and is determined by multiplexer states at frame arrival instants, see (1). Hence, the system response is ($n_0 = 0$)

$$Y(l) = \sum_{t=1}^{N} \max(0, n_{t-1} + \mathbf{X}_{\mathbf{k}_t} \cdot \mathbf{1} - d - B) \tag{5}$$

which gives expected system response $E(Y) = \sum_{\forall l} Y(l)P(l)$. Hence, an unbiased estimator of $E(Y)$ from R direct simulation experiments is

$$\bar{Y} = \frac{1}{R} \sum_{r=1}^{R} Y(l_r) \tag{6}$$

where the alignment A_{l_r} is sampled according to the uniform distribution of (4).

3.3 Importance Sampling Fundamentals

Importance sampling has been used with success to yield speed-up in rare event simulation, see (Heidelberger 1995) for an excellent overview. Simulation of packet losses will require extremely long simulation periods to obtain stable estimates.

The theoretical fundamentals of importance sampling can shortly be described by the following. Consider Y as an observation of the quantity of interest to be a function $g(X)$ where X is sampled from $f(x)$. Assume that nonzero values of Y are rarely observed in a direct simulation. The basic idea is simply to change the underlying sampling distribution to $f^*(x)$ to make Y more likely to occur, in which the following relation holds

$$E_f(Y) = E_{f*}(Y \cdot \frac{f(X)}{f^*(X)}) = E_{f^*}(Y \cdot \Lambda(X)) \tag{7}$$

where $\Lambda(X)$ is the likelihood ratio between $f(X)$ and $f^*(X)$. Thus, the property of interest, Y, can be estimated by taking R samples from $f^*(x)$, accumulate $\Lambda(X)$, and use the following unbiased estimator

$$\bar{Y}^* = \frac{1}{R}\sum_{r=1}^{R} Y_r \cdot \Lambda(X_r). \tag{8}$$

The main challenge is to choose a new distribution, f^*, that minimizes the variance to the estimator, $\bar{Y}^*$. If an unsuited distribution is used, it is observed that simulation is inefficient and is producing inaccurate results, see e.g. (Devetsikiotis and Townsend 1993, Heegaard and Helvik 1999).

In the trace simulations in this paper, the X is an alignment A_l of frame sequences originally sampled from a uniform distribution, $f(x)$. The system response $Y = Y(l)$ is the number of lost packets, see (5). The following section will discuss heuristics specific for homogeneous frame sequences which determine the new sample distribution $f^*(x)$.

3.4 Changing the Sampling Distribution

3.4.1 Heuristics

Using a direct trace driven simulation approach for traffic sources modelled by sequence of frames, the alignments are sampled according to a uniform distribution. When a large number of frames in the sequence have few packets, and/or the overall mean load are low, this will result in an enormous number of alignments with system response $Y_l = 0$, i.e. the observation $Y_l > 0$ is a rare event. Hence, very long simulation experiments are required in order to estimate the packet loss ratio or long packet delays.

Importance sampling is introduced to this trace driven approach with the goal to reduce the rarity problem by increasing the frequency of alignments with $Y_l > 0$. This objective is achieved by changing the $f^*(X)$ by the following heuristics.

1. *Load distribution*: Sample the starting (frame) position of each of the multiplexed sources proportionally to its load, instead of the uniform distribution, see Figure 6. This will provoke the heavy load frames to coincide in an alignment.
2. *Load selection*: The conditional starting positions of the sources are restricted to those which makes a temporary overload (and hence packet loss) feasible.

3.4.2 Alignment Probabilities

Because an alignment is selected through sampling of one of its frame constellations, the alignment probability $P(l)$ is the sum of frame constellation probabilities, $P(l, j)$. The frame space is reduced for every source, dependent on the accumulated load of the frames sampled by the previous sources, and the maximum load which is possible to obtain by the remaining sources. Let $J(x) = \{i \mid X_i > x\}$ be the set of frame positions having more than x packets. Then, according to the heuristics from previous section, the change of frame constellation probabilities to $P^*(l, j)$ can be expressed as follows.

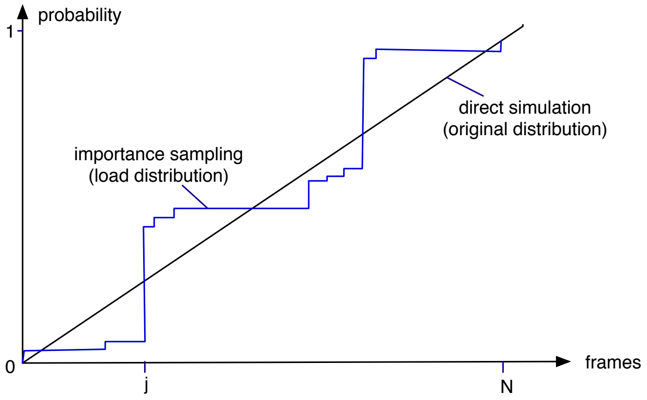

Figure 6: Cumulative Load Distribution

In sequence $i = 1$ to M:

$$x_i = d - \sum_{m=1}^{i-1} X_{k_j(m)} - (M - i) \cdot X_{\max} \tag{9}$$

$$P_i^*(l, j) = (X_{k_j(i)} \cdot I(x_i)) / \sum_{m \in J(x_i)} X_{k_j(m)} \tag{10}$$

and finally

$$P^*(l, j) = \prod_{i=1}^{M} P_i^*(l, j) \tag{11}$$

where

$$P^*(l) = \sum_{j=1}^{N} P^*(l, j). \tag{12}$$

This operation is $O(N)$-complex and is the most critical operation in the algorithm with respect to computer efficiency. However, if the sequence is sorted by frame size in decreasing order, and the frame references to the original unsorted sequence are known, the evaluation can be reduced to a $O(1)$-complex operation, see (Andreassen 1997).

Finally, the likelihood ratio is the ratio between the alignment probabilities in (4) and (12)

$$\Lambda(l) = \frac{P(l)}{P^*(l)} = \frac{1/N^{M-1}}{\sum_{j=1}^{N} P^*(l,j)}. \tag{13}$$

4 GENERALIZATIONS

This section describes some generalizations of the trace driven importance sampling strategy presented in the previous section, which makes it applicable for evaluation of a wider range of settings. In addition, the requirement of identical frame rates and trace lengths may obviously be somewhat relaxed. Frame rates, which are integer fractions of the highest rate, may be chopped into a series of shorter frames. Traces with lengths that are integer fractions of the longest, may be repeated, or more general, the least common multiple may be used at the expense of an increased computational effort.

4.1 Heterogeneous Traces

Typically, a node is not offered traces with the same content, i.e. sequence of frames. The assumption of homogeneous traces was made in the previous section to simplify the description and notation. It can easily be generalized to heterogeneous sources. Let the number of packets at frame position i, X_i, be generalized to $X_i^{(f)}$ where the new index (f) refers to the trace type. Substituting this into (9) and (10) gives

$$x_i = d - \sum_{m=1}^{i-1} X_{k_j(m)}^{(f_m)} - \sum_{m=i}^{M} X_{\max}^{(f_m)} \tag{14}$$

$$P_i^*(l,j) = (X_{k_j(i)}^{(f_i)} \cdot I(x_i)) / \sum_{m \in J(x_i)} X_{k_j(m)}^{(f_m)}. \tag{15}$$

The ordering of traces in sequence $m = 1, \cdots, M$ is given by $\mathbf{f} = \{f_1, \cdots, f_M\}$. In order to avoid biased frame constellation sampling, the ordering in $\mathbf{f}$ must be random an rearranged for each frame period. In the homogeneous case, this sequence is fixed, i.e. the frame position of source 1 is always sampled first, then 2, and finally source M.

4.2 Non-Synchronized Sources

In establishing (1), the assumption was made that frames from different sources arrive simultaneously at the multiplexer. When this assumption is removed as in Figure 3, each source will in addition to the frame starting point also be associated with a specific frame phase. Such a generalization will influence both the system response and the calculations of the likelihood ratio.

The smoothest multiplex of sources is obtained when frame phases are evenly spaced in the frame period. Each of the M sources may occupy one of M distinct and different frame phases, such that a in sequence of length N, there will then be $N \cdot M$ discrete starting positions that can be chosen in $M! \cdot N^M$ ways. The uniformly distributed probability of choosing any alignment with respect to frame and frame phases is then given by

$$P(l) = \frac{N \cdot M}{M! \cdot N^M} = 1/[(M-1)! \cdot N^{M-1}]. \tag{16}$$

The probability of choosing a specific phase-constellation is $1/M!$, so the biased alignment probability may be expressed as

$$P^*(l) = \frac{1}{M!} \sum_{j=1}^{N \cdot M} P^*(l,j) \tag{17}$$

The likelihood ratio of the l'th alignment is

$$\Lambda(l) = \frac{M/N^{M-1}}{\sum_{j=1}^{N \cdot M} P^*(l,\ j)}. \tag{18}$$

Letting the number of discrete phase values increase, a Riemann sum can be formed such that the likelihood ratio in the limit of continuously varying frame phases can be expressed as

$$\Lambda(l) = \frac{1/N^{M-1}}{1/T_f \int_0^{NT_f} P^*(l,t)dt} = \frac{1/N^{M-1}}{\sum_{\forall i} d_i/dP^*(l,t)}. \tag{19}$$

Here T_f denotes the frame duration, the sum in the last term is performed over all fixed-rate intervals in the simulation sequence and is the length of the i'th fixed-rate interval relative to the frame duration. It should be noted that the complexity of calculations for unsynchronised sources is a factor M higher than for synchronised ones.

The system response is $(n_0 = 0)$

$$Y(l) = \sum_{j=1}^{N \cdot M} \max(0, n_{j-1} + \frac{d_j}{d} \mathbf{X}_{\mathbf{k}_t} \cdot \mathbf{1} - d_j - B). \tag{20}$$

4.3 Network of Nodes

Evaluation of the performance in one node is interesting e.g. when studying bottleneck routers or access links. If end-to-end performance is of interest where no single bottleneck is expected, a network of nodes must be considered.

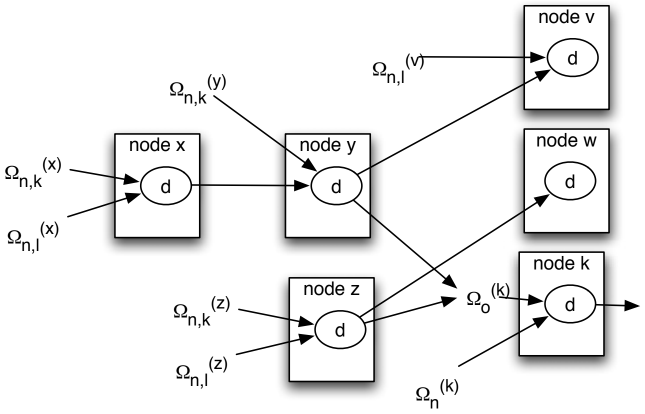

Figure 7: Network Of Nodes

The current importance sampling approach can be applied to a general network topology with no looping traffic. If the packet loss ratio is low, it can be assumed that the total loss ratio over several nodes can be calculated by decomposition, i.e. by considering the loss ratio at one node at the time because the simultaneously loss of packets in two or more nodes at nearly the same time are neglectable.

Under these assumptions the frame alignment described in the previous can be applied to node k considering all traces, both the new traces in $\Omega_n^{(k)}$ and the traces from other nodes in $\Omega_o^{(k)}$. These are the traces in the sets $\Omega_{n,k}^{(i)}$, $i = x, y, z$ in Figure 7. The extra index k indicates that the trace will be routed to node k. Observe that the nodes x, y, z in Figure 7 are offered other new traces that are not routed to the node k. This is indexed by $\Omega_{n,l}^{(i)}$, $i = x, y, z$ where $l \neq k$. These sets should not be manipulated while considering the node k. This strategy can be repeatedly applied to each node and the total loss ratio can be estimated by summation of each of these individual calculations if the losses in a node are independent of the loss in other nodes.

5 APPLICATION EXAMPLES

To study the applicability and efficiency of the proposed importance sampling (IS) strategy, this section contains two studies of trace driven simulation of multiplexing of video streams over one access line (single deterministic server process). The first study compares stratified sampling and the IS approach in a homogeneous case, while the second case shows the applicability of the IS approach on multiplexing of heterogeneous video streams (traces).

5.1 Compared to Stratified Sampling on Homogeneous MPEG Streams

In previous work (Andreassen, Emstad, and Riksaasen 1995), an alternative speed-up technique for MPEG simulation was applied, which was based on the use of *stratified sampling*, see e.g. (Lewis and Orav 1988) for an introduction. This section compares the importance sampling with stratified sampling and the direct simulation approach for application to a multiplex of homogeneous MPEG video sources. The comparisons are made based on an efficiency measure m, considering both the variability σ^2, the number of replications, R, and the CPU-time consumption t_{cpu}, over experiments, see (Heegaard 1995). Note that the most efficient technique will have the lowest measure.

$$m = \sigma^2/R \cdot t_{cpu}.$$

The comparison was carried out for two film sequences, MrBean and Bond, having different characteristics. A representative sample of the results are presented in Figure 8 showing the efficiency measure in a logarithmic scale for the three approaches. The main observations are:

- For high load values, direct simulation is always better than importance sampling and at least as good as stratified sampling. For high loads there are no rare events associated with the estimates, and the additional computer cost introduced by a speed-up technique is wasted.
- For a small number of sources, stratified sampling is at least as good as importance sampling.
- Importance sampling is better that stratified sampling even for a small number of sources of film-sequences having rather low maximum to mean ratio ($S_{\max}/S$) like the Bond sequence.
- Importance sampling is better than stratified sampling when the number of sources become large. As presented in (Andreassen, Emstad, and Riksaasen 1995), the stratified sampling model assumes synchronized and homogeneous sources, while importance sampling does not have these restrictions. Hence, in the following section, importance sampling will be used to speed-up the simulation of heterogeneous sources.

The relative efficiency of importance sampling to stratified sampling are illustrated in Figure 9, using load of 0.25 on the MrBean film-sequence and varying the number of sources from $M = 5$ to 12.

5.2 Applied to Heterogeneous MPEG Streams

In the simulation experiments, the number of simultaneous sources was regulated by multiplying the number of each source type, giving multiplexing scenarios with multiples of 19 sources. Thematic content and some traffic characteristics of the traces are described in (Rose 1995). Each calculation is based on 5000 independent simulations, and in figures, error-bars indicate the obtained 95% confidence intervals.

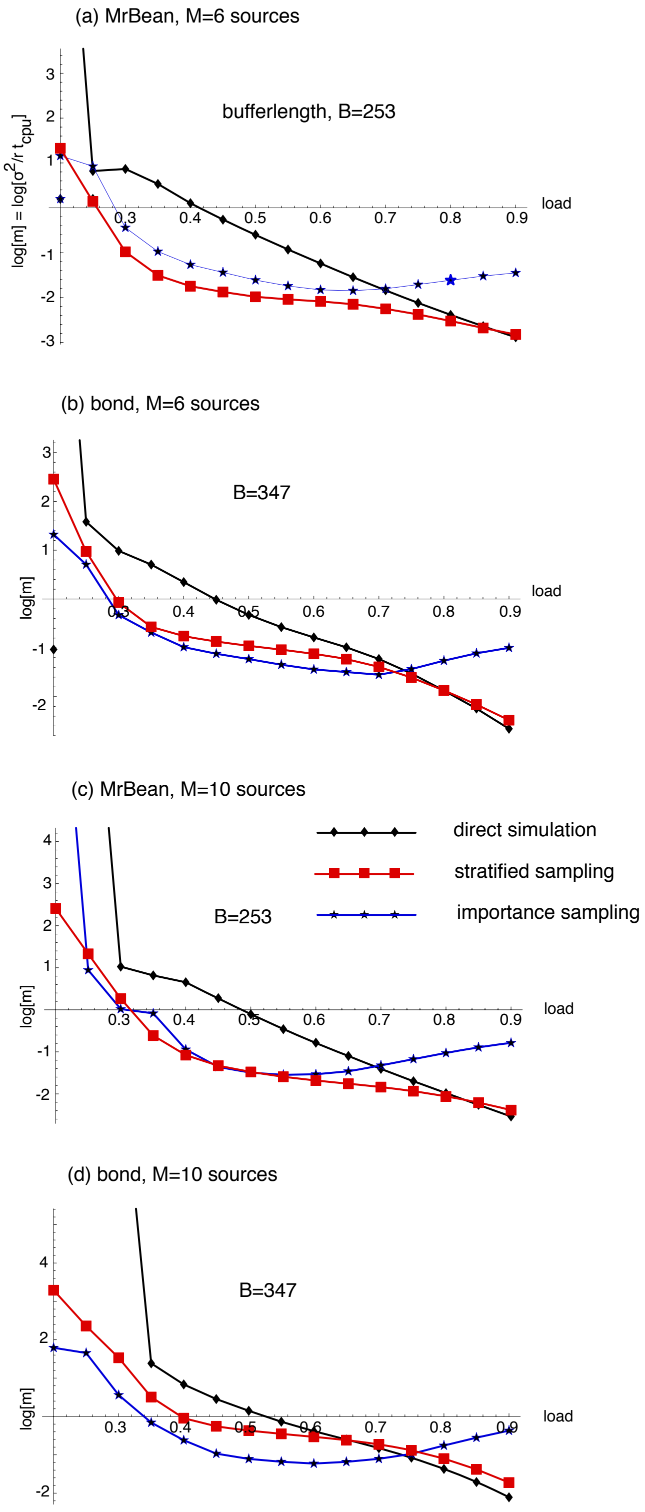

Figure 8: Comparison of Direct Simulation, Importance Sampling and Stratified Sampling With Respect to The Efficiency Measure [logarithmic scale]

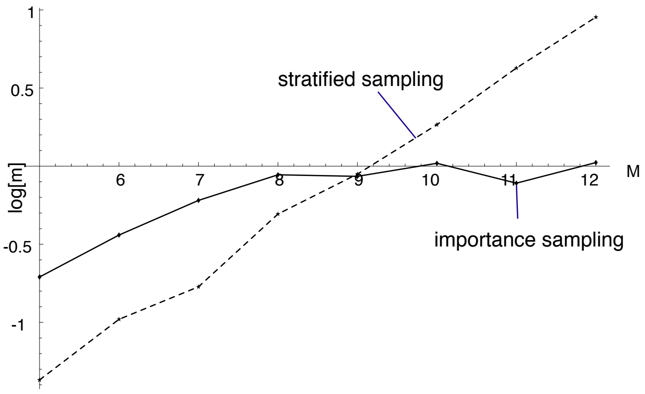

Figure 9: Efficiency Comparison As The Number of Source Increases

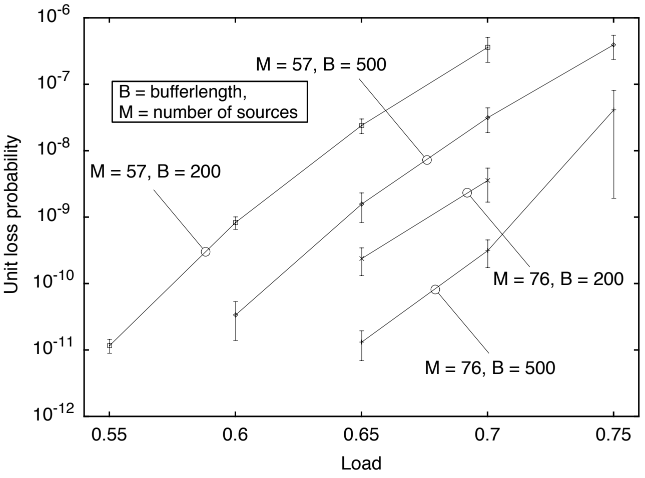

Figure 10: Loss Probabilities Of Multiplexed Sources, 95% Confidence Intervals

We have concentrated on multiplexers with moderately sized buffers.

From the results in Figure 10 it can be seen that as the load increases, the sampling bias tend to decrease the confidence. This can be observed for the highest loss values in the case with $M = 76$ and $B = 500$ [packets]. However, for high loss probabilities, direct simulation is a better option. As the number of sources increases, it is increasingly difficult to obtain good results, but it should be feasible at least when reducing the buffer size.

The method is sensitive to the buffer size even when using the load selection. The reason is that the load selection does not consider the buffer size and with large buffers the number of observations of system response where $Y_r = 0$, will be significant and the loss of packets are rare events also under the changed sampling distribution.

6 CLOSING REMARKS

Performance evaluation of realistic traffic from multimedia traffic sources with real-time requirements is very challenging. Analytical solutions can not handle a large number of sources and the complexity introduced by the diversity in the source behaviours and requirements. In addition, the analytically tractable parametric source models do not capture all source properties like the long lived correlations in video streams. Using simulations relaxes the constraints related to size and complexity, but faces the same problem with the parametric models. In addition, simulation can be rather inefficient due to the enormous number of events (e.g. packets) that has to be simulated for each event of interest (e.g. packet loss).

The importance sampling strategy presented in this paper will significantly increase the efficiency of simulation of traffic sources modelled as traces, i.e. sequences of packets. The current strategy is applicable to simulation of multiplexing of non-synchronised, heterogeneous traces on a single node. The importance sampling heuristics for changing the underlying sampling distribution are based on a combination of likely contributions to packet losses from the individual frames, and a systematic exclusion of irrelevant samples. The applicability is demonstrated on a case of multiplexing of MPEG encoded video sources.

The method can be applied to a general network topology where the loss of packets in each node can be assumed to be independent. Further testing is required in order to check the applicability and efficiency of the method in a network, e.g. applied on a cases with real-time video and VoIP traffic over a IP access network. Inclusion of congestion aware (e.g. TCP traffic) sources will improve the realism. This is an important issue for further study.

ACKNOWLEDGMENTS

This work is partially carried out at the Centre for Quantifiable Quality of Service in Communication Systems, Centre of Excellence appointed by The Research Council of Norway, funded by the Research Council, NTNU and UNINETT, `<http://www.ntnu.no/Q2S/>`.

REFERENCES

Andreassen, R. Ø. 1997, June. *Traffic performance studies of MPEG variable bitrate video over ATM*. PhD thesis, Norwegian University of Science and Technology.

Andreassen, R. Ø., P. J. Emstad, and T. Riksaasen. 1995, 22 - 24 August. Cell losses of multiplexed VBR MPEG sources in an ATM-multiplexer. In Norros, I., and J. Virtamo. (Eds.) 1995, 22 - 24 August. *The 12th Nordic Teletraffic Seminar (NTS-12)*, Espoo, Finland, 83–95.

Andreassen, R. Ø., P. E. Heegaard, and B. E. Helvik. 1996, 20 - 22 August. Importance sampling for speed-up simulation of heterogeneous mpeg sources. In *The 13th Nordic Teletraffic Seminar (NTS-13)*, ed. P. J. Emstad, B. E. Helvik, and A. H. Myskja, 190–203. Trondheim, Norway: Tapir Trykk.

Chang, C.-S., Y.-M. Chiu, and W. Song. 2001. On the performance of multiplexing independent regulated inputs. In *Proc. ACM Sigmetrics 2001/Performance 2001*, Volume 29, 184–193.

Devetsikiotis, M., and J. K. Townsend. 1993, June. Statistical optimization of dynamic importance sampling parameters for efficient simulation of communication networks. *IEEE/ACM Transactions on Networking* 1 (3): 293 – 305.

Floyd, S., and V. Paxson. 2001. Difficulties in simulating the internet. *IEEE/ACM Trans. Netw.* 9 (4): 392–403.

Glasserman, P., and Y. Wang. 1997. Counterexamples in importance sampling for large deviation probabilities. *The Annals of Applied Probability* 7 (3): 731 – 746.

Heegaard, P. E. 1995, 22 - 24 August. Comparison of speed-up techniques for simulation. In Norros, I., and J. Virtamo. (Eds.) 1995, 22 - 24 August. *The 12th Nordic Teletraffic Seminar (NTS-12)*, Espoo, Finland, 407–420.

Heegaard, P. E., and B. E. Helvik. 1999, March 11-12. On the use of likelihood ratio as indicator of the accuracy of importance sampling estimates. In *In Proceedings of Workshop on Rare Event Simulation (RESIM99)*. Twente University, The Netherlands.

Heidelberger, P. 1995, January. Fast simulation of rare events in queueing and reliability models. *ACM transaction on modeling and computer simulation* 5 (1): 43–85.

Jain, R. 1991. *The art of computer systems evaluation; techniques for experimental design, measurement, simulation and modelling*. Wiley: Wiley Professional Computing.

Lewis, P. A. W., and E. J. Orav. 1988.. *Simulation methodology for statisticians, operations analysts and engineers*, Volume I. Wadsworth & Brooks/Cole Advanced Books & Software.

Paschalidis, I. C., and S. Vassilaras. 2004, october. Importance sampling for the estimation of buffer overflow probabilities via trace-driven simulations. *IEEE/ACM Transactions on Networking* 12 (5): 907–919.

Rose, O. 1995, February. Statistical properties of MPEG video traffic and their impact on traffic modeling in ATM systems. Technical Report 101, University of Wuerzburg, Institute of Computer Science. The traces obtained from: `<ftp-info.informatik.uni-wuerzburg.de/pub/MPEG>`.

AUTHOR BIOGRAPHIES

POUL E. HEEGAARD is an associate professor at Department of Telematics at Norwegian University of Science and Technology (NTNU) and a senior scientist at Telenor R&D. He received his MSc (Siv. Ing.) in 1988 and his PhD (Dr.

Ing.) in 1998 from NTNU. His research interest is within the areas of performance and dependability evaluation of Telematics systems. He has special interests in speedup simulation techniques, and adaptive, distributed monitoring and management techniques in dynamic networks. His email address is <poul.heegaard@item.ntnu.no>.

BJARNE E. HEVIK is professor at Department of Telematics at Norwegian University of Science and Technology (NTNU). He is principal academic at the Norwegian Centre of Excellence (CoE) for Quantifiable Quality of service in Communication systems. He received his Siv.ing. degree (MSc) in 1975 and Dr. Techn. in 1982 from NTNU. His field of interests includes QoS, dependability modeling, measurements, analysis and simulation, fault-tolerant computing systems and survivable networks. His current research focus is on distributed, autonomous and adaptive fault-management in telecommunication systems, networks and services. His email address is <bjarne.helvik@q2s.ntnu.no>.

RAGNAR Ø. ANDREASSEN is senior scientist at Telenor R&D. dept of market and regulation. He received a MSc (Cand. Scient.) in Physics from the University of Oslo in 1988, a PhD (Dr. Ing.) in Telecommunications from the Norwegian University of Science and Technology in 1997. His research interests covers various areas such as network traffic and dependability performance, network architectures, charging, accounting and pricing for communication services. His email address is <ragnar.andreassen@telenor.com>.