Modelling of Unconstrained and Constrained H.26x Traffic over IP Networks

S. Kouremenos[1], S. Domoxoudis[2], V. Loumos[2], and A. Drigas[1]

July 31, 2008

1. National Center for Scientific Research “DEMOKRITOS”, Institute of Informatics and Telecommunications, P.O. Box 15310 Gr. Ag. Paraskevi, Attiki, Greece

2. National Technical University of Athens “NTUA”, School of Electrical and Computer Engineering, Multimedia Technology Laboratory, P.O. Box 15780 Gr. Zographou, Attiki, Greece

Abstract

In this manuscript, methods for modelling and parameter assessment of unconstrained and constrained videoconference traffic are proposed. In the case of unconstrained traffic the encoder operates in an independent of the network mode (open-loop) while in constrained traffic the encoder has knowledge of the networking constrains and operates using rate-control algorithms (in the loop). The analysis of extensive data that were gathered during experiments with popular videoconference terminals, as well as, of traffic traces available in literature, suggested that while the unconstrained traffic traces exhibited high short-term correlations, the constrained counterpart patterns appeared to be mostly uncorrelated, in a percentage not affecting queueing. On the basis of these results, this study discusses methods for accurate modelling and analytical treatment of both types of traffic. Extensive model-based queuing results, in single-source and multiplexed environments, using continuous methods, compared to trace-driven results, confirm the validity of our modelling proposals.

Keywords: videoconference traffic, unconstrained, constrained, network performance, VBR encoders, modelling, simulation, queueing, H.261, H.263, H.263+, DAR, C-DAR, multiplexing

Introduction

H.26x videoconference traffic is expected to account for large portions of the multimedia traffic in future heterogeneous networks (wire, wireless and satellite). The videoconference traffic models for these networks must cover a wide range of traffic types and characteristics because the type of the terminals will range from a single home or mobile user (low video bit rate), where constrained video traffic is mainly produced, to a terminal connected to a backbone network (high video bit rate), where the traffic is presented to be both constrained and unconstrained. Furthermore, successful videoconference traffic modelling can lead to a more economical network usage (improved traffic policing schemes), leading to lower communication costs and a more affordable and of higher quality service to the end-users.

Partly due to the above reasons, the modelling and performance evaluation of videoconference traffic have been extensively studied in literature and a wide range of modelling methods exist. The results of relevant early studies [2],[4],[5],[6],[7],[8],[9],[10] concerning the statistical analysis of variable bit rate videoconference streams being multiplexed in ATM networks, indicate that the histogram of the videoconference frame-size sequence exhibits an asymmetric bell shape and that the autocorrelation function decays approximately exponentially to zero. An important body of knowledge, in videoconference traffic modelling, is the approach in [7] where the DAR(1) [1] model was proposed. More explicitly, in this study, the authors noted that AR models of at least order two are required for a satisfactory modelling of the examined H.261 encoded traffic patterns. However, in the same study, the authors observed that a simple DAR(1) model, based on a discrete-time, discrete state Markov Chain performs better - with respect to queueing - than a simple AR(2) model. The results of this study are further verified by similar studies of videoconference traffic modelling [9] and VBR video performance and simulation ([8] and [12]). In [15], Dr Heyman proposed and evaluated the GBAR process, as an accurate and well performed single-source videoconference traffic model.

The DAR(1) and GBAR(1) models provide a basis for videoconference traffic modelling through the matching of basic statistical features of the sample traffic. On this basis and towards the modelling of videoconference traffic encoded by the Intra-H261 encoder of the ViC tool, the author in [18] proposed a DAR(p) model using the Weibull instead of the Gamma density for the fit of the sample histogram. In [19], the authors introduced a Continuous Markov chain model, called C-DAR, which is based on the DAR model and is suitable for theoretical analysis. In the same study, the authors concluded that Long Range Dependence (LRD) has minimal impact on videoconference traffic modelling (conclusion also declared in [14]). Looking at the C-DAR model as a Markov modulated rate process, as in [3], the same study's authors applied the fluid-flow method to compare the C-DAR versus a trace-driven simulation. The C-DAR(1) model, via the fluid-flow method, has the advantage of being analytically treated and as a result can be directly applicable to VBR video traffic engineering studies (used as a modelling validation method in the current study).

Relevant newer studies of videoconference traffic modelling reinforce the general conclusions obtained by the above earlier studies by evaluating and extending the existing models and also proposing new methods for successful and accurate modelling. An extensive public available library of frame size traces of unconstrained and constrained MPEG-4, H.263 and H.263+ off-line encoded video was presented in [21] along with a detailed statistical analysis of the generated traces. In the same study, the use of movies, as visual content, led to frames generation with a Gamma-like frame-size sequence histogram (more complex when a target rate was imposed) and an autocorrelation function that quickly decayed to zero (a traffic model was not proposed though in the certain study).

Of particular relevance to our work is the approach in [22], where an extensive study on multipoint videoconference traffic (H.261-encoded) modelling techniques was presented. In this study, the authors discussed methods for correctly matching the parameters of the modelling components to the measured H.261-encoded data derived from realistic multipoint conferences (in ”continuous presence” mode).

The above studies certainly constitute a valuable body of knowledge. However, most of the above studies examine videoconference traffic traces compressed by encoders (mainly H.261) that were operating in an unconstrained mode and as a result produced traffic with similar characteristics (frame-size histogram of Gamma form and strong short-term correlations). Today, a large number of videoconference platforms exist, the majority of them operating over IP-based networking infrastructures and using practical implementations of the H.261 [25], H.263 [26],[27] and H.263+ [26],[27] encoders[1]. The above encoders operate on sophisticated commercial software packages that are able of working in both unconstrained and constrained modes of operation. In unconstrained VBR mode, the video system operates independently of the network (i.e. using a constant quantization scale throughout transmission). In the constrained mode, the encoder has knowledge of the networking constraints (either imposed off-line by the

[1] Although a newer encoder, namely, H.264, exists, it is not in a compatible version yet and only two commercial video systems where found to support it, which could not establish a common H264 communication.

user or on-line by an adaptive bandwidth adjustment mechanism of the encoder) and modulate its output in order to achieve the maximum video quality for the given content (by changing the quantization scale, skipping frames or combining multiple frames into one). Furthermore, most of the previous studies have dealt with the H.261-encoding of movies (like Starwars) that exhibit abrupt scene changes. However, the traffic patterns generated by differential coding algorithms depend strongly on the variation of the visual information. For active sequences (movies), the use of a single model based on a few physically meaningful parameters and applicable to a large number of sequences does not appear to be possible. However, for videoconference, this is more probable as the visual information is a typical head and shoulders content that does not contain abrupt scene changes and is consequently more amenable to modelling. Moreover, an understanding of the statistical nature of the constrained VBR sources is useful for designing call admission procedures. Modelling constrained VBR sources, to the best knowledge of these authors, is an open area for study. Our approach towards this direction was to gather video data generated by constrained VBR encoders that used a particular rate control algorithm to meet a particular channel constraint and then model the resulting trace using techniques similar to those used for unconstrained VBR. The difficulty with this approach is that the resulting model could not be used to understand the behaviour of a constrained VBR source operating with a different rate control algorithm or a different channel constraint. However, given that in constrained VBR the encoder is in the loop, it is more likely that network constraints are not violated and that the source operates closer to its maximum allowable traffic. This may make constrained VBR traffic more amenable to modelling than unconstrained VBR traffic. The basic idea is that we can assume worst case sources (i.e. high motion contents), operating close to the maximum capacity and then characterize these sources.

Taking into account the above, it is important to examine whether the models established in literature are appropriate for handling this contemporary setting in general. It is a matter of question whether all coding strategies result in significantly different statistics for a fixed or different sequence. Along the above lines, this study undertook measurements of the videoconference traffic encoded, during realistic low and high motion head and shoulders experiments, by a variety of encoders of popular commercial software modules operating in both unconstrained and constrained modes. Moreover, the modelling proposal was validated with various traces available in literature [21] (to be referred as "TKN traces" from now on).

The rest of the manuscript is structured as follows: section 1 describes the experiment characteristics and presents the first-order statistical quantities of the measured data. Section 2 discusses appropriate methods for parameter assessment of the encoded traffic. In subsection 2.3, our modelling results are validated through the comparison of model-based and trace-driven simulations. Finally, section 3 culminates with conclusions and pointers to further research.

1 The experimental and measurement work

The study reported in this manuscript employed measurements of the IP traffic generated by different videoconference encoders operating in both unconstrained and constrained modes. More explicitly, we measured the traffic generated by the H.26x encoders[2] included in the following videoconference software tools: ViC (version v2.8ucl1.1.6) [32], VCON Vpoint HD [33], France Telecom eConf 3.5 [34] and Sorenson EnVision [35]. These are: H.261, H.263 and H.263+. All traces examined in the current study are representative of the H.26x family video systems. Especially, the ViC video system uses encoders implemented by the open H.323

[2] The NV, NVDCT, BVC and CellB encoders [11] were examined in [24] and it was found that they resulted in similar traffic patterns with the H.261 encoder. Thus the modelling proposal for H.261, in the current study, is applicable for these encoders.

community [36]. These encoders are based on stable and open standards and as a consequence their examination is more probable to give reusable modelling results.

For all the examined encoders, compression is achieved by removing the spatial (intraframe) and the temporal (interframe) redundancy. In intraframe coding, a transform coding technique is applied at the image blocks, while in interframe coding, a temporal prediction is performed using motion compensation or another technique. Then, the difference or residual quantity is transform coded. Here, we must note that the ViC H.261 encoder [13], [16] performs only intraframe coding oppositely to the H.261 encoders of Vpoint, eConf and EnVision, where blocks are inter or intra coded. The above encoding variations influence the video bit rate performance of the encoders and as a consequence the statistical characteristics of the generated traffic traces. It is a matter of question, consequently, if the different encoders' traffic can be captured by a common traffic model.

At this point, we may discuss about the basic functionality of the examined video systems which is a fundamental factor in the derived statistical features of the encoded traffic and a basic reason of the experiments' philosophy we followed. The rate control parameter (bandwidth and frame rate) sets a traffic policy, i.e. an upper bound on the encoded traffic according to the user's preference (obviously depending on his/her physical link). An encoder's conformation to the rate control of the system is commonly performed by reducing the video quality (and consequently the frame size quantity) through the dynamic modulation of the quantization level. In the case of ViC, a simpler method is applied. The video quality remains invariant and a frame rate reduction is performed when the exhibited video bit rate tends to overcome the bandwidth bound. In fact, in ViC, the video quality of a specific encoder is a parameter determined a priori by the user. In the case of Vpoint, eConf and EnVision, the frame rate remains invariant and a video quality reduction is performed when the exhibited video bit rate tends to overcome the bandwidth bound. This threshold can be set through the network setting of each client. Moreover, Vpoint utilizes adaptive bandwidth adjustment (ABA). ABA works primarily by monitoring packet loss. If the endpoint detects that packet loss exceeds a predefined threshold, it will automatically drop to a lower conference data rate while instructing the other conference participant's endpoint to do the same.

Two experimental cases were examined in the current study as presented in Table 1 (TKN traces are also included). Case 1 included experiments where the terminal clients were operating in unconstrained mode while Case 2 covered constrained-mode trials. In both Cases, two "talking-heads" raw-format video contents were imported in the video systems through a Virtual Camera tool [37] and then peer-to-peer sessions of at least half an hour were employed in order to ensure a satisfactory trace length for statistical analysis. These contents were offline produced by a typical webcam in uncompressed RGB-24 format: one with mild movement and no abrupt scene changes, "listener", (to be referred as VC-L) and one with higher motion activities and occasional zoom/span, "talker" (VC-H). The video size was QCIF (176x144) in both Cases and all scenarios (VC-H and VC-L). In Case 1, no constraint was imposed either from a gatekeeper or from the software itself. The target video bit rates that were imposed in Case 2 are shown in Table 1. In each case, the UDP packets were captured by a network sniffer and the collected data were further post-processed at the frame level[3] by tracing a common packet timestamp. The produced frame-size sequences were used for further statistical analysis.

Specific parameters shown in Table 1, for the VC-H and VC-L traces, depend on the particular coding scheme, the nature of the moving scene, and the confidence of the measured statistics. Moreover, traffic traces available in literature where used for further validation. Specifically, the traces used were: "office cam" and "lecture room cam" (from the TKN library). These

[3] It is important to note, here, that analysis at the MacroBlock (MB), as in [20], level has been examined and found to provide only a typical smoothing in the sample data. We believe that the analysis at the frame level is simpler and offers a realistic view of the traffic.

traces were offline H.263 encoded in a constrained and unconstrained (no target bit rate was set during offline encoding) mode.

Some primary conclusions, as supported by the experiments' results (see Table 1), arise concerning the statistical trends of the encoders' traffic patterns. Specifically, H.263+ produces lower video bit rate than H.263 and H.261 do. This was expected, since the earlier encoder versions have improved compression algorithms than the prior ones (always with respect to the rate produced). Finally, for all the encoders, the use of the VC-H content led to higher rate results (as reasonable). Similar results were observed for the mean frame size and variance quantities. In all cases, the variance quantities of the VC-H content were higher than that of VC-L with the exception of the ViC H.263+ encoder (Case 1 – Traces 5,6) where the opposite phenomenon appeared.

The encoders used for the production of the TKN traces tend to adjust their quality in a "greedy" manner so as to use up as much of the allowed bandwidth as possible. At this point, we must note that Trace 4 of Case 2 is semi-constrained (i.e. the client did not always need the available network bandwidth). However, this particular case can be covered by the "worst-case" Case 2 – Trace 3, where the target rate is reached (full-constrained traffic).

Taking into account the above context, the following questions naturally arise:

- What is the impact of the encoders' differences on the generated videoconference traffic trends?
- Can a common model capture both types of traffic, unconstrained and constrained?
- Are the traffic trends invariant of the constraint rate selected?
- How does the motion of the content influence the generated traffic - for each encoder - and the parameters of the proposed traffic model?
- Can a common traffic model be applied for all the above cases?

The above questions pose the research subject which is thoroughly examined in the context to follow. Their answers will be given along with the respective analysis.

2 Traffic analysis and modelling assessment

The measured traffic analysis for all experimental sets confirms the general body of knowledge that literature has formed concerning videoconference traffic. Traffic analysis was employed for all experimental cases. More explicitly, in all cases, the frame-size sequence can be represented as a stationary stochastic process, with a frequency histogram of an approximately bell-shaped (more narrow in the case of H.263 and H263+ encoding) Probability Distribution Function (PDF) form, see figures 1((a)-(c)), 2((a)-(c)) more complex in the TKN traces as their content (office and lecture cam) probably contained more scene changes than our contents VC-L and VC-H. Examining more thoroughly the sample histograms, we noted that the smoothed frame-size frequency histograms of the H.261 encoder have an almost similar bell-shape (see figures 1((a),(b)) and figures 2((a),(b))) while a more narrow shape appears in the H.263 and H.263+ histograms (figures 1(c) and 2(c). The VC-H frequency frame-size histograms appeared to be more symmetrically shaped than the correspondent VC-L histograms. This is reasonable as the rate of the H.26x encoders depends on the activity of the scene, increasing during active motion (VC-H) and decreasing during inactive periods (VC-L).

Furthermore, the AutoCorrelation Function (ACF) of the unconstrained traffic (for all traces of Case 1) appeared to be strongly correlated in the first 100 lags (short-term) and slowly decaying to values near zero (see some indicative figures 3((a)-(c)) of the traces of Case 1). On the contrary, the ACFs of the constrained traffic (Case 2) decayed very quickly to zero denoting the lack of short-term correlation (see figures 3((d)-(f)). This conclusion is very critical in queueing as the short-term correlation parameter has been found to affect strongly buffer occupancy and overflow probabilities for videoconference traffic. In fact, to verify this assumption, we measured the buffer occupancy of the constrained traces in queueing experiments of different traffic intensities. Buffering was found to be very small at a percentage not affecting queueing. On this basis, it is evident, that for the purpose of modelling of the two types of traffic not a common model can be applied. More explicitly, a correlated model is needed for the case of unconstrained traffic while a simpler non-correlated model is enough for constrained traffic.

The DAR model, proposed in [7], has an exponentially matching autocorrelation and so matches the autocorrelation of the data over approximately hundred frame lags. This match is more than enough for videoconference traffic engineering. Consequently, this model is a proper solution for the treatment of unconstrained traffic. When using the DAR model, it is sufficient to know the mean, variance and autocorrelation decay rate of the source, for admission control and traffic forecasts. A negative feature of the DAR model is that it exhibits "flat spots'" which make its sample paths "look" different from those of the data when comparisons are made for a single source (for multiplexed data sources they are indistinguishable). Though these flat spots may not affect traffic engineering, there is another model which is more specialized for modelling accurately the short-term fluctuations of single teleconference sources, namely, GBAR. However, the GBAR model cannot be applied in our study as it is exclusively based on the Gamma density (except from the cases where the Gamma density is proposed).

For the constrained traffic traces, a simple random number generator based on the fit of the sample frame-size histogram can be directly applied. The DAR model with the autocorrelation decay rate value equal to zero can also be a solution. This feature turns constrained videoconference traffic more amenable to traffic modelling than its counterpart unconstrained as only two parameters are needed, the mean and the variance of the sample.

The rest of the paper discusses methods for correctly matching the parameters of the modelling components to the data and for combining these components into the DAR model (to be analytical treated via the C-DAR and the fluid-flow method for unconstrained traffic).

2.1 Fitting of the frame-size frequency histograms of the traces

A variety of distributions was tested for fitting the sample frame-size frequency histograms. These are the following: Gamma, Inverse Gamma (or Pearson V), Loglogistic, Extreme Value, Inverse Gauss, Weibull, Exponential, Lognormal. The most dominant ones found to be the first three. Even though the Inverse Gauss density performed similarly to the Gamma distribution, it is not included in the analysis to follow, as the Gamma distribution is more popular and simpler. Finally, the Extreme Value distribution performed, in total, worse than the other ones.

For the purpose of fitting the selected distributions' density to the sample frame-size sequence histogram, although various full histogram-based methods (e.g. [22]) have been tried in literature, as well as maximum likelihood estimations (MLE), we followed the approach of the simple moments matching method. This method has the positive feature of requiring only the sample mean frame size and variance quantities and not full histogram information. Thus, taking into account that the sequence is stationary - and as a result the mean and the variance values are almost the same for all the sample windows - it is evident that only a part of the sequence is needed to calculate the corresponding density parameters. Furthermore, this method has the feature of capturing accurately the sample mean video bit rate, a property that is not ensured in the case of MLE or histogram-based models. However, in the cases of not satisfactory fit by none of the examined distributions (as in the case of the TKN traces) a histogram-based method can be applied as an unconvential fitting method.

If m the mean, v the variance of the sample sequence s and m_l the mean and v_l the variance of the logarithm of the sample s, then the distribution functions and the corresponding parameters derived from the moments matching method are given by the following equations, for each distribution correspondingly, Eq. (1): Gamma, Eq. (2): Inverse Gamma, Eq. (3): Loglogistic.

$$f(x) = \frac{1}{\beta\Gamma(\alpha)}\left(\frac{x}{\beta}\right)^{\alpha-1} e^{-\frac{x}{\beta}} \tag{1}$$

where $\alpha = m^2/v$, $\beta = v/m$ and $\Gamma(\alpha) = \int_0^\infty u^{\alpha-1}e^{-u}du$

$$f(x) = \frac{1}{\beta\Gamma(\alpha)} \cdot \frac{e^{-\frac{\beta}{x}}}{\left(\frac{x}{\beta}\right)^{(a+1)}} \tag{2}$$

where $a = m^2/v + 2$ and $\beta = m\,(m^2/v + 1)$

$$f(z) = \frac{e^{z(1-\sigma)-m_l}}{(1+e^z)^2\sigma} \tag{3}$$

where $z = (\ln(x) - m_l)/\sigma$, $m_l = \mathbb{E}[\ln(s)]$ and $\sigma = \sqrt{3\mathbb{E}[\ln(s)]}/\pi$

Given the dominance of the above distributions, modelling analysis and evaluation will be presented for the above three densities. The numerical results (densities' parameters) from the application of the above parameters-matching methods appear in Table 2. At this point, we must note that the Loglogistic density (3), although it provides better fits in the H.263+ cases (as will be commented upon later) exhibits mean and variance values that slightly deviate from the sample counterpart values. However, this is negligible with respect to queuing as concluded by the fluid-flow simulations presented in Section 3.

The modelling evaluation of the above methods has been performed from the point of queueing. As a consequence, we thoroughly examined fits of cumulative distributions. This was done as follows: we plotted the sample quantiles from the sample cumulative frequency histogram and the model quantiles from the cumulative density of the corresponding distribution. The Q-Q plot of this method refers to cumulative distributions (probabilities of not exceeding a threshold).

Figures 1((a)-(c)) and figures 2((a)-(c)) present Q-Q plots for all traces of both Case 1 and 2 respectively. The results suggest that for fitting videoconference data, the coding algorithm used should be taken into consideration. There seems to be a relationship between the coding algorithms and the characteristics of the generated traffic. For instance, for H.261, in most cases, the dominant distribution is Gamma (1), as can be verified from the Q-Q plots depicted in figures 1((a),(b)), and for H.263 and H.263+, the Loglogistic density (3) has a more "stable" performance than the other two (Q-Q plots shown in figures 1(c) and 2(c). The Inverse Gamma density (2) seems to be suitable for H.263 traffic (see figure 1(c)) although it was outperformed by the Loglogistic density in some cases. However, as will be commented upon later, it did not provide a solution in all cases of constrained traffic.

We must note that in Case 2, where a constrain was imposed, the moments matching method for calculating the distribution's parameters did not always provide a good fit, and performed as shown in figures 2((a)-(c)) (Inverse Gamma and Loglogistic are depicted. The Gamma density provided similar fit). To provide an acceptable fit, a histogram-based method proposed for H.261 encoded traffic in [22], known as C-LVMAX, was used. This method relates the peak of the histogram's convolution to the location at which the Gamma density achieves its maximum and to the value of this maximum. The values of the shape and scale parameters of the Gamma density are derived from: $a = (2\pi x_{max}^2 f_{max}^2 + 1)/2$ and $b = 1/(2\pi x_{max} f_{max}^2)$ where f_{max} is the unique maximum of the histogram's convolution density at x_{max}. Numerical values for this fit appear also in Table 2 (for Case 2 only). Figures 2((a)-(c)) show how the three distributions fit the empirical data using the method of moments (Inverse Gamma, Loglogistic) and the C-LVMAX method (Gamma C-LVMAX). The Inverse Gamma density could not be calculated for all the constrained traces (Case 2 – Traces 5,6,8,9), due to processing limitations (for large a, b parameters the factor $\left(\frac{1}{\beta}\right)^{(a+1)}$ in Eq. (2) is very small, near zero, and consequently its inverse quantity could not be calculated[4]).

Summing up the above analysis, it is evident that the Gamma density is better for H.261 unconstrained traffic, the Loglogistic for unconstrained H.263, H.263+ traffic and the C-LVMAX method for all cases of constrained traffic. However, if a generic and simple model needed to be applied for all cases then the most dominant would be the Loglogistic density.

2.2 Calculation of the autocorrelation decay rate of the frame-size sequences

At this point, we may discuss about the calculation of the autocorrelation decay rate of the frame-size sequence of the unconstrained traces (as denoted in the previous sections, constrained traffic appeared to be uncorrelated and as a result the decay rate of its autocorrelation function can be set to zero). From the figures 3((a)-(d)), it is observed that the ACF graphs of unconstrained traffic exhibit a reduced decay rate beyond the initial lags. It is evident that unconstrained video sources have very high short term correlation, feature which cannot be ignored for traffic engineering purposes. This is a behaviour also noted in earlier studies [6].

To fit the sample ACF, we applied the model proposed in [22] that is based on a compound exponential fit. This model fits the autocorrelation function with a function equal to a weighted sum of two geometric terms:

$$\rho_k = w\lambda_1^k + (1-w)\lambda_2^k \tag{4}$$

where λ_1, λ_2 are the decay rates with the property: $|\lambda_2| < |\lambda_1| < 1$. This method was tested with a least squares fit to the autocorrelation samples for the first 100, since the autocorrelation decays exponentially up to a lag of 100 frames (short-term behavior) or so and then decays less slowly (long-term behavior). This match is more than enough for traffic engineering, as also noted in [28]. What is notable is that using this model, the autocorrelation parameter ρ is chosen *not* at $lag-1$, as in DAR model. For each encoder (in Case 1), the parameter numerical values of the above fit appear in Table 2.

In the section to follow, the discussed modelling components are combined into complete traffic models with the C-DAR method. Furthermore, the different modelling parameters are validated comparing sample-based against model-based fluid-flow simulations in a single-server queueing system. This analysis is performed only for unconstrained traffic as for constrained traffic, the Q-Q plots are enough for modelling validation purposes.

2.3 Queueing analysis via the C-DAR model and the fluid-flow method – modelling validation

The C-DAR model that was proposed and used analytically in [19] can be directly applied for full modelling and analytical treatment of H.26x unconstrained (correlated) traffic over IP networks. This model is defined as a continuous-time discrete-state Markov chain with a transition rate matrix Q of the form:

$$Q = f_c(P - I) \tag{5}$$

[4] However, the values of the parameters of the Inverse Gamma density for Case 2 – Traces 5,6,8,9 are given in Table 2.

where $f_c = \dfrac{\ln\rho}{\rho-1}f$, $P = \rho\mathrm{I} + (1-\rho)A$ from the DAR(1) model [7] with ρ the autocorrelation decay rate derived from Eq.(4), f is the frame rate of the videoconference traffic,I is the identity matrix and A is a rank-one stochastic matrix with all rows equal to the probabilities resulting from the fit of the selected distribution. The C-DAR model demands the representation of the frame-size sequence with a constant number of states, whose probabilities values will fill the rows of the stochastic matrix A. These states can be easily chosen by dividing the interval between the maximum and the minimum frame size of the sequence into M frame-size states. So, if $x_{\min}$ is the minimum and $x_{\max}$ the maximum frame-size value then a reasonable state step n is $n = (x_{max} - x_{min})/M$, with n rounded to the nearest integer. The rate of each state can be easily calculated by the relative mean rate of a histogram window, as follows: if $\mathbb{P}_i$ is the probability mass of frame size S_i (derived from the corresponding density) then the rate value of the state value is equal to $f\sum_{i=1}^{n}\mathbb{P}_i S_i/\sum_{i=1}^{n}\mathbb{P}_i$. The value of the autocorrelation decay rate ρ should be chosen equal to the parameter λ_1 of the model used to fit the ACF in Eq.(4) (see Table 2) and the elements for the rows of table A should be determined through the fit produced by the PDF models (with parameters chosen from Table 2).

Following the approach in [19], the C-DAR model - as a continuous-time Markov chain model - is suitable for theoretical analysis using the fluid flow method (see also [3],[29],[30]. The above scheme is a very fast and simple queueing analysis method for VBR video traffic. Dr K. Kontovasilis provided us with a Matlab implementation of the above scheme, namely, ”genflow”. The “genflow” program takes as input the characteristics of N statistically identical fluid-flow Markov-Modulated sources (with global matrices $Q_g = Q \otimes Q \otimes ... \otimes Q$ from Eq.(5) and state space compressed to $\begin{pmatrix} \mathrm{N+M} \\ \mathrm{M} \end{pmatrix}$ states due to the statistical identical feature of the superposed streams) and solve the congestion problem of those N sources being statistically multiplexed over a multiplexer with infinite buffering capabilities. This program has been used in other studies too (see [31]). This method is analyzed as follows: consider a single server queueing system fed by videoconference traffic $r(t) \geq 0$ as a Markov modulated rate process according to the C-DAR model with a finite number of M states and transition rate matrix Q_g. More explicitly, in each state $i = 1 \dots m$, we correspond a video rate r_i. If $\mathbf{\Pi}$ is the corresponding steady state probability vector, then the mean input rate $\overline{r}$ is calculated as follows: $\overline{r} = \sum_{i=1}^{\mathrm{M}} \mathbf{\Pi}_i r_i$. The mean rate of the calculated rate vector captures always the mean rate of the sample (with a slight deviation in the case of the Loglogistic density). Let $R = diag\{r_1 \dots r_{\mathrm{M}}\}$ and C be the constant server capacity. When $r(t) > C$, the input traffic cannot be served entirely and its excess part is stored into a buffer in order to be served later. Let $\{X(t), t \geq 0\}$ be the stochastic process that represents the buffer occupancy. It is noted that the traffic intensity of the system is equal to $\overline{r}/C$. Define the steady state distribution $F_i(x)$ as the joint probability that the buffer occupancy is less than or equal to x when in the i state of the source model. Let: $F(x) = [F_1(x), F_2(x), \dots, F_{\mathrm{M}}(x)]^T$. Then from [29],[30], we have the differential equation:

$$\frac{dF(x)}{dx}D = F(x)Q_g \tag{6}$$

where $D = R - CI$. Given the infinite buffer assumption, we determine a buffer threshold B and define the buffer overflow probability as follows:

$$P_{overflow} = 1 - F(B)\mathbf{1} \tag{7}$$

where $\mathbf{1} = (1, ..., 1)^T$. From Eq. (6) and the boundary conditions for the infinite buffer size approach in [3],[29],[30], the following relation holds:

$$F(x) = \sum_{i=1}^{\mathrm{M}} \alpha_i e^{z_i x}\phi_i \tag{8}$$

where the coefficients a_i must be calculated from the boundary conditions and z and ϕ are, correspondingly, the eigenvalue and the left eigenvector of the matrix $Q_g D^{-1}$. Given the infinite buffer assumption, the solution of Eq. (8) is given as follows:

$$F(x) = \mathbf{\Pi} + \sum_{i \in S_o} a_i e^{z_i x} \phi_i \tag{9}$$

where $S_o \overset{\triangle}{=} \{j | r_j > C\}$, $z_i < 0$ and $z_1 = 0$.

Using the above method (with the assumption of a finite buffer), the authors in [19], proved experimentally (comparing the analytical model versus trace-driven simulation) that the C-DAR model provides accurate queueing results (mean cell loss rate, mean queue length) and therefore is suitable for theoretical analysis of videoconference traffic. To validate the modelling proposals of the previous sections, we present experimental queueing results comparing the complementary distribution of the buffer overflow given by the C-DAR Markov chain as derived from the calculation of Eq. (7) and (9) for any value of buffer threshold B versus the one given by a discrete-event simulation [17] using the actual traces (trace-driven simulation e.g. [21]). For a variety of Case 1 traces[5], the complementary buffer size densities from the results of the fluid-flow method for all the examined distribution models (for different values of multiplexed sources N) and the corresponding sample (derived from the discrete simulation of the trace being multiplexed[6] and frame interarrival times equal to $1/f$) are plotted together (see figures 4((a)-(n)). The probabilities values are always assigned at the logarithmic scale. The traffic intensity was chosen equal to 0.85[7], the autocorrelation decay rate properly chosen from Table 2 and the number of states of the Markov chain M equal to 5 (increasing the number of states higher than five led to identical results). The comparison between the simulation and the analytical results gives a clear indication of the queueing performance of the proposed models for unconstrained videoconference traffic. As can be seen there is quite good agreement between all curves, apart from the fact that the curve deviates from those derived from by analysis from small buffer sizes. This is physical as with the fluid-flow method the discreteness of the buffer occupancy is neglected. Moreover, in all plots where more than one sources were multiplexed the multiplexing gain property led to more conservative results for the models (asymptotically tight though). More explicitly, it is concluded that, in most cases of single source service (e.g. see couples of figures (1(a), 4(a)) - (1(c), 4(d))) the models that exhibited the best fits in Q-Q plots provided closely accurate queueing results. This fact constitutes the selection of the ACF decay rate at the first 100 lags valid for unconstrained videoconference traffic engineering purposes. However, there are some obvious deviations from the above conclusion where the sample results are roughly fitted by the models. This phenomenon, which is less intense in the case of multiplexing of the same sources due to the multiplexing gain property, claims the existence of notable long-term trends in the ACFs of the respective traces (Case 1 – Traces 9, 11, 12). It is evident that in these particular cases, more than 100 lags (this was also remarked in [22] where the authors proposed a fit in the first 500 lags of the ACF) are needed to capture the strong correlation structure of the traffic. This can be also verified by the bad queueing performance of the Loglogistic density, despite its better fitting behavior in the corresponding Q-Q plots. Though for the majority of the examined cases, where a 100-lag fit was found to be accurate, a network administrator could choose a fit at 500 lags in order to ensure the conservativeness of single-source queueing results. For multiplexing this is already ensured.

[5] Queueing analysis for VC-H and VC-L traces of the same encoder and video system was found to be similar. Consequently, for brevity reasons, only the "worst-case" VC-H traces are examined.

[6] In trace-driven multiplexing, the first frame occurrence of each source was randomized over the interval of a frame and then the source kept its individual frame synchronization.

[7] If the models retain close to the sample at high traffic intensities, their applicability is ensured for lower values of the traffic intensities (as declared in [2]). This property, though, was tested and expected results were found.

Regarding the constrained (uncorrelated) traffic, it is repeated that there is no need to perform simulations, since the buffering is done inside the encoder. The traffic can be captured by the Gamma density calculated via the C-LVMAX method, as can be validated from the Q-Q plots shown in figures 2((a)-(c)).

3 Conclusions

The current study is a contribution of modelling and simulation results for a variety of existing videoconference encoders for talking heads communication. An extensive analysis of the measured data, a careful but simple modelling of the frame-size sequences and the extensive evaluation of the modelling components, led us to the general conclusion that the traffic can be distinguished into two categories: unconstrained and constrained. In the unconstrained traffic, strong correlations between successive video frames can be found (and a large buffer has to be used for better performance). On the other hand, where bandwidth constraints are imposed during the encoding process, the generated traffic is uncorrelated.

We used the measured data to develop statistical traffic models for unconstrained and constrained traffic. These models were further validated with different videoconference contents (low motion and high motion, TKN library). Different statistical models for fitting the empirical distribution (method of moments and C-LVMAX method) were examined.

For fitting the videoconference data, the coding algorithm used should be taken into consideration. There seems to be a relationship between the coding algorithms and the characteristics of the generated traffic. For instance, for H261, in most cases, the dominant distribution is Gamma, and for H263 and H263+, Loglogistic has a more ”stable” performance. Moreover, the Inverse Gamma density could not be calculated for all constrained traces, due to processing limitations. This fact constitutes the Inverse Gamma density as impractical as a generic model for H.263 traffic.

Regarding the unconstrained traces, a careful but simple generalization of the DAR model can simulate conservatively and steadily the measured videoconference data. The model was further verified using the Continuous version of the DAR model, namely, C-DAR model (analytical solution). For the constrained traces, the traffic can be captured by the C-LVMAX method via a random number generator, producing frames at a time interval equal to the sample. On the other hand, if a moments matching method needed to be applied, then the Loglogistic density is a direct solution. Another interesting assumption is that the traffic trends remain invariant when a different network constrained is selected, as evident from the TKN traces. So, the proposed model for the constrained traffic can be applied without taking into account the specific network constraint.

It is evident that if a generic and simple model needed to be applied for all cases of videoconference traffic then the most dominant would be the DAR model based on the fit of the Loglogistic density with a decay rate properly assigned to the fit of the sample ACF at the first 100 lags (although a 500 lags fit would lead to a more conservative queueing performance), for the case of unconstrained traffic, and to zero for the constrained traffic.

Future work includes the integration of the proposed models in dynamic traffic policy schemes in real diffserv IP environments. Careful analysis and modelling of cases of semi-constrained traffic, although their counterpart ”worst-case” full-constrained cases cover their traffic trends, is of particular interest, too.

Acknowledgements

The authors would like to thank Dr K. Kontovasilis for valuable suggestions and for providing the “genflow” program. Finally, we wish to thank the anonymous reviewers for their helpful

comments.

References

[1] P.A. Jacosb and P.A.W.Lewis, Time series generated by mixtures, J. Time Series Analysis, vol. 4, no. 1, pp. 19-36, 1983.

[2] B. Maglaris, D. Anastassiou, P. Sen, G. Karlsson and J. D. Robbins, Performance models of statistical multiplexing in packet video communications, IEEE Trans. Commun., 36: 834–843, 1988.

[3] D. Mitra, Stochastic theory of a fluid model of producers and consumers coupled by a buffer, Adv. Appl. Prob., vol. 20, p. 646-676, 1988.

[4] R. Kishimoto, Y. Ogata, and F. Inumaru, Generation interval distribution characteristics of packetized variable rate video coding data streams in an ATM network, IEEE JSAC, 7:833–841, 1989.

[5] H. S. Chin, J. W. Goodge, R. Griffiths and D. J. Parish, Statistics of video signals for viewphone-type pictures, IEEE JSAC, 7:826–832, 1989.

[6] M. Nomura, T. Fujii and N. Ohta, Basic characteristics of variable rate video coding in ATM environment, IEEE JSAC, 7:752–760, 1989.

[7] D. P. Heyman, A. Tabatabai and T. V. Lakshman, Statistical analysis and simulation study of video teleconference traffic in ATM networks, IEEE Trans. Circuits Syst. Video Technol., 2:49–59, 1992.

[8] D. M. Cohen and D. P. Heyman, Performance modelling of video teleconferencing in ATM networks, IEEE Trans. Circuits Syst. Video Technol., 3:408–422, 1993.

[9] D.P. Heyman, T.V. Lakshman, Modelling Teleconference Traffic from VBR Video Coders, IEEE ICC, pp.1744-1748, 1994.

[10] D.M. Lucantoni, M.F. Neuts, Methods for Performance Evaluation of VBR Video Traffic Models, IEEE/ACM Transactions on Networking, Vol.2, No.2, pp. 176-180, 1994.

[11] R. Frederick, Experiences with real-time software video compression, Xerox Parc, 1994.

[12] A. Elwalid, D. Heyman, T. V. Lakshman, D. Mitra, and A. Weiss, Fundamental bounds and approximations for ATM multiplexers with applications to video teleconferencing, IEEE JSAC, 13:1004-1016, 1995.

[13] S. R. McCanne, Scalable Compression and Transmission of Internet Multicast Video, Report No. UCB/CSD-96-928, Computer Science Division (EECS), University of California, Berkeley, California 94720, 1996.

[14] A. Erramilli, O. Narayan, and W. Willinger, Experimental queueing analysis with long-range dependent packet traffic, IEEE/ACM Trans. Networking, 4:209-223, 1996.

[15] D. P. Heyman, The GBAR source model for VBR videoconferences, IEEE/ACM Trans. Networking, 5: 554-560, 1997.

[16] L. D. McMahan, Video Conferencing over an ATM Network, Thesis, California State University, Northridge, 1997.

[17] M. Law and W. D. Kelton, Simulation Modelling and Analysis - 3nd Edition, McGraw-Hill Higher Education, 1999.

[18] B. Ryu, Modelling and Simulation of Broadband Satellite Networks: Part II-Traffic Modelling, IEEE Communications Magazine, 1999.

[19] S. Xu, Z. Huang, and Y. Yao, An analytically tractable model for video conference traffic, IEEE Trans. Circuits Syst. Video Technol., 10:63–67, 2000.

[20] G. Sisodia, L. Guan, M. Hedley, S. De, A New Modelling Approach of H.263+ VBR Coded Video Sources in ATM Networks, RealTimeImg, No. 5, pp. 347-357, 2000.

[21] F. Fitzek and M. Reisslein, MPEG-4 and H.263 video traces for network performance evaluation, IEEE Network, vol. 15, no. 6, pp. 40-54, 2001.

[22] C. Skianis, K. Kontovasilis, A. Drigas and M. Moatsos, Measurement and Statistical Analysis of Asymmetric Multipoint Videoconference Traffic in IP Networks, Telecommunications Systems, Volume 23, Issue 1. pp. 95-122, 2003.

[23] F. Fitzek, P. Seeling, M. Reisslein, Using Network Simulators with Video Traces, web site.

[24] S. Domoxoudis, S. Kouremenos, V. Loumos and A. Drigas, Measurement, Modelling and Simulation of Videoconference Traffic from VBR Video Encoders, Second International Working Conference, HET-NETs '04, Performance Modelling and Evaluation of Heterogeneous Networks, Ilkley, West Yorkshire, U.K., 26 - 28 July, 2004.

[25] ITU Recommendation, H.261: Video codec for audiovisual services at 64 kbit/s, 1993.

[26] ITU Recommendation, H.263: Video coding for low bit rate communication, 2005.

[27] Ç.263 Standard, Overview and TMS320C6x Implementation, White Paper, www.ubvideo.com.

[28] T. Lakshman, A. Ortega, and A. Reibman, Variable-Bit-Rate (VBR) Video: Tradeoffs and Potentials, Proceedings of the IEEE, 1998.

[29] D. Anick, D. Mitra, M. M. Sondhi, Stohastic theory of a data handling system with multiple sources, Bell Systems Technical Journal, vol.61, no.8, pp.10-18, 1974.

[30] T. E. Stern and A. I. Elwalid, Analysis of separable Markov-modulated rate models for information-handling systems, Advances in Applied Probability, 23:105-139, 1991.

[31] N. Mitrou, S. D. Vamvakos, K. P. Kontovasilis, Modelling, Parameter Assessment and Multiplexing Analysis of Bursty Sources with Hyper-Exponentially Distributed Bursts. Computer Networks and ISDN Systems 27: 1175-1192, 1995

[32] The ViC Tool, http://www-mice.cs.ucl.ac.uk/multimedia/software/vic/

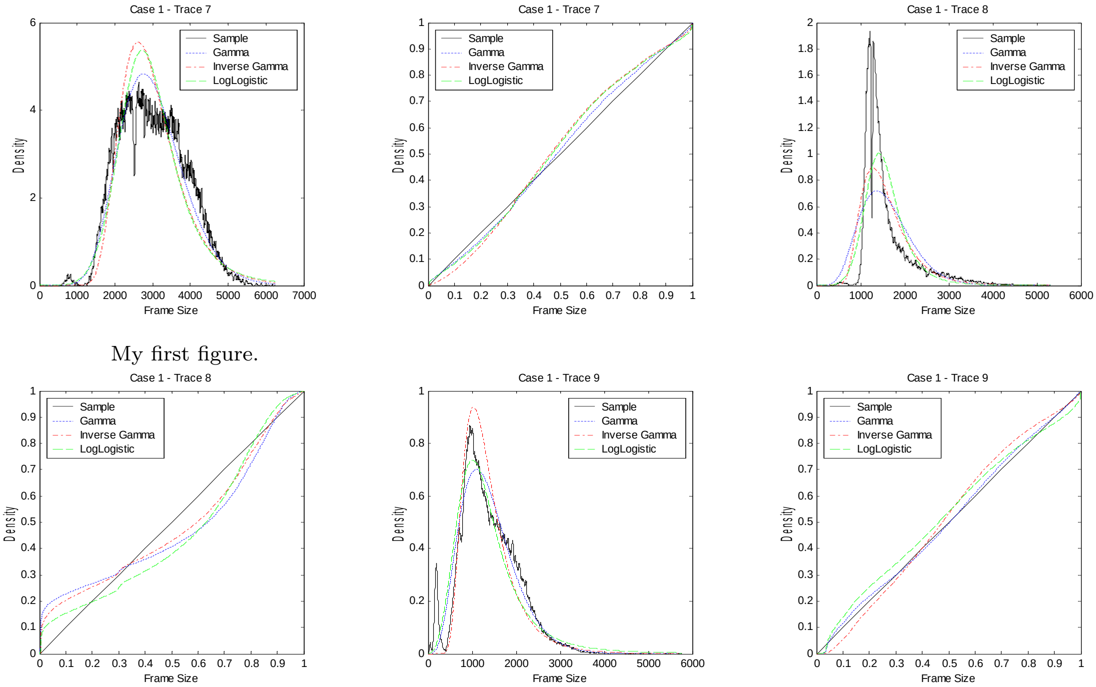
[33] VCON Vpoint HD, http://www.vcon.com

[34] France Telecom eConf, http://www.rd.francetelecom.com

[35] Sorenson EnVision, http://www.sorensonvrs.com

[36] OpenH323 Project, http://openh323.org

[37] MorningSound, http://www.soundmorning.com/



My first figure.

Figure 1: Frame-size histograms vs moment fit and the respective Q-Q Plots for unconstrained traces

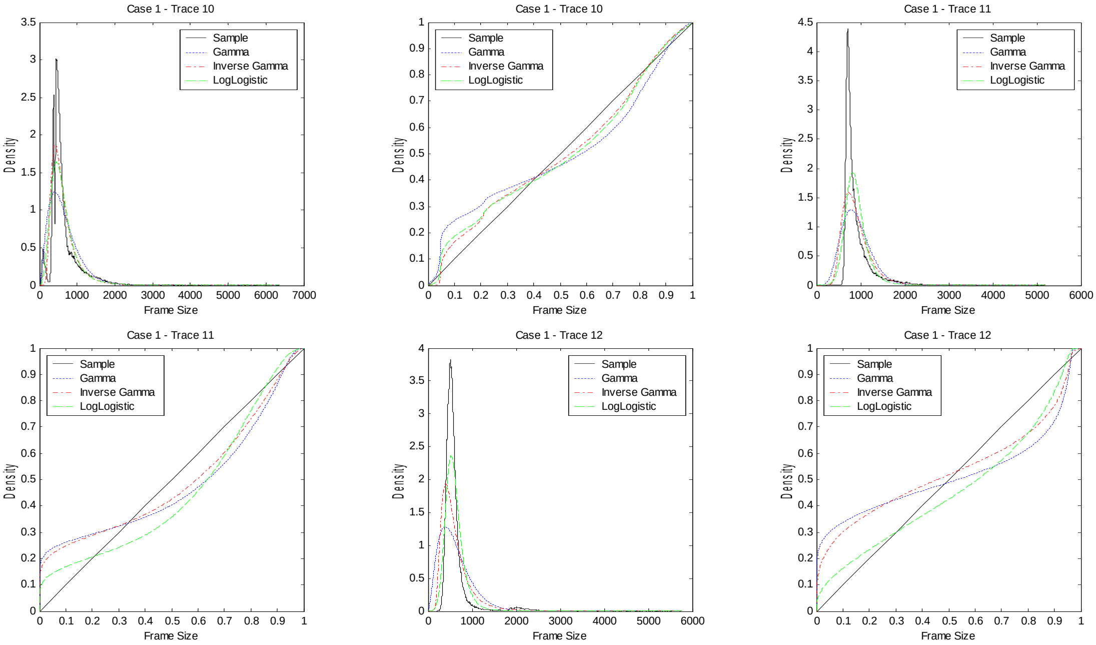


Figure 2: Frame-size histograms vs moment and C-LVMAX fit and respective Q-Q Plots for constrained traces

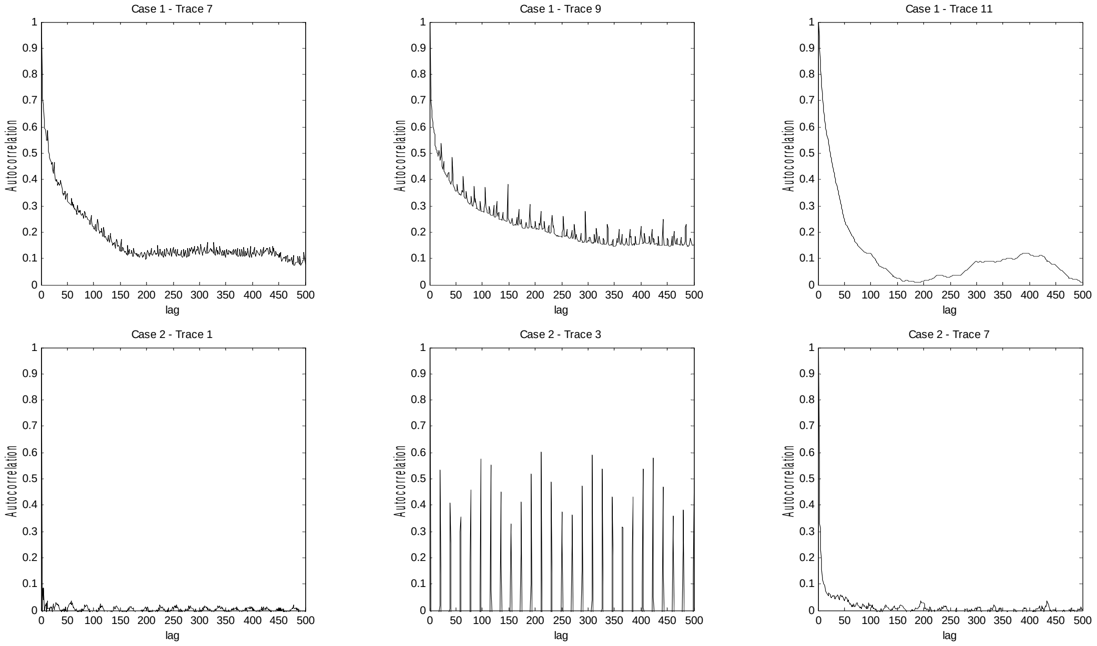


Figure 3: Autocorrelation Graphs for unconstrained and constrained traces

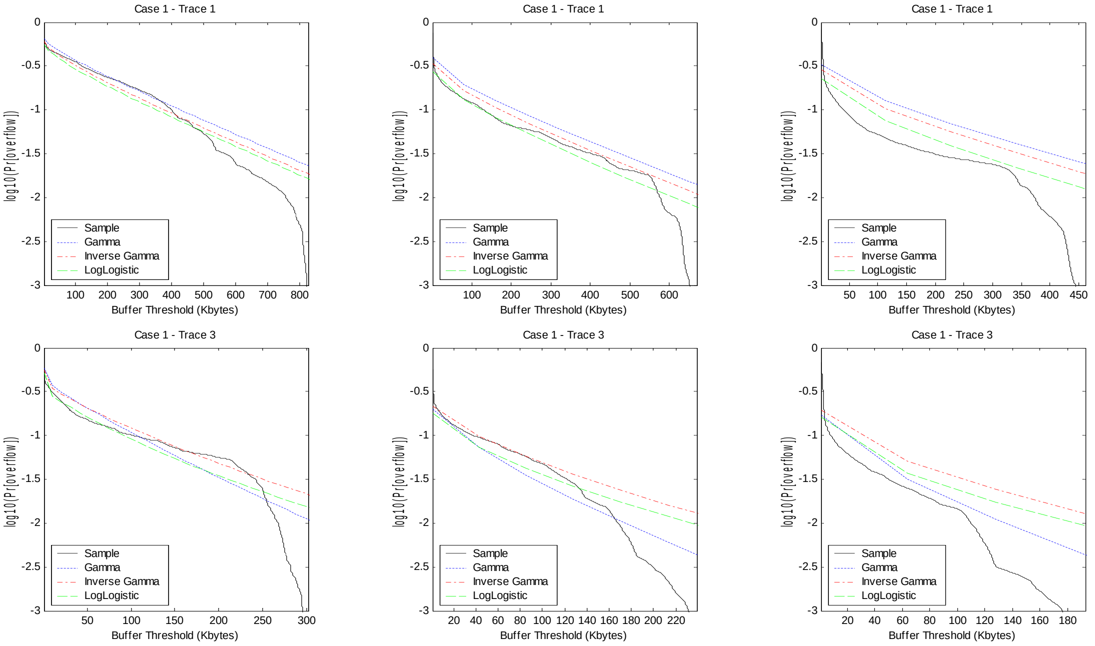


Figure 4: Complementary buffer overflow density plots of model vs sample

Table 1: Statistical quantities of the sample frame-size sequences

Case 1 (UNCONSTRAINED)										
Trace	Client	Encoder	Duration	Content	Target Rate	No Frames	Frame Rate	Rate	Mean	Variance
1	ViC	Intra H261	3600	VC-H	700	54006	15	111	921	249670
2				VC-L		53937	15	63	527	174130
3		H263		VC-H		54011	15	72	603	56981
4				VC-L		53453	15	54	457	24588
5		H263+		VC-H		53679	15	39	327	167780
6				VC-L		53633	15	27	224	205200
7	Vpoint	H261	1800	VC-H	700	27275	15	365	3009	731270
8				VC-L		27276	15	193	1592	347080
9	eConf	H263+	3600	VC-H	444	96805	27	298	1387	389300
10				VC-L		91710	25	130	640	147790
11	TKN	H263	2700	OFFICE		33825	13	91	904	107160
12			3600	LECTURE		45459	13	62	618	136850
Case 2 (CONSTRAINED)										
1	Vpoint	H261	3600	VC-H	155	49596	14	144	1306	48718
2				VC-L		50636	14	143	1275	29285
3	eConf	H263+		VC-H	82	51331	14	84	733	8267
4				VC-L		48848	14	67	619	22758
5	EnVision	H263	1800	VC-H	256	22501	13	239	2394	27275
6	TKN		2700	OFFICE	64	13800	5	64	1565	37316
7			3600	LECTURE		16788	5	64	1715	105530
8			2700	OFFICE	256	13936	5	256	6200	456790
9			3600	LECTURE		17707	5	256	6505	1540500

Table 2: Parameter values of the modeling components

	Gamma		Inverse Gamma		Loglogistic		C-LVMAX		Exponential		
Trace	α	β	α	β	m_l	σ	α	β	w	λ_1	λ_2
Case 1 (UNCONSTRAINED)											
1	3.3971	271.1005	5.3971	4049.5	6.648	0.3549			0.5057	0.9925	0.7757
2	1.5944	330.4804	3.5944	1367	5.9603	0.4449			0.5575	0.9905	0.8709
3	6.3878	94.4476	8.3878	4457.1	6.3382	0.1909			0.4368	0.9915	0.7528
4	8.5059	53.7656	10.5059	4347.3	6.0789	0.1609			0.3083	0.9906	0.8269
5	0.6385	512.6355	2.6385	536.2591	5.5827	0.3051			0.0683	0.992	0.5528
6	0.2454	914.5151	2.2454	279.4386	5.1082	0.318			0.0454	0.994	0.5493
7	12.3817	243.024	14.3817	40266	7.9663	0.1661			0.5322	0.991	0.8192
8	7.2998	218.0509	9.2998	13211	7.3199	0.1695			0.6018	0.9935	0.8781
9	4.9385	280.7675	6.9385	8234.1	7.1117	0.3038			0.5373	0.9932	0.7762
10	2.7722	230.8974	4.7722	2414.5	6.3161	0.3023			0.6022	0.9935	0.8059
11	7.6226	118.5656	9.6226	7792.9	6.7613	0.1533			0.7935	0.9785	0.8769
12	2.787	221.5946	4.787	2338.8	6.3424	0.193			0.7793	0.999	0.9316
Case 2 (CONSTRAINED)											
1	35.017	37.2996	37.017	47043	7.1602	0.0954	5.8689	236.9217			
2	55.4924	22.9723	57.4924	72016	7.1415	0.0752	9.1345	145.5206			
3	65.0183	11.2758	67.0183	48400	6.5901	0.0673	28.1127	25.8396			
4	16.8458	36.7556	18.8458	11050	6.3966	0.1456	2.7951	261.8669			
5	210.0779	22.7889	212.0779	1010500	8.4713	0.0388	294.3836	16.3024			
6	65.6089	23.8487	67.6089	104220	7.3455	0.0845	245.985	6.422			
7	27.8637	61.5424	29.8637	49496	7.4205	0.1407	187.5797	9.7124			
8	84.1382	73.6822	86.1382	527810	8.7237	0.0796	319.6409	19.6355			
9	27.4707	236.8108	29.4707	185210	8.7524	0.1455	249.7543	27.7889			