

Mixture Models for WWW Usage: Traffic Modelling for CATV Networks

Dee Denteneer & Verus Pronk

Philips Research
Prof. Holstlaan 4
5656 AA Eindhoven
the Netherlands

e-mail:

{dee.denteneer, verus.pronk}@philips.com

Abstract

Traffic modeling for HFC networks differs from traffic modeling for queueing systems in the greater emphasis on individual connections and on short-range dependencies. Mixture models form an appropriate class of models that take these characteristics into account whilst being, at least qualitatively, consistent with long-range dependence. These issues are illustrated by means of a model for WWW usage and a simulation of a CATV network, which is also known as a hybrid fiber coax (HFC) network, that serves a number of web clients.

1 Introduction

Tele-traffic modeling is currently one of the hot topics in the telecommunications world. The body of work devoted to this topic is vast and rapidly expanding. The models are firmly based in traffic measurements. These measurements establish that traffic on modern communication systems (e.g., an Ethernet or the Internet) differs in significant ways from assumptions about traffic that have been traditionally made in performance analysis.

With some simplification, one may say that the focus in this area is on models for *aggregate traffic* that exhibit *long-range dependence*. Interest in traffic due to a single source is secondary and is motivated by a search for a physical explanation of this long-range dependence: this explanation is often formulated by means of heavy-tailed distributions (for examples, see [17] or the papers in [1]). Here, long-range dependence means that time correlations between traffic loads exist for very long periods of time, i.e., they tail off hyperbolically instead of exponentially as with the classical short-range dependent models. It is this characteristic of the traffic that is

responsible for the excessive waiting times in queueing systems such as the Internet.

The relevance of these new models has become particularly clear from recent results in queueing analysis. In e.g. [6] it is shown that long-range dependence has an enormous impact on both waiting times and cell-loss probabilities in queueing systems. In the same vein, in [2], it is shown that waiting times in queues with heavy-tailed service times are considerably larger than waiting times in queues with light-tailed service times.

So, traffic models for queueing systems rightly stress the long-range dependence of the traffic, possibly neglecting short-range dependencies.

Our interest in traffic modeling, however, is due to cable networks that are currently being standardized (e.g. see [5] or [13]). In these networks, data transfer is more complex than in ordinary queueing systems. This is due to a sequential procedure for data transfer from a station at the customer premises to a central node, which consists of two stages. Firstly, a request stage is carried out, in which a station requests a number of data slots in contention with other stations. Secondly, a data transfer stage is carried out, in which the data is transferred in the data slots that have been reserved for this station. For a detailed description of data transfer in such cable networks see e.g. [12] or [7].

Traffic modeling for HFC networks is, in two respects, different from traffic modeling for traditional queueing systems. Firstly, in HFC networks, the short-range dependencies have a considerable impact on the efficiency of data transfer: request mechanisms, such as multiple requests and piggybacking, make it plausible that packets generated 'close to each other' are relevant for throughput and delay. Bursts of traffic can effectively be dealt with, as it is not necessary to go through the

request period for each packet within a burst. Secondly, traffic modeling for HFC networks is more directly geared to the traffic generated by individual sources than to traffic from an aggregate of users, because it is the single-user traffic itself that forms the basis for simulation.

The sensitivity of HFC performance to traffic characteristics has not yet been thoroughly investigated. However, studies that are already available indicate that correct traffic models are of great concern and that both long-range dependencies and short-range dependencies are relevant; [11] establishes the importance of long-range dependence in simulations using Ethernet traces. However, [9] establishes the additional relevance of short-range dependencies in simulations of HFC networks with artificial traffic, with either exponentially or Pareto distributed times between successive packets.

The central tenet of this paper is that *mixture models*, i.e., models based on mixture distributions, constitute a suitable class of models to describe network traffic due to a single source. In particular, the log time between successive packets in internet traces can be successfully described by means of a mixture of Gaussian distributions. With these models it is possible to describe the short-range dependencies in network traffic in a physically plausible way. This allows a view of the fine time structure of network traffic, which is relevant in our context and which has been largely ignored until now (see [15]).

Traffic due to aggregates of such models is not, in a formal sense, long-range dependent. However, by considering the variance plot evaluated for finite traces, we will show that such an aggregation has at least qualitative similarities to long-range dependent traffic.

This suitability of mixture models is illustrated in our analysis of web traffic. Web traffic will be split into two distinct components: the *user*-process and the *TCP*-process. In the user-process, events are directly generated by the user who browses the web. Each page request constitutes an event, and the modes of the mixture distribution refer to such user states as actively navigating, thinking, and having a break. The TCP-process concerns the events that are due to the protocols. Here, packet transmissions constitute the events, and the modes of the mixture models relate to networks characteristics, such as bandwidth and round-trip time.

Next, we apply these models in comparison with the traditional Poisson process to investigate the efficiency of data transfer in HFC networks. The results show the sensitivity of the simulation results to traffic characteristics; stressing the relevance of correct traffic models. Moreover, in-line with [9], they illustrate the fact

that short-range dependencies are beneficial for HFC networks.

The rest of the paper is organized as follows. In section 2, we give a brief introduction to mixture models and their relation to long-range dependence. The next section then presents a model for web usage. The section thereafter applies the results by simulating a typical HFC network that serves web clients. We end with some concluding remarks.

2 Mixture distributions

Mixture distributions are a popular tool to describe distributions that have more than one mode; for an extensive review, see [16]. In this paper, we will restrict ourselves to mixtures of Gaussians, or normal, distributions. Even such mixtures are extremely flexible; it can be shown that any distribution can be approximated arbitrarily well using such a mixture of Gaussians. We give a brief introduction to mixture models below, followed by a description of their relation to long-range dependence.

2.1 A brief introduction

Probability distributions are the standard tool for describing variation mathematically. A wealth of such distributions are available; the normal or Gaussian, the exponential, and the lognormal distributions are well known examples. In some applications, however, even this wealth does not suffice for an accurate description of variability. This is particularly the case if the quantity of interest depends on some unknown state, which in turn governs the variation.

Examples abound, also in traffic modeling. In traffic modeling for example, one can hypothesize two distinct states of the traffic on link: high and low load. Given the state of the traffic, the time between successive packets follows an exponential distribution, with either a high or a low intensity parameter. Not given this state, however, the traffic follows a distribution with two distinct modes, which is a mixture of the two original distributions.

The formal recipe to create mixture distributions closely follows this simple example: mixture distributions are weighted sums of standard distributions, where the weights sum to one. The weights are usually interpreted as probabilities of some unknown factor and describe the relative frequency with which this factor is in one of several states. Thus, in the example above, the weights describe the relative frequency with which traffic is high or low.

An example of a mixture distribution is given in figure 1, where a mixture of three Gaussians is displayed. In the graph, the first mode corresponds to the first element in the mixture. The second and third components are close together and combine to form the second mode in the graph, which is not symmetric, but skewed to the right. These two features illustrate the increased flexibility in modeling with mixture distributions. Firstly, they can be used to describe distributions that have multiple modes. Secondly, they allow for increased flexibility in the description of the shape of the individual modes.

In order to use a mixture distribution, one must choose the component distributions, the number of components, and the numerical values that characterize the component distributions. Given the component distributions and the number of components, it is possible to fit the mixture distribution to a given data set by estimating numerical values for the parameters of the distributions. There is a procedure available for doing this. It is usually simple and it is called the EM, expectation-maximization, algorithm. It is also possible to estimate the number of components rather than to consider this a fixed, known, quantity. In such cases, a penalty function, such as Akaike's information criterion or the Bayesian information criterion, is used to decide upon the number of components. These issues are thoroughly explored for web traffic in [4], to which we refer for the statistical details.

2.2 Relation to long-range dependence

There is currently wide spread interest in long-range dependence as it appears to be one of the few invariants of network traffic (e.g. see [15]). Hence, the relation of mixture models to long-range dependence is of relevance. First of all, it must be stressed that traffic generated by means of mixture models is not long-range dependent in any formal, mathematical sense. However, there is an interesting relation between these two concepts. Informally, traffic models that give rise to long-range dependence are bursty at *all* time scales, including those that approach infinity. The mixture models display burstiness at a finite number of time scales and can hence not be long-range dependent. However, a finite traffic trace from such a mixture model may be bursty on all time scales that are present in the trace and so be similar to long-range dependent traffic.

We now make this informal observation more precise. To this end assume that traffic is condensed to a series of counts: we observe a sequence of counts, where each count represents the number of packets that appear on a link in some small time interval. The variability of such a sequence of counts can be expressed through its

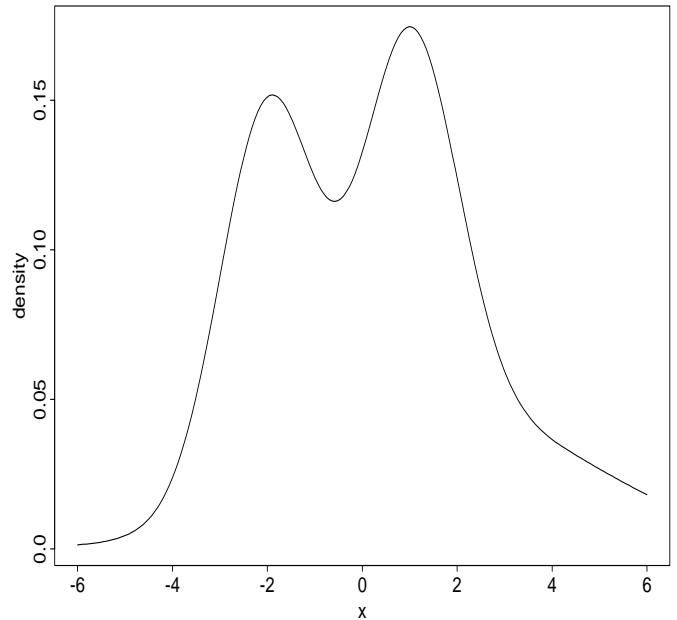


Figure 1: Mixture of three Gaussian distributions, with means $(-2, 1, 2)$, standard deviations $(1,2,3)$, and weights $(.33, .33, .33)$

variance: the average squared deviation from the mean. This sequence of counts can be further reduced by averaging the counts in m successive intervals to obtain a new, reduced series of counts. It is well known that the variance of the reduced series of counts equals $1/m$ times the variance of the original series of counts, if the counts in successive intervals are independent. Also, this reduction in variance by the factor $1/m$ holds for large m , if these counts are short-range dependent. For long-range dependent traffic, however, averaging will reduce the variance only by a factor m^r , with $r > -1$.

This observation leads to a popular graphical technique to discover long-range dependence: the variance plot, see e.g. [17] and references therein. In this technique, one plots the log of the block size (m) against the log of the variance of the averaged series for various values of m . If the slope of this plot equals -1 , overall or for large m , one concludes to independence or short-range dependence. If the slope tails off as $r > -1$, one concludes that long-range dependence is present.

We now apply this technique to two series of artificial data. The first series is constructed from an aggregate of 100 Poisson sources. Here, the individual processes are such that the time between successive events follows an exponential distribution with $\lambda = 2$. The aggregate traf-

fic is converted to a series of 180000 counts, by counting the number of packets due to these sources in time intervals of .01 seconds during half an hour. The variance plot for this sequence is displayed in figure 2 and is denoted by the symbols "e". This variance plot shows a decay with a slope equal to -1, that characterizes independent traffic.

Another sequence was constructed by simulating an aggregate of traffic due to 100 identical mixture sources. Here, the time between successive events is not derived from an exponential distribution but from a mixture of lognormals. The parameters of the lognormals were chosen so that overall traffic intensity matches the intensity of the Poisson sources above. Again, the aggregate traffic due to these sources was converted to a sequence of 180000 counts. The variance plot for this sequence is also displayed in Figure 2 and is denoted by the symbols "m". This aggregate of mixtures is not long-range dependent. However, the decay in the variance plot appears to be linear with a slope that is distinctly above -1, so is similar to long-range dependent traffic in this important respect.

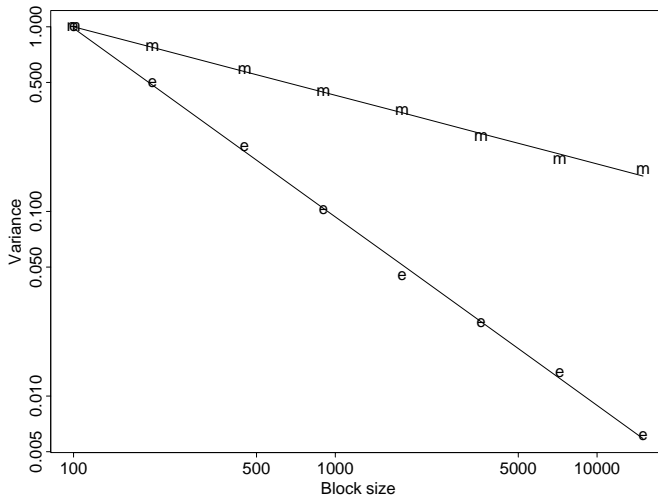


Figure 2: Variance plot for aggregate traffic due to 100 Poisson sources (e) and aggregate traffic due to 100 mixture sources (m).

3 A model for web usage

In this section, we outline a model for web usage by means of mixture models. Basically, such a model has

the following ingredients:

- The times at which packets are generated by the web client. This can equivalently be described by means of the inter-arrival times: the time between successive packets.
- The size of these packets.

The model then consists of the probability distribution functions that describe the variability of both inter-arrival times and packet sizes, and a description of the correlation between these. The familiar Poisson process is an example of a candidate model which substitutes an exponential distribution for the inter-arrival times, substitutes a constant distribution for the packet sizes, and assumes independence of all quantities.

However, the Poisson process falls short of our goals. Actual web client traffic is much more variable than a realization of a Poisson process, and artificial traffic can easily be distinguished from actual traffic.

To understand the deficiencies of the Poisson process, observe that web client traffic is governed by two processes:

The user-process This is the process as perceived by users who are browsing the web. A web browser requests a succession of web-pages, e.g., by clicking with the mouse.

The TCP-process This is the actual packet exchange that goes on between the web client and the web server. After a user requests a page, one or several (such as when a requested page contains several images) TCP connections are opened. The web client traffic in each connection consists of an *open connection*, an *information request*, a series of *acknowledgements*, and a *close connection*.

Each of these two processes has its own time scale, as users typically act much slower than computers. Hence, the existence of two time scales makes it untrue that just one, uni-modal probability distribution function will suffice to describe the time between successive events.

Now this argument can be extended. Again, the TCP-process does not consist of a homogeneous generation of packets with identically distributed inter-arrival times. Rather, the traffic at this level consists of *flights* of packets. The time between packets in the same flight is largely determined by the speed with which a packet is handled by a computer; the time between successive flights is determined by the round-trip time of the network.

In addition, the user-process will not be homogeneous in time as a user does not request pages at a constant

rate. Rather, he will alternate between various states and these states can be characterized, e.g., as *actively navigating*, *thinking*, and *having a break*. The time between successive page requests will depend on the user's state. So, in order to describe the inter-arrival time at the user level, one probability distribution function is required for each such possible state, and mixture distributions are appropriate.

These states can be argued about theoretically, but they can also be observed in measurements: see Figures 4 and 3. They are both based on publicly available data, that can be found at the Internet traffic archives [10]. The TCP-process will be analysed using the LBL-TCP packet traces, see [14]. For the user-process, we have used the UCB home IP usage study. Both data sets contain time stamps of all events, due to a community of users, on a given network connection. The analysis of both data sets was performed along similar lines:

- Disaggregation of the data into a number of traces due to individual source destination pairs, as in [17].
- Fitting a mixture of Gaussian distributions to the log time between events, in each of these traces.
- Comparison of the results, i.e. comparisons of the number of component distributions and the numerical values for the parameters of these distributions for all these traces.

In the following section, we summarize the conclusions, first for the TCP-process, next for the user-process.

As for the TCP-process, figure 3 shows the histogram of the log inter-arrival times of the packets in one of the individual traces, many traces being similar to the one shown. The multi-modality makes it clear that there are distinct time scales that play a role, and that mixture distributions are appropriate. Here, the reason for multiple time scales as explained above, is motivated by the nature of the protocols, and should cause no surprise. The mixture distributions are a convenient tool for extracting the information relevant to each of the time scales in each the individual traces.

Two further remarks concern the models for the TCP-process that we derive from these fitted mixture distributions. Firstly, the sequence of times between successive packets does not form an independent sequence. This is ignored when using mixture distributions. However, in building the actual simulation models, autocorrelations are incorporated using knowledge about the protocols, rather than using statistics. Secondly, the fact that the time between flights of packets corresponds roughly to the round trip time can be exploited: the round trip time can be shortened (artificially) so that faster networks or caching-techniques can be investigated.

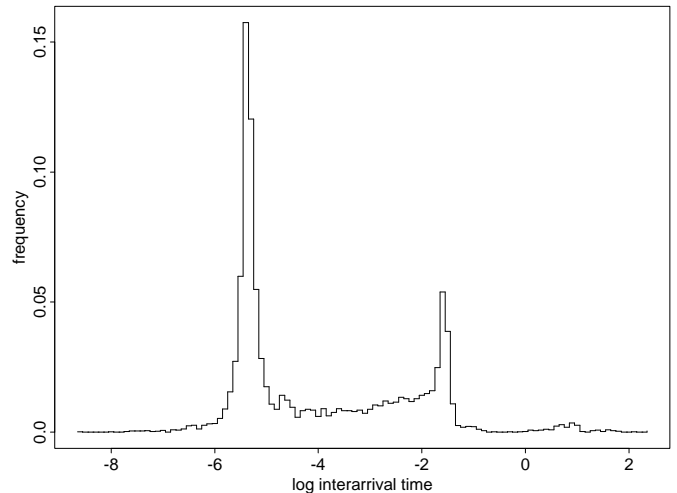


Figure 3: Histogram of the log inter-arrival times of successive packets in the TCP-process, based on one trace from LBL-TCP-3. The multi-modality of the histogram reflects the various time scales of the TCP-process.

Next, we turn to the user-process. Figure 4 displays the histogram of the inter-arrival times for page requests, as observed in a collection of 100 individual traces extracted from the UCB home IP usage study. Clearly, the histogram has several modes. Each mode reflects one of the time scales of the user process and each mode can be labelled with one of the user states.

A closer analysis was performed by fitting separate mixture models to each of the individual traces. It was found that each of these traces could be fitted well with mixture distributions with between two and five components, and that the majority fit well using only three components. In addition, these three component distributions were present in most of the traces, making them an appropriate candidate for the list of invariants of web-traffic. These components correspond to the following time scales:

- .15 sec.** : multiple page requests, such as when a web document requires other documents.
- 1.5 sec.** : actively navigating from a small web page.
- 15. sec.** : actively navigating from a large web page.

These three components represent the bulk of the time between events in these traces. An occasional component, present in only a few of the traces, was found at the very small time scale of .0025 seconds. This was always the time between requests for various images from

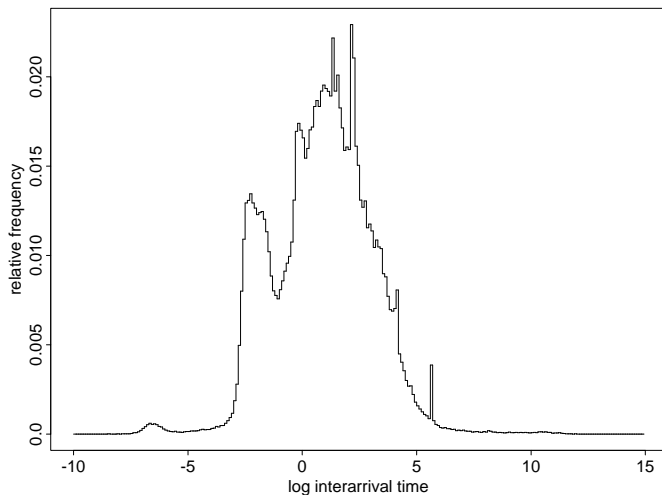


Figure 4: Histogram of the log inter-arrival times of successive page requests in the user-process, based on a set of traces, selected from the UCB home IP usage study. The multi-modality of the histogram reflects the various time scales of the user-process.

a requested web page. Larger components at time scales around 150 seconds were also incidentally encountered, representing downloading of larger web pages or having a break.

A final remark concerns the numerical values, such as the time scales above, that must go into every model used for traffic simulations. We are well aware of the ever changing nature of network traffic (see [15]), and that no absolute importance can be attached to these values. Having said this, however, it is still so that traces within a given data set show considerable similarity. E.g. in case of the home IP usage study, we extracted 100 traces corresponding to the users that generated most of the traffic, and 100 traces from randomly chosen users. Comparison of the distribution of the time between events for the heavy users versus the randomly chosen users revealed no significant difference, see [4].

To this we add that the ever changing nature of network characteristics necessitates frequent 're-calibration' necessary of the numerical values that go into the distributions. The attractiveness of an approach based on mixture distributions is enhanced by the fact that estimation of these numerical values can be automated with relative ease: re-fitting the models to new data is straightforward and can be done routinely as new measurements are being performed.

4 Application to HFC networks

The sensitivity of the performance of HFC networks to traffic characteristics has been noted in [11] and [9]. They establish that both long-range dependencies and short-range dependencies play a role. As to the relevance of long-range dependence, in [11], the authors compare simulations with actually observed Ethernet traffic traces to simulations with artificial traffic, that was obtained by time-permuting these observed traces. Here, the traffic used in these two simulations is identical in one respect: the same data values are used in the simulations. However, the traffic streams differ in their time structure: the observed Ethernet traffic is long-range dependent, whereas the permuted traces are independent. Therefore, differences obtained in these simulations can be attributed to this time dependency. Their simulations show that the performance of an HFC network (measured in terms of average transmission delay) in the case of the actual Ethernet traffic is much worse than the performance in the case of the artificial traffic. They conclude that the correlation structure also plays an important role for HFC networks.

In [9], the authors investigate the relevance of short-range dependencies: they compare the performance of HFC networks for traffic streams with exponential inter-arrival times to the performance for traffic streams with Pareto inter-arrival times. Again, the traffic assumption (exponential or Pareto) has an enormous impact on the outcomes of the simulations; the delays in simulations with Pareto inter-arrival times compare favorably with the delays in the simulations with exponential inter-arrival times. In this section, we extend the investigation into the impact of short-range dependencies on network performance.

We consider an HFC network that is largely compliant with the forthcoming Digital Video Broadcast (DVB) and Digital Audio Visual Council (DAVIC) standard, with a single upstream and a single downstream channel and a transmission capacity of approx. 3 and 30 Mbit/s, respectively. Within this context, we compare the delay as experienced by upstream traffic resulting from a number of Poisson sources with the delay as experienced by a number of web clients, simulated by means of mixture models as outlined in Section 3.

In the network, the transmission of application data upstream from the application in the homes to a central point, called the head end (HE), is governed by a request-grant mechanism. Requests for transmission slots are sent by the applications to the HE in contention, and the resulting grants guarantee contention-free transmission of the application data in the allocated slots. Contention resolution is done using a ternary, blocked tree algorithm

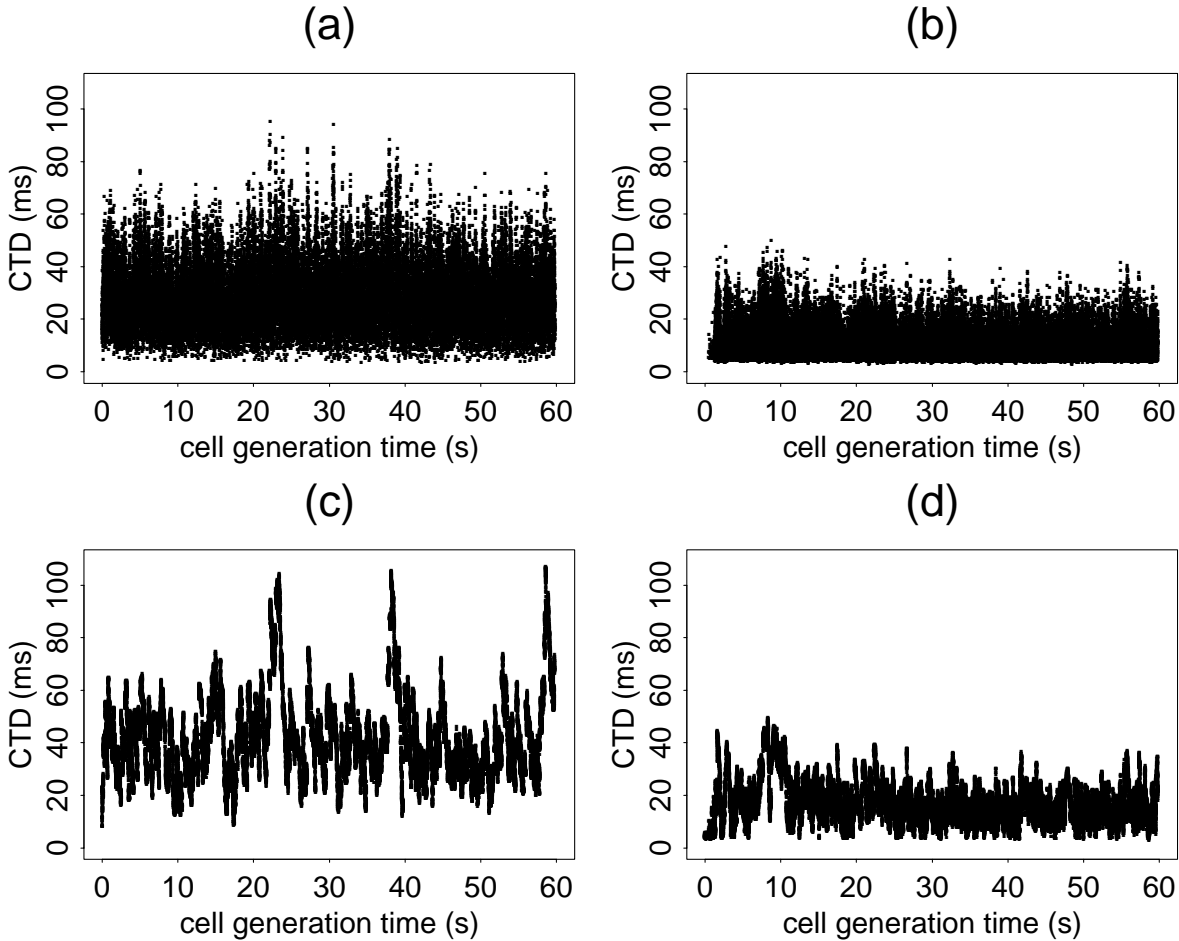


Figure 5: Time series of cell transmission delays (CTDs) (a) equivalent Poisson client traffic, (b) web-client traffic, (c) bulk Poisson traffic that accompanies equivalent Poisson client traffic, and (d) bulk traffic that accompanies web-client traffic.

(see [3]) When a request has to be retransmitted, the request size in terms of the number of requested slots may be updated to reflect current need.

As a single web client only produces a moderate amount of upstream traffic, i.e., in the order of 1 or 2 ATM cells per second on average, a bulk Poisson source, generating single ATM cells, is used to consume the bulk of the upstream bandwidth. The remaining bandwidth is (partly) consumed either by a number of web clients or an equal number of *equivalent* Poisson clients. An equivalent Poisson client generates on average the same amount of ATM cells as a web client, but with exponentially distributed inter-arrival times. In this way, pairs of corresponding simulations were carried out.

Table 1 lists the simulation settings. Note that the simulation corresponds to a network condition with heavy traffic load: the system has to transfer about 5100

ATM cells of application data per second, on a maximum of 6000 cells per second.

<i>simulation</i>	<i>sources</i>	<i>cell rate</i>
Web client	1 bulk Poisson	4400/s
	400 web clients	1.72/s
Poisson	1 bulk Poisson	4400/s
	400 eq. Poisson	1.72/s

Table 1: Simulation settings. Note that the web-client cell rates are time averages observed in simulations and these deviate from the long-term average of 1.76/s.

Figure 5 illustrates time series of individual cell trans-

mission delays (CTDs) for the bulk Poisson traffic, web client traffic and equivalent Poisson traffic relating to the simulations.

The figure shows a number of notable differences. Firstly, the CTD of the aggregate web-client traffic is significantly lower on average. Secondly, the bulk Poisson traffic CTD that accompanies the web clients is lower and less variable than the bulk traffic that accompanies the Poisson sources.

However, what is most striking is that the differences are caused by a change in only a relatively small portion of the total traffic.

The large differences in mean CTD are to be put down to the influence of the more bursty behavior of the web clients primarily on the contention resolution process. This bursty behaviour generally causes successive web-client cells to be generated by a relatively small number of web clients as compared to the uncorrelated generation of successive cells by the equivalent Poisson clients. As a result, fewer web clients with larger requests will contend in each tree, causing less delay in getting the requests to the scheduler.

5 Conclusions

In this paper, we have illustrated the importance of accurate traffic modeling for HFC networks. Only using Poisson processes only to describe traffic does not give a clear picture of HFC network performance. There is a clear need for application-specific traffic models for an accurate prediction of QoS versus load in service scenario studies. Most notable in this context is the need to also consider short-range dependencies in traffic, as well as single sources, as they significantly influence the contention-resolution process in HFC networks. Models based on mixture distributions form an appropriate class of models for doing this.

References

- [1] Adler, R., R. Feldman, and M.S. Taqqu [1998], *A practical guide to heavy tails. Statistical techniques for analyzing heavy-tailed distributions*, Birkhäuser, Basel.
- [2] Boxma, O.J., and J.W. Cohen [1998], The M/G/1 queue with heavy-tailed service time distribution, *IEEE Journal on Selected Areas in Communication*, Vol. 16, No. 5, pp. 749-763.
- [3] Capetanakis, J.I. [1979], Tree algorithms for packet broadcast channels, *IEEE Transaction on Information Theory*, Vol. 25, No. 5, pp. 505-515.
- [4] Denteneer, D. [1999], Random versus heavy browsers: toward a model for WWW usage, in H. Friedl, A. Bergold, G. Kauerman (eds), *Statistical Modelling*, pp. 165-172.
- [5] DVB [1999], Digital Video Broadcasting, DVB Interaction channel for Cable TV distribution systems (CATV), working draft (version 2), March 12, 1999, based on European Telecommunications Standard ETS 300 800 (Mar, 1998).
- [6] Erramilli, A., O. Narayan, and W. Willinger [1996], Experimental queueing analysis with long-range dependent packet traffic, *IEEE/ACM Transactions on Networking*, Vol. 4, No. 2, pp. 209-223.
- [7] Golmie, N., S. Masson, G. Pieris, and D.H. Su [1997], A MAC protocol for HFC networks: Design issues and performance evaluation, *Computer Communications*, vol. 20, pp. 1042-1050.
- [8] Golmie, N., Y. Saintillan, and D.H. Su [1999], A Review of contention resolution algorithms, *IEEE Communication surveys*, first quarter, pp. 2-12.
- [9] Harpantidou, Z., and M. Paterakis [1998], Random multiple access of broadcast channels with Pareto distributed packet inter-arrival times, *IEEE Personal Communications*, April, pp. 48-55.
- [10] <http://www.acm.org/sigcomm/ITA/>
- [11] Ivanovich, M.V., and M. Zukerman [1998], Evaluation of priority and scheduling schemes for an IEEE 802.14 MAC protocol loaded by real traffic, *Proc. Infocom '98/97*.
- [12] Limb, J.O. and D. Sala [1997], A protocol for efficient transfer of data over hybrid Fiber/Coax Systems, *IEEE/ACM Transactions on Networking*, vol. 5, pp. 872-881.
- [13] MCNS Holdings [1999], Data-over-Cable Service Interface Specification, Radio frequency Interface Specification, Ref. Nr. SP-RFIV1.1-I01-990311.
- [14] Paxson, V., and S. Floyd [1995], Wide-area traffic: the failure of Poisson modeling, *IEEE/ACM Transactions on Networking*, Vol. 3, No. 2, pp. 226-244.
- [15] Paxson, V., and S. Floyd [1997], Why we don't know how to simulate the Internet, Proc. of the 1997 Winter Simulation Conference, Atlanta, GA.
- [16] Titterington, D.M., A.F.M. Smith, and U.E. Makov [1985], *Statistical analysis of finite mixture distributions*, Wiley, New York.
- [17] Willinger, W., M. Taqqu, R. Sherman, and D. Wilson [1997], Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level (extended version), *IEEE/ACM Transactions on Networking*, vol. 5, No. 1, pp. 71-86.