ELSEVIER

Computer Communications 25 (2002) 1094–1102

computer communications

www.elsevier.com/locate/comcom

Analytical modeling of primary and secondary load as induced by video applications using UDP/IP

B.E. Wolfinger*, M. Zaddach, K.D. Heidtmann, G. Bai[1]

Department of Computer Science, Telecommunications and Computer Networks Division (TKRN), Hamburg University, Vogt-Kölln-Straße 30, D-22527 Hamburg, Germany

Received 1 August 2001; revised 30 October 2001; accepted 28 November 2001

Abstract

A particular challenge, when trying to analyze and predict the behavior of subnetworks of the global Internet, refers to the task of elaborating a sufficiently realistic workload characterization. In particular, it is necessary to specify (work)load at different system interfaces. This paper presents a generalized proceeding for load modeling, which can be used to formally describe sequences of (communication) requests at well-defined interfaces within a network. At first, the basic proceeding is applied by way of example to the modeling of primary load, i.e. load at an interface close to end-users, whereby we focus on video sources. We then tackle the challenging problem of characterizing secondary load, i.e. load as it is occurring at a lower layer interface within a protocol/service hierarchy, and for this purpose, we suggest a new approach for analytical modeling of load transformations as they are typical for communication networks. The broad applicability and the high validity of our approach to model load transformations is exemplified by means of a comprehensive case study assuming video sources and considering some load transformation corresponding to the impact within a RTP/UDP/IP protocol hierarchy.

Keywords: Workload characterization; Load transformation; Traffic engineering; Internet; Video communication

1. Workload characterization and load modeling for computer networks

When modeling service-systems, such as computers, communication networks, database systems, etc. it is a common practice to clearly separate the requests to be processed/served from the service-system properly which serves the requests. The set of requests to be served over time is denoted as load or workload of the service-system. If we apply this view to computer systems the requests to be served may correspond, e.g. to user programs to be executed, and in database systems the transactions to be processed may represent the load. In computer networks [17] files, e-mails, audio or video streams, WWW pages, etc. could represent the load if we consider an application-oriented interface, whereas at an interface to a packet-switched network the packets to be transmitted would constitute the load which is handed over to the network. It is well known that a sufficiently realistic load characterization, in most cases, is an indispensable prerequisite in order to obtain valid results when predicting the expected system behavior under a given load (e.g. by means of modeling or measurement studies) [5].

In the case of computer networks analyses, realistic load characterization is required, e.g. when applying analytical models or when executing simulation experiments. Moreover, load characterization is needed for the construction of artificial load generators, which e.g. may generate some synthetical load for an existing communication network. Using such load generators network behavior, under various load conditions, can then be investigated by means of measurements.

As we will explain in Section 2 in some more detail, load characterization on one hand has to specify the single requests which in their totality represent the overall load and on the other hand, it has to specify the stream of requests (arrival process) at a well-defined interface of the service-system considered. Depending on the desired range of use of load characterization very different degrees of freedom may exist for this characterization.

In this paper as a revised version of Ref. [19] we want to tackle the difficult problem of load characterization for the Internet. In particular, our contribution will introduce a flexible method for characterizing load at very different

* Corresponding author.
E-mail address: wolfinger@informatik.uni-hamburg.de (B.E. Wolfinger).
[1] Presently with Department of Computer Science, University of Calgary.

0140-3664/02/$ - see front matter
PII: S0140-3664(02)00023-3

interfaces within a UDP/IP-based protocol hierarchy. Our new method of load characterization will be based on the analytical modeling of load transformations. To demonstrate its practical relevance, we will apply the method by way of example.

Section 2 summarizes some of the features and properties of the Internet which make characterization of load for this network and its users extremely difficult. Moreover, we indicate the state-of-the-art in load measurements and load modeling of the Internet. We apply a generalized proceeding for load modeling as suggested in Ref. [17] during a first case study (Section 3) to characterize load as it could occur at an application-oriented interface within an Internet host computer. In the study, by way of example, we will model video streams starting from detailed load measurements and assuming standards for video encoding, such as H.261/H.263 (for MPEG cf. Refs. [3,4,18–21]).

Our view is that load at an application-oriented interface, which we call *primary load* (PL), is transformed by parts of the communication system into a different load, called *secondary load* (SL), which we can observe at a lower layer interface within the given protocol hierarchy. The topic of load transformation and an innovative method for its analytical modeling will be discussed in the context of Internet (cf. Section 4). In Section 5, our second case study will illustrate our new approach to model load transformation processes. We will demonstrate these transformations for the RTP/UDP/IP stack because of its importance for real-time applications. Thus, we will be able to characterize in a highly realistic manner the SL (IP traffic at an LLC interface) as it would be induced by a set of video traffic sources corresponding to the PL, e.g. within a video server. Our load transformation approach will be successfully validated in Section 6 by comparing measured SL (as observed in an Internet subsystem) with SL as it is predicted after transforming a given PL in the modeling domain.

2. Special aspects of modeling Internet load

In this paper, we are going to use the following definition of load, cf. Refs. [3,11,14,17].

The (*offered*) *load* or *workload* $L = L(E, S, IF, T)$ denotes the total sequence of requests which is offered by an environment E to a service-system S via a well-defined interface IF during the time-interval T. We call L the load generated by E for S at IF during T. Let us shortly discuss the strong dependencies of L on E, S, IF and T for the case of a computer network:

- E: all the requests to be served by S are created within the environment which, in particular, comprises the set of (human) network users as well as the (distributed) applications.
- S: as the service-system is responsible for serving the requests originated by E, the characterization of requests has to specify among others, the resource requirements of each request during its processing/service by S.
- IF: the interface chosen is extremely important as it reflects the decomposition of the computer network and its users into E and S; IF also directly determines the type of requests which can be part of the workload and, moreover, it limits the set of possible sequences of requests.
- T: evidently, the load observed in an existing network is highly dependent on the choice of T.

In the following, we want to focus on load characterization for the Internet. In Ref. [19], e.g. we debated the question why load characterization for the Internet is so much more difficult than characterizing load in networks like corporate networks or conventional LANs.

As a direct consequence of our definition of load, load characterization always assumes a well-defined interface. Unfortunately this is quite often not taken into account in existing publications. In the context of Internet we could, in particular, choose the following interfaces for load characterization: an application-oriented interface (e.g. interface to services/protocols such as FTP, Telnet, HTTP, SMTP,…) [2,8,12]; interface to the transport services, based on TCP or UDP, within the endsystems [22]; the packet interface to IP; or the LLC interface, e.g. in an Ethernet-based Intranet.

For most of the Internet interfaces mentioned, a large number of publications exist presenting load measurements for these interfaces. In particular, load measurements for application-oriented interfaces in the Internet have been presented in Refs. [2,22], measurements for the transport layer interface cf. Ref. [15] covering TCP and [22] covering UDP. Load measurements referring to IP interface have been summarized in Refs. [6,7,10].

Load measurements for specific interfaces in communication networks can be used to look for stochastic processes which are able to reflect, with sufficiently good accuracy, the main characteristics of the arrival process observed. In order to supplement the existing approaches to approximately measured Internet load by means of stochastic processes, we argue for a more general proceeding which is not restricted to mathematical modeling but allows us some detailed load characterization and load modeling for simulation models and artificial load generators of IP-based networks, too.

Our approach to load characterization and modeling will be presented in detail in the following sections and it will be applied in the context of the Internet. The approach comprises a generalized proceeding for load modeling directly based on load measurements. We suggest tackling the problem of load modeling for communication networks starting with modeling of the PL, as it exists at an application-oriented interface. We start with modeling the PL because this allows us a straightforward modeling of SL in a rather flexible way (e.g. for various choices of SL interfaces). Quite often PL can be observed in a relatively simple and direct manner and load at such application-oriented

interfaces typically is created quite independently of the communication network's state. Moreover, a complex PL can be conceived as an (mutually independent) overlay of elementary single sources of PL (e.g. single MPEG source in video communications, single file transfer using FTP, transmission of a sequence of PCM samples resulting of a single voice source, etc). Once we have solved the problem of characterizing the single sources of PL (for various types of sources, cf. examples of source models in the context of ATM networks as introduced in Ref. [16]), by means of overlaying single sources, we can produce mixes of complex PL. Thereafter, we can apply our approach for load transformation (cf. Section 4), in order to obtain a realistic characterization of the SL which exists at some arbitrarily chosen lower layer interface and which is induced by the complex mix of PL.

For this innovative approach to characterize SL, of course, we have to assume some sufficiently detailed knowledge regarding the process of load transformation as it occurs in the communication network and, moreover, that the load transformation is not too strongly dependent on the network's state. One of the main advantages of our approach to SL characterization is that it is not necessary to measure at the SL interface. Therefore, this method of load characterization can also be applied during the design of an innovative communication network under the assumption that the kind of load transformation in the newly designed network is known sufficiently precisely and that PL will be created in the future network in the same or in a similar way as in the present network. Another important advantage is that it is not necessary to create a possibly very complex mix of PL in an existing network; it is sufficient to consider this mix of PL in the modeling domain.

Our subsequent modeling of primary traffic loads for the Internet will be based on the generalized procedure proposed in Refs. [3,11]. The main purpose of this procedure is to present a unified description technique which allows us to formulate models of load (mainly for simulation experiments) for different degrees of detail in modeling and for various kinds of system interfaces. In particular, we want to cover load, which reflects requirement of communication resources.

For a presentation of our generalized load modeling method, we refer the reader to Ref. [17]. Quite comprehensive experiences in applying the modeling method have been reported by Bai [3]. A load description language based on extended finite automata has been elaborated by Kim [11] and a load specification technique based on extended Petri nets, is presented in Ref. [14].

3. Modeling of Internet traffic at an interface of primary load in video communications

In video communications via packet-switched networks, e.g. the following two classes of applications can be distinguished, video conferences and video-on-demand (VoD) services. While both types of applications do have different constraints referring to quality and real-time context, there are several similarities. In video conferencing, the video sequence as collected by a camera at discrete time instants leads to an isochronous stream of data units (uncompressed video frames) over time. The video frames are passed to a video encoder for compression at a frequency of ω frames/s. To achieve encoding in real-time the encoder has to execute compression of one frame in less than $1/\omega$ s. The video frames are passed to a transport system for RTP-based transmission again as an isochronous stream. The delay ξ (between provisioning of the uncompressed and transfer of the compressed frame) of the transport system is strongly determined by the speed of the video encoder. The class of VoD applications has only to transmit already encoded, i.e. compressed streams. In this case, the delay ξ could be significantly higher, but in order to realize high QoS, the time instances related to frame transmissions at sender and receiver should be synchronized as well.

In the following, we want to model PL as it is generated in video communications at an interface close to the video source. In particular, we want to observe the compressed video stream at the interface (IF_P) between application-oriented services and the transport system. As we are going to strictly base load modeling on load measurements, we have to collect measurements at IF_P. The arrival process of the video streams at IF_P is very regular because of its isochronous nature. As is usual in modeling video load [13], in the following, we assume that length of frames is the only attribute of interest for the requests observed at IF_P. Therefore, we have to measure the length x_i (in bytes) of the ith video frame being passed via IF_P at time $t_i = t_0 + (i/\omega)$ s, if t_0 denotes the start of the observation interval. The trace of frame lengths $X = \{x_i | i = 1, 2, \ldots, n\}$ describes the load and leads to the empirical distribution function $\mathcal{H}(s) = 1/n \sum_{i=1}^{n} \mathbf{1}_{\{x_i \le s\}}(s)$, $s \in \mathbb{R}$, where $\mathbf{1}_{\Omega}$ denotes the indicator function for the set $\Omega \subseteq \mathbb{R}$.

We carried out comprehensive load measurements [3,21] based on well-established standards for video encoding, such as H.261, H.263 and MPEG, in order to obtain results of general interest. The series of experiments referred to in this section cover 52 different video sequences, varying the quantization levels from 1 to 18. We exemplify the results by discussing one series of experiments in some more detail, in particular choosing the widespread reference sequence Claire. The sequences, we have analyzed, lead to the hypothesis that lengths of frames can be closely approximated by a normal distribution as illustrated, e.g. in Fig. 1 as well as in our publications [19–21]. The empirical mean $\hat{a} = 1/n \sum_{i=1}^{n} x_i$ and estimated variance $\hat{\sigma}^2 = 1/n \sum_{i=1}^{n} (x_i - \hat{a})^2$ were determined by the maximum likelihood method.

In order to quantitatively judge the accuracy of the maximum likelihood estimates, by means of a χ^2-test, we tested the empirical distribution for normal distribution according

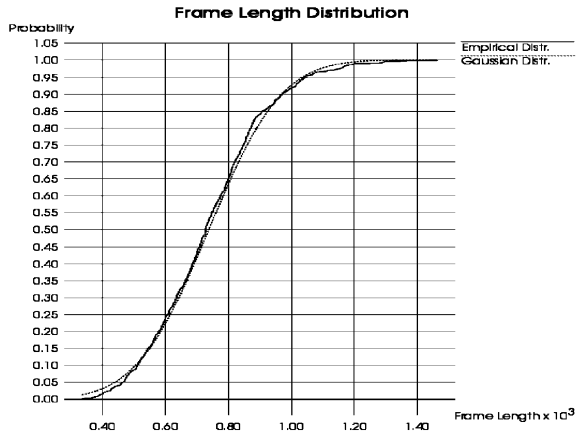

Fig. 1. Frame length distributions of the sequence Claire, quantization level 4, H.261 encoding.

to $\mathcal{N}(\hat{a}, \hat{\sigma}^2; x) = \Phi((x - \hat{a})/\hat{\sigma})$ and, in all cases, the results suggested to accept the hypothesis of normal distribution (details, cf. Refs. [18,20]). Thus, it seems acceptable to characterize the marginal distribution function of video frame lengths by approximate normal distributions. An important advantage of this approach results from the fact that the normal distribution is determined by only two parameters and it allows a straightforward derivation of quantiles and other statistical quantities.

Because of predictive encoding the frame length generating process could be strongly autocorrelated. We have carried out comprehensive measurements of the autocorrelation function $\hat{\rho}(\tau)$ (coefficient of correlation with lag τ) and we have observed [18–20] that the autocorrelation structure is based on long term dependencies, that it collapses, if these dependencies are eliminated, referring to short term measurements.

Our restriction to homogeneous video sequences and the strict separation between I- and P-frames evidently eliminates self-similar structures [20].

4. A new approach to model load transformations

In Section 3, by way of example, we have investigated PL as generated by video encoders. We can interpret the processing of data units within protocol layers as a process of transformation effective on the PL and producing the so-called SL. Let us denote components which transform load as (*load*) *transformers*. Load transformers change the properties of the load, e.g. in such a way that, on one hand, data units corresponding to the SL may become larger or smaller than those of the PL or that, on the other hand, the interarrival times of requests may be changed.

In communication systems a PL at some interface IF_P within the protocol hierarchy induces a SL at some lower layer interface IF_S. Characterization of SL in many cases is as important as or even more important than characterization of PL. In characterizing SL, as it would be induced by some given PL, the following two approaches can be distinguished: direct measurement of the (real) load at interface IF_S in an existing communication system or modeling of SL.

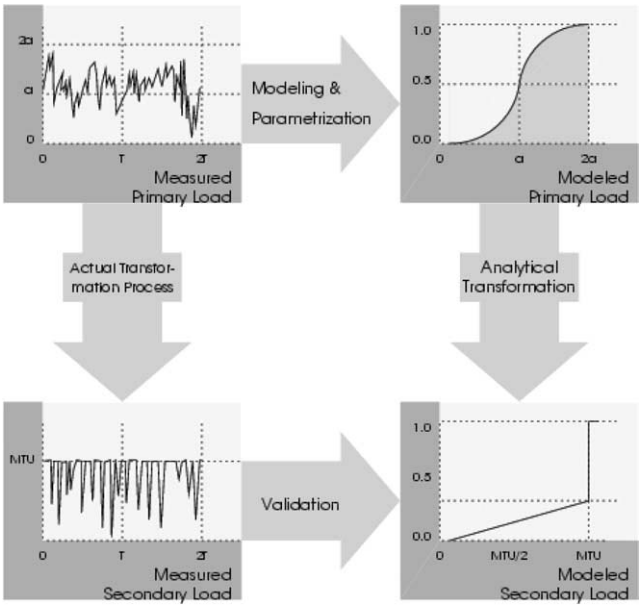

Fig. 2. Analytical modeling of load transformations.

In case of the first approach, we would have to generate the PL of interest in the real network and measure the SL which arrives at IF_S after having passed the real transformation process (e.g. the protocol processing). This approach is not feasible if the interface IF_S is not accessible for measurements in an existing network or during design or early development of a new communication system, when IF_S would not yet be implemented. Moreover, it could be necessary to investigate a SL as it would be induced by a very special mix of single sources on the level of PL and it could be impossible to generate this mix in the existing network.

In the second approach based on modeling, the influence of parts of a communication system on a given PL is reflected by a model. Here, the real transformation process is replaced by a so-called *artificial transformer* (transformation in the modeling domain). The purpose of an artificial transformer is to convert the attributes (and their values) of PL into those of SL as well as to transform the arrival process of PL requests into the one for SL requests. If the artificial transformer used is a sufficiently valid model of reality, we can obtain a realistic prognosis of SL to be expected.

Load characterization has to cover the specification of the arrival process of requests as well as the specification of the values of request attributes. The characterization can be deterministic if we use, e.g. some trace or it may be probabilistic if we use, e.g. some distribution to reflect the interarrival times and the attribute values of the requests generated over time. This implies the following levels of abstractions for the load: the actual load, its description as a

trace, or its probabilistic characterization by means of distributions.

Some measured load can be approximated by a distribution (e.g. to characterize lengths of data units) which may be directly used as a model of PL by an artificial transformer. The artificial transformer may then reflect the transformation process just by changing (recalculating) the given distribution into a new one to approximate the induced SL (cf. Section 5, for example). The validity of the predictions of the artificial transformer can be determined by means of comparisons with measured SL (cf. Fig. 2 for some graphical illustration of the proceeding).

In an earlier work, as opposed to this paper, we already investigated load transformation by means of simulation [3,4,11].

In this contribution, we want to advocate for analytical load transformation. As is common in modeling, in the special case of modeling load transformations too, analytical modeling offers the important advantages of only minor programming and calculation effort in evaluating the required formulas as well as leading to more comprehensive possibilities of model evaluations and result interpretations. Transformation processes reflecting the behavior of complete protocol hierarchies as they exist in nowadays networks are very complex as a consequence of all the various (layered) services they have to support. So, there does not seem to exist any hope that a sufficiently realistic analytical modeling of load transformation should be feasible. To solve this problem, nevertheless, we suggest to map the complete process of load transformation onto a sequence of elementary load transformations which take place one after the other and thus to achieve analytically tractable transformations.

Moreover, we distinguish two classes of elementary transformations. A first class of elementary load transformations only allows the modification of the interarrival times between requests. Examples of such transformations are algorithms for smoothing traffic such as *Leaky Bucket*-Algorithms. A second class of elementary transformations modifies request types and attributes but keeps constant the request interarrival times. To give an example for a transformation of the second class, generation of packet headers typically does not significantly change the packet interarrival times whereas it modifies the attribute *packet length*. Astonishingly, a large variety of transformations of the second class which are relevant in real networks are still analytically tractable in a very realistic way as we could prove in Ref. [20], namely transformations such as fragmentations of protocol data units, generation of CRC checksums, bit stuffing, some kinds of compression algorithms, etc.

And even transformations which at the same time modify interarrival times between requests and attributes may still be accessible by analytical models to reflect load transformations, if we conceptually separate transformation of timing from transformation of request attributes.

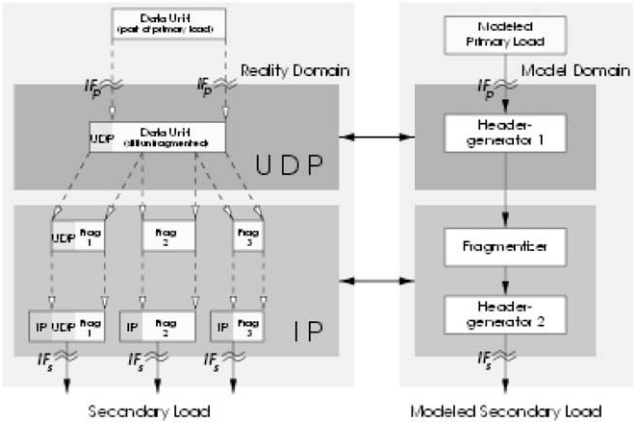

Fig. 3. Modeling SL using load transformers.

5. Modeling of IP traffic at an LLC-interface as an example of secondary load characterization in a video server

The general concept of load transformation (e.g. by protocol layers) as introduced in Section 4 will now be exemplified in looking at load transformations being typical for the Internet. As we want to make use of our results for PL modeling as presented in Section 3, in the following we assume video communication directly based on UDP (RTP) and IP protocols (cf. Fig. 3).

The interface IF_P which we choose to observe PL, as it is generated by the video sources, corresponds to the UDP transport service-access-point (TSAP). The interface IF_S chosen by us for SL observation and modeling corresponds to the interface between IP layer and the network adapter (i.e. Ethernet adapter in our case). The interface IF_S chosen is still sufficiently of high level in order to make results of load modeling not too much dependent on the network technology used and of the adapter's implementation (e.g. its buffer management). Let us shortly discuss the load transformation now, as it is effective between interface IF_P and IF_S. To keep the case study sufficiently simple let us assume that the PL characterization at IF_P refers to the arrival process of requests at IF_P. Only one type of request is to be considered, namely data units (e.g. the encoded video frames), which have to be transmitted via UDP, and so the data unit length is the only attribute of requests.

So, the load transformation resulting from (RTP)UDP/IP functionality implies a modified arrival process of requests (now IP packets) at IF_S. As the packet length at IF_S is the only attribute of interest, it is sufficient to focus on manipulations of data units within UDP/IP layers which have an impact on length. Length of data units is changed by UDP when adding the headers with UDP specific control information. On the Network layer, fragmentation by IP changes not only the length but also the number of data units. Maximum data unit length used for fragmentation results from an agreement between IP and network adapter, being kept fixed

during network operation. Evidently, after fragmentation, IP too adds IP specific control information to packets. Fig. 3 summarizes transformations which have an impact on lengths of data units being processed by UDP/IP. The figure also suggests the modeling of UDP/IP transformations by mapping these transformations (in the model domain) onto three elementary transformers, namely two Header-Generators and one Fragmentizer, which are placed in series.

In Section 3, we concluded that a single source of video traffic can be adequately characterized by the distribution of lengths of video frames which are generated according to the video display frequency used. Let $\mathscr{L}(x)$ be the probability distribution of data unit lengths which resulted during modeling of PL, observed at IF_P. In the following, we want to derive mathematically the impact which transformers of type *Header-Generator* and *Fragmentizer* have, in particular leading to new distributions of lengths.

Transformation of single requests by RTP and UDP implies that a header of $\nu = 12 + 8$ byte length is added to data units. Here also a checksum is calculated for the user data leading to a delay being proportional to data unit length. According to measurements (for a Pentium-166 PC under Linux) this delay varies between 50 and about 250 μs and thus is rather small compared to the interarrival times of video frames at IF_P of, e.g. 33 ms. So, in SL modeling, we will neglect the delay resulting of RTP and UDP processing.

The new distribution of data unit lengths $\mathscr{U}(x)$ which is a consequence of header generation by UDP is reflected by the equation $\mathscr{U}(x) = \int_{-\infty}^{x} \mathrm{d}\mathscr{L}(s-\nu) = \mathscr{L}(x-\nu)$. UDP datagrams are passed to IP which, by means of fragmentation, has to make sure that the maximum transmission unit (MTU)-length of the next lower layer, LLC (logical link control) is respected. Therefore, if a data unit handed over to IP has a length larger than the value MTU-length minus IP-header-length it will be fragmented. Measurements prove that the time for fragmentation can be neglected. If Θ denotes MTU-length in bytes (e.g. $\Theta = 1500$ byte for Ethernet) and ψ denotes IP-header-length in bytes (e.g. $\psi =$ 20 byte in case of IPv4 or $\psi = 40$ byte for IP-version 6) the maximum length of fragments ϑ can be calculated as $\vartheta = \Theta - \psi$ and so we can directly calculate the expected number of fragments β, a value which is of great significance in characterizing the burstiness of a traffic source. Let $\beta(n)$, $n = 1, 2, \ldots$ denote the probability that a UDP data unit is fragmented into exactly n segments, then $\beta(n) = \mathscr{U}(n\vartheta) - \mathscr{U}((n-1)\vartheta)$, which allows straightforward calculation of β, namely

$$\beta = \sum_{i=1}^{\infty} i\beta(i) = \sum_{i=1}^{\infty} i(\mathscr{U}(i\vartheta) - \mathscr{U}((i-1)\vartheta)). \tag{1}$$

As for every distribution F, we have $\lim_{x\to\infty} F(x) = 1$, it is evident that for every error bound $\epsilon > 0$ we find an $l \in \mathbb{N}$ with $l = \min_{k\in\mathbb{N}} \{\mathscr{U}(k\vartheta) \geq 1 - \epsilon\}$, so that the approximation $\mathscr{U}(k\vartheta) \to 1$ is valid for $k \geq l$. Thus, we obtain the following approximation for β:

$$\beta \approx l\mathscr{U}(l\vartheta) - \sum_{i=0}^{l-1} \mathscr{U}(i\vartheta) \approx \sum_{i=0}^{l-1} 1 - \mathscr{U}(i\vartheta). \tag{2}$$

Eq. (2) now allows us to calculate the distribution $\mathscr{F}(x)$ of lengths for fragments generated by IP:

$$\mathscr{F}(x) = \begin{cases} 0, & x \leq 0, \\ \dfrac{1}{\beta} \displaystyle\sum_{i=0}^{l-1} \mathscr{L}(x + i\vartheta - \nu) - \mathscr{L}(i\vartheta - \nu), & 0 < x < \vartheta, \\ 1, & x \geq \vartheta. \end{cases} \tag{3}$$

After fragmentation the IP header is created and added to the corresponding fragment. According to measurements on today's PCs the CPU processing time required for header creation by IP is slightly lower than 30 μs. This value is of interest as it strongly influences the packet interarrival times within bursts of packets, resulting from IP fragmentation, at the interface between IP and LLC layer, i.e. at interface IF_S in our load modeling example. Concerning lengths of IP packets at interface IF_S we obtain the corresponding distribution $\mathscr{I}(x)$ of lengths directly as $\mathscr{I}(x) = \mathscr{F}(x - \psi)$ which can be composed with Eq. (3).

Thus, we have successfully completed our search for the distribution of data unit lengths at the SL interface IF_S. Moreover, our results also cover characterization of the mutual dependencies between fragments, in particular, the probability that a fragment is followed by another one referring to the same UDP datagram is determined by the array $\vec{\beta} = (\beta_1, \beta_2, \ldots)$, as well as its expectation β. Our solution for calculating β directly (for a given distribution of data unit lengths at a PL interface) is of important practical relevance: among others dimensioning of resources, such as appropriate choice of buffer sizes, and model-based quality-of-service (QoS) management may be considerably supported by knowledge of β.

This transformation can be generalized in order to achieve valid results for an overlay of m single sources. In video communication this situation could correspond to a video server with load produced by m independent video sources S_i. Referring to Section 3 we could characterize the single sources by m load models $\mathscr{L}_i(x) = \Phi((x - \hat{a}_i)/\hat{\sigma}_i)$, $\forall i = 1, \ldots, m$, and assume requests (video frames) of source S_i being generated with periodicity T_i beginning at starting instant τ_i, i.e. generation of requests at instants $\tau_i, \tau_i + T_i, \tau_i + 2T_i, \ldots$ Let be $\mathbb{T}_i = \{\tau_i + nT_i | n \in \mathbb{N},\ n < N_{\max}\}$ the set of all observed arrival times of stream i. Thus, the relative proportion α_i of arrivals concerning the ith stream to the overall arrival process can be determined, i.e.

$$\alpha_i = |\mathbb{T}_i| \left(\sum_{j=1}^{m} |\mathbb{T}_j| \right)^{-1},$$

the relative proportion α_i^* of the departure process of the

complex SL is given by $\alpha_i^* = \alpha_i/\beta_i$, where β_i are determined by Eq. (1) for all streams $i = 1, \ldots, m$. So, Eq. (3) implies

$$\mathscr{I}(x) = \begin{cases} 0, & x \le \psi, \\ \sum_{i=0}^{l-1} \sum_{j=1}^{m} \alpha_j^* (\mathscr{L}_j(x + i\vartheta - \nu - \psi) - \mathscr{L}_j(i\vartheta - \nu - \psi)), & \psi < x < \vartheta + \psi, \\ 1, & x \ge \vartheta + \psi. \end{cases} \quad (4)$$

Eq. (4) now allows us to characterize SL (in terms of distribution of packet lengths) which is induced by a complex PL representing e.g. the overlay of single video sources in a video server. For more details regarding the derivation of our analytical formulae, cf. Refs. [19,20].

Among others, our results would allow us to directly use and easily parameterize a packet train model [9] (with deterministic intertrain- and intercar-times) as a realistic description of the SL to be expected.

6. Validation of our transformer approach in secondary load characterization by means of analytical modeling

We now want to validate the accuracy of our method for SL prediction based on application of a load transformer. To prepare the validation we start with measuring both, PL as generated by single video sources as well as the SL which is induced by PL and observed at the interface (IF_S) between IP and LLC layer. Measurements have been carried out for Pentium PCs (166 MHz, under Linux) assuming typical Internet MTU sizes of 576 and 1500 byte. The sources of PL active during the measurements have to be modeled to allow load transformation in the modeling domain. Each single source is mapped onto a load generator creating a sequence of requests (video frames to be transmitted) with a single attribute *length* based on the measurements of Section 3. To perform the transformation we just have to apply Eq. (4) of Section 5 to the given normal distribution. A χ^2-test is again used in order to validate the predicted distribution for SL with respect to the actual, measured lengths at IF_S.

To stress our analytical modeling approach for SL characterization, in the following we will assume a complex PL resulting from an overlay of m single sources (in particular: $m \in \{3, 30\}$). First, we consider $m = 3$, i.e. a video server communicating with three video clients. As video sequences we choose Claire, Foreman and Carphone. The server is sending sequence Claire on a quantization level of 10 ($\hat{a} \approx 301.52$, $\hat{\sigma} \approx 51.34$) with a video frame frequency of 12 frames/s, Carphone on a quantization level of 4 ($\hat{a} \approx 3071.75$, $\hat{\sigma} \approx 950.58$) with 15 frames/s and Foreman on a quantization level of 1 ($\hat{a} \approx 14,309.19$, $\hat{\sigma} \approx 2424.25$) with 30 frames/s. We calculated $\beta = \sum_{i=1}^{m} \alpha_i \beta_i$ according to our analytical model for mixed traffic load transformation and found $\beta \approx 3.0891$ (for MTU size of 1500 byte) and $\beta \approx 7.3922$ (for MTU size of 576 byte). Note that β is characterizing the burstiness of load at interface IF_S. Comparisons between empirical and analytically determined distributions are presented in Fig. 4.

Besides the excellent conformity between empirical and analytical distribution it is remarkable that the distribution function characterizing the pure fragments is nearly linear for MTU size = 576 byte and assuming exact linearity we can approximate the complete distribution function just by linear interpolation of $\{(\psi, 0.0), (\Theta, 1/\beta), (\Theta, 1.0)\}$.

The strong non-linearity in the empirical distribution function (for an MTU size of 1500 byte), cf. Fig. 4, for a value of the IP packet size of about 300 byte is a consequence of the videosubstream Claire which is part of the PL and remains completely unfragmented because of the already very small original video frames, as observed at IF_P. Even in this case, the analytically determined distribution still predicts highly precise results whereas the first order approximation leads to small deviations. Nevertheless, for practical purposes even the accuracy of the first order approximation should be acceptable in most cases, especially as the error in calculating β (according to our analytical model) is neglectable here again. Results of the χ^2-test lead to a value of 1.4767 (with MTU size of 1500 byte) and 0.9763 (with MTU size of 576 byte) for the distribution determined by analytical load transformation. For the first order approximation, χ^2-test provides values of 9.0654 (MTU size: 1500) and 3.8722 (MTU size: 576). These values clearly do not argue for rejection of the hypothesis

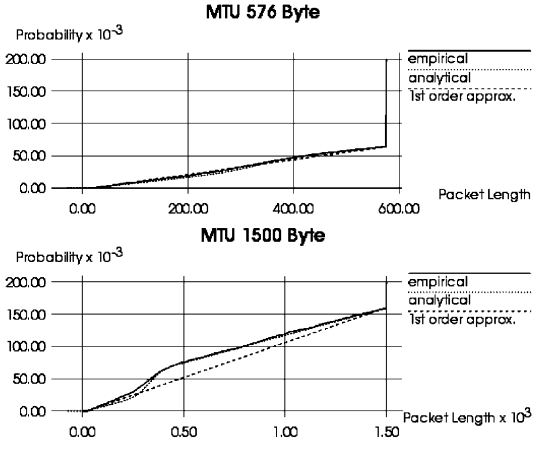

Fig. 4. Comparison between measurement results and analytical modeling of SL (overlay of three sequences, H.261 encoding).

that the empirical packet length distribution is adequately approximated by both distributions suggested (analytical and first order approximation).

Typically a video server simultaneously serves a large number of clients. In Ref. [18], we present a detailed study of a video server operating with an overlay of $m = 30$ sources. The validation leads to results as accurate as the ones presented here, and moreover, substantiates the hypothesis that an overlaying of a large number of single sources intensifies the effect of linearity in the distribution function of IP packet lengths up to the MTU size.

As a general conclusion of our numerous validation experiments, cf. Refs. [19–21], in all comparisons between the empirical packet length distribution (based on the actual measurements) and the mathematically predicted length distribution (based on our analytical load transformation model) we observed excellent agreement between both distributions. Predicted SL characterizations were found to remain valid even when substantially increasing the error introduced in PL modeling. This results from the fact that, quite often, already the precise prediction of the expected number of fragments allows one to produce a sufficiently realistic characterization of SL.

Though it might still have been expected that our approach for analytical load transformation is able to reflect the elementary transformations without significant loss of reality, it is quite surprising, however, that the sequencing of elementary transformations does not introduce inaccuracies and that, moreover, one can completely neglect the additional packets containing just control information, which may result from processing the offered load within RTP/UDP and IP layers.

7. Load transformation induced by TCP

The transformation of load as it is executed by the UDP/IP protocol stack can be modeled much more easily than in the case of TCP/IP, because TCP behavior is strongly dependent on the state of the network [8] (e.g. TCP's congestion control leads to a transformation of the packet arrival process which is strongly network state dependent, i.e. there exists some feedback between network and load generating environment). Thus, in case of TCP, load modeling can no longer be done by assuming an environment (comprising the TCP senders), which reacts independently of the underlying service provider (comprising the IP service). So, one could expect that analytical modeling of load transformation as induced by TCP is completely unfeasible. However, the situation is much better than that: luckily, transformation of request attributes (e.g. packet length) by TCP too, is still largely independent of the network's state, and therefore, can be modeled as in case of UDP (cf. modeling of segmentation and header generation as in Sections 5 and 6). Nevertheless, the problem remains of how to model the transformation of the packet arrival process as induced by TCP (i.e. how to model the transformation of timing behavior by TCP for a sequence of packets). At present, we see the following methods for modeling TCP's impact on packet interarrival times:

- One could have some knowledge on how TCP influences the packet timing during different states of the network (e.g. network overload versus low network utilization). In particular, if network utilization is low, analytical modeling of TCP's impact on packet timing still seems directly feasible).
- In case that analytical modeling of TCP's impact on timing is unfeasible, we still could characterize transformation of packet timing by TCP based on simulations. One should note that even if we rely on simulations the simulation experiments are simplified significantly because they just have to solve transformation of timing and not transformation of request attributes (for which the analytical approach still could be applied).

So, we can conclude that, even in case of TCP, the analytical approach to load transformation is strongly useful. The present limitations in covering the transformations regarding TCP's impact on packet timing analytically are not necessarily very serious, in particular, for those situations, where exact packet interarrival times at the IP interface are not very relevant. This could be the case in configurations, where we are interested in the packet arrival process at a SL interface IF_2 below IP layer and the arrival process of IP packets at TCP/IP interface (IF_1) is completely destroyed by some smoothing or access control process active before the packets enter at interface IF_2.

8. Summary and outlook

In this contribution, we have addressed the challenging problem of real-time workload characterization and modeling for complex computer networks such as the Internet. Though quite a few researchers share the opinion that the Internet cannot be modeled at all, our view is slightly different. We agree that the Internet in its totality seems indeed to be much too complex to be modeled but we claim that adequately chosen subsystems or special aspects of the Internet can still be modeled in a sufficiently realistic manner, at least if we choose an appropriate level of abstraction.

The focal point of our paper concerns characterization of SL. For this purpose, we suggested to start with some sufficiently valid characterization of PL and to map the transformation processes, which transform this primary into a SL, onto some analytical models. This new approach of investigating load transformation in a modeling domain seems to be much more flexible and should lead to more insight than characterizing SL in the conventional way, namely, approximating some load measurements collected in a real network. We were able to show as a result of our

comprehensive experiments that most of the transformation processes referring to *length* attribute are still tractable in a sufficiently realistic manner by analytical means.

On one hand, we have demonstrated how to analytically model load transformations in principle, and on the other hand, have also applied our modeling approach in comprehensive case studies. In order to obtain boundary conditions, which should be at the same time realistic and relevant from a practical point of view, we have used Intranet configurations as underlying communication networks and video applications (with standard compression algorithms) as examples of relevant real-time applications. Not only we were able to demonstrate a realistic characterization of PL for video sources, but we also could prove (by comparisons with real SL measurements) that our method to predict SL (by using analytical modeling of load transformations) leads to SL characterizations, which are astonishingly realistic and thus highly valid.

The realistic load characterization, which is enabled by our load modeling and load transformation approach, covering various interfaces within an IP-based protocol hierarchy can be used in a straightforward manner in combination with analytical or simulation models as well as a constituent of an experimental infrastructure for dedicated performance measurements regarding subsystems of the Internet. Limitations of our approach concern, e.g.

- the possibly high expenditure which may result in characterizing the large variety of single sources of PL in the Internet as well as the mutual dependencies which may exist between these sources of load;
- the complexity of some load transformation processes which may not be easily modeled in a sufficiently realistic manner, especially if the transformation would depend strongly on the network's state.

Some of the still open load modeling problems mentioned may only be very hardly – if at all – solvable for the Internet as a global network in its full complexity. Nevertheless, we hope that our load modeling approach and, in particular, our new method for analytical load transformation can and will be used to derive valid models for innovative communication networks including real-time oriented subsystems of the Internet, focusing on RTP and UDP. The application of such models would allow one to identify and possibly eliminate (some of the many) bottlenecks in parts of the Internet, to study—by means of modeling—the impact of changes in the Internet protocol stack (e.g. inclusion of protocols to support real-time communication or additional QoS management functionality) and, last not least, to investigate and predict the behavior of Internet subsystems for some of the load situations to be expected in the future.

References

[2] M.F. Arlitt, L. Williamson, Internet web servers: workload characterization and performance implications, IEEE/ACM Transactions on Networking 5 (5) (1997).

[3] G. Bai, Load Measurements and Modeling for Distributed Multimedia Applications in High-Speed Networks, Uni Press, Hochschulschriften, 1999.

[4] G. Bai, B.E. Wolfinger, Possibilities and Limitations in Smoothing MPEG-coded Video Streams: A Measurement based Investigation, MMB'97, VDE-Verlag, 1997 pp. 119–135.

[5] A.B. Downey, D.G. Feitelson, The elusive goal of workload characterization, ACM SIGMETRICS Performance Evaluation Review 26 (4) (1999) 14–29.

[6] R. Epsilon, J. Ke, C. Williamson, Analysis of ISP IP/ATM network traffic measurements, Performance Evaluation Review 27 (2) (1999) 15–24.

[7] A. Feldmann, A.C. Gilbert, P. Huang, W. Willinger, Dynamics of IP traffic: a study of the role of variability and the impact of control, ACM SIGCOMM'99 Conference, Computer Communication Review 29 (4) (1999) 301–313.

[8] G. Haring, Chr. Lindemann, M. Reiser (Eds.), Performance Evaluation: Origins and Directions, LNCS, vol. 1769, Springer, Berlin, 2000.

[9] R. Jain, S.A. Routhier, Packet trains—measurements and a new model for computer network traffic, IEEE Journal on Selective Areas in Communication SAC-4 (6) (1986) 986–995.

[10] J.L. Jerkins, J. Monroe, J.L. Wang, A measurement analysis of internet traffic over frame relay, Performance Evaluation Review 27 (2) (1999) 3–14.

[11] J.J. Kim, Formale Lastbeschreibung und eine Methode zur Lastmodellierung für innovative Kommunikationssysteme, Verlag Shaker, Reihe Informatik, Aachen, 1993.

[12] G. Kotsis, K. Krithivasan, S. Raghavan, A workload characterization methodology for WWW applications, Proceedings of the International Conference on the Performance and Management of Complex Communication Networks, Tsukuba, 1997.

[13] A.A. Lazar, G. Pacifici, D.E. Pendarakis, Modeling video sources for real-time scheduling, ACM Multimedia Systems 1 (1) (1994) 253–266.

[14] J. Magott, B.E. Wolfinger, Formal description technique to support load modelling for innovative communication systems, Applied Mathematics and Computer Science Journal 4 (4) (1994) 605–633.

[15] V. Paxson, Automated packet trace analysis of TCP implementations, ACM SIGCOMM'97 Conference, Computer Communication Review 27 (4) (1997) 167–179.

[16] G.D. Stamoulis, M.E. Anagnostu, A.D. Georgantas, Traffic source models for ATM networks: a survey, Computer Communications 17 (6) (1994) 428–438.

[17] B.E. Wolfinger, Characterization of mixed traffic load in service-integrated networks, Systems Science Journal 25 (2) (1999) 65–86.

[18] B.E. Wolfinger, M. Zaddach, K.-D. Heidtmann, G. Bai, Modeling of primary and secondary load in the Internet, Technical Report No. 227/00, Department of Computer Science, Hamburg University, 2000.

[19] B.E. Wolfinger, M. Zaddach, K.-D. Heidtmann, G. Bai, Analytical modeling of primary and secondary load as induced by video applications using UDP/IP, Proceedings of SPECTS'01, Orlando, Florida, 2001, pp. 276–286.

[20] M. Zaddach, Modellierung, Charakterisierung und Transformation von Videoverkehrslasten, PhD Thesis, Department of Computer Science, Hamburg University, 2001; published also by: Shaker Verlag, Aachen, Germany, 2001.

[21] M. Zaddach, K.-D. Heidtmann, Measurement and traffic characterization of H.26x-coded video streams, Proceedings of MMB'01, Aachen, Germany, 2001, pp. 89–102.

[22] W. Zhu, Characterizing wide area conversations on the Internet, MSc Thesis, Department of Computer Science, University of Saskatchewan, Canada, 1994.