Peakedness Characterization in Teletraffic

S. Molnár, Gy. Miklós
High Speed Networks Laboratory, Dept. of Telecommunications and Telematics, Technical University of Budapest
H–1111, Sztoczek u. 2, Budapest, Hungary Tel: +36 1 463 3889,
Fax: +36 1 463 3107, Email: {molnar,miklos}@ttt-atm.ttt.bme.hu

Abstract

The bursty nature of traffic over many time scales is one of the most challenging characteristics of high speed networks. In this paper we deal with the generalized peakedness as a promising candidate measure of this poorly understood phenomenon. An extension of the framework of the theory of generalized peakedness in discrete time with the applications for the most important traffic models are developed and the results are demonstrated in the paper. A new model fitting technique is also given in this framework with examples. Finally, the engineering aspects of the measurement of peakedness and applications for various real traffic (MPEG video, aggregated ATM, Ethernet) are presented.

1 INTRODUCTION

An important experience from recent measurement studies (including Ethernet, ATM LAN/WAN networks [7, 14, 16]) regarding the nature of broadband traffic is that traffic exhibits bursty properties over many time scales.

One of the key concepts for capturing the bursty character of traffic is self-similarity which resulted in active research on fractal characterization [7, 14]. So far it is not clear how successfully we can utilise self-similarity from a practical traffic engineering point of view but one thing is for sure: burstiness seems to be the most important yet poorly understood characteristic of traffic in high-speed networks. Our work is motivated by this need. In this paper we focus on peakedness as one of the most promising candidate measures of traffic burstiness.

The simplest burstiness measures take only the first-order properties of the traffic into account. A set of candidates are the moments of the inter-arrival time distribution. In practice the peak to mean ratio and the squared coefficient of variation are the most frequently used first-order measures [13, 15].

Measures expressing second-order properties of the traffic are more complex. The autocorrelation function, the indices of dispersion [4, 18] and the generalized peakedness [2, 3] are the most well known measures from this class.

Moreover, there are a number of burstiness measures based on different concepts, e.g. we can use burst length measures [15, 19] or parameters of a leaky bucket for burstiness characterization [12]. By the concept of self-similarity the Hurst parameter and other fractal parameters are also candidates for burstiness measures [7, 14].

In this paper we review the theory of generalized peakedness and further develop the basic concept by introducing the generalized peakedness in discrete time. The advantage of this approach is that it allows us to apply the general framework of peakedness for traffic engineering. We provide the computation of peakedness for a number of important discrete time models including the Markov modulated batch Bernoulli process and the batch renewal process. The relationship between IDC and peakedness is also presented. We discuss the challenges of measuring peakedness in practice. Moreover, we show a technique how Markov modulated traffic models can be fitted to a measured peakedness curve. Finally, the practical applicability of peakedness and our modeling technique are demonstrated by examples based on measured MPEG video, aggregated ATM and Ethernet traffic.

2 PEAKEDNESS MEASURES

Peakedness of a traffic stream has been found a useful characterization tool in blocking approximations and in trunking theory [5]. It has been defined as the variance to mean ratio of the number of busy servers in an infinite hypothetical group of servers to which the traffic is offered, where the service times of the servers are independent and exponentially distributed with a common parameter.

2.1 Generalized peakedness

Eckberg [2] extended this definition by allowing arbitrary service time distribution and defined *generalized peakedness* as a functional which maps holding time distributions into peakedness values. For a given complementary holding time distribution $F^c(x) = \mathrm{P}\{\text{holding time} > x\}$, Eckberg defines the peakedness functional $z\{F^c\}$ as the variance to mean ratio of the number of busy servers in a hypothetical infinite group of servers with independent holding times distributed according to F^c. The general definition provides a way to characterize the variability of an arrival stream with respect to a given service system.

Let us have a stationary arrival process S in continuous time with counting function $N(t)$ = the number of arrivals in $(0, t]$ for $t \geq 0$. The mean arrival intensity is denoted by $m = \mathrm{E}\{N(t)\}/t$, which is independent of t due to the stationarity of S.

Arrivals are allowed to come in batches of random size B. We define the batchiness parameter as $b = \mathrm{E}\left\{B^2\right\}/\mathrm{E}\left\{B\right\}$ which can be shown to be the mean size of a batch that an arbitrary arrival finds itself in. The differential process [1] $\Delta N(t)$ is defined for a fixed Δt as the number of arrivals in $(t, t + \Delta t]$, that is, $N(t + \Delta t) - N(t)$. We define the covariance density of the arrival process $k(s)$ for $s > 0$ as the covariance of the differential process as Δt goes to zero: $k(s) = \lim_{\Delta t \to 0} \frac{\mathrm{Cov}\{\Delta N(t), \Delta N(t+s)\}}{(\Delta t)^2}$ which is independent of t due to the stationarity of S. For $s < 0$ we let $k(s) = k(-s)$.

We offer the arrival process S to an infinite server group where the service times are independent and have a complementary holding time distribution of $F^c(x)$ ($x \geq 0$; for $x < 0$, we define $F^c(x) = 0$), mean holding time of $1/\mu = \int_{-\infty}^{\infty} F^c(x)dx$ where μ is the service rate, and finally the autocorrelation of F^c is $\rho_{F^c}(x) = \int_{-\infty}^{\infty} F^c(s)F^c(s + x)ds$.

Denoting the number of busy servers at time t by $L(t)$, the generalized peakedness functional is defined as

$$z\{F^c\} = \frac{\mathrm{Var}\left\{L(t)\right\}}{\mathrm{E}\left\{L(t)\right\}}. \tag{1}$$

If the arrival stream is defined for the whole time axis $(-\infty, \infty)$, it is independent of t due to the stationarity of S. In practice, we never have an arrival process for an infinitely long time; in this case, we have to define the peakedness for a t which is large enough for the initial transient period in the service system to be negligible. (More precisely, $z\{F^c\} = \lim_{t\to\infty} \mathrm{Var}\left\{L(t)\right\}/\mathrm{E}\left\{L(t)\right\}$.)

With the notation introduced above, the peakedness of the arrival stream can be expressed in terms of the covariance density function as [2]

$$z\{F^c\} = 1 + \frac{\mu}{m}\int_{-\infty}^{\infty} (k(s) - m\delta(s))\rho_{F^c}(s)ds \tag{2}$$

where $\delta(s)$ is the Dirac delta function.

The important case of exponential service time simplifies to

$$z_{\exp}(\mu) = \frac{b+1}{2} + \frac{1}{m}k^*(\mu) \tag{3}$$

where $k^*(\mu) = \int_{0+}^{\infty} k(s)e^{-\mu s}ds$, the Laplace transform of the covariance density function. Here we have the peakedness of a given arrival stream as a function of the service rate μ.

It is shown [2] (and is suggested by eq. (3)) that the peakedness function $z_{\exp}(\mu)$ together with m determines $k(s)$ and therefore the pair $(z_{\exp}(\mu), m)$ is a complete second order characterization of the arrival process.

The peakedness function $z_{\exp}(\mu)$ can be used to compute the peakedness functional for a large class of holding time distributions as shown in [2]. The method is elaborated in [11] to give the peakedness functional for Coxian holding time distributions. The importance of Coxian holding times lies in the fact that any holding time distribution can be approximated with arbitrary accuracy by Coxian distributions. Eckberg also investigated the application of generalized peakedness in delay systems [3]. Eckberg's definition of generalized peakedness for point processes has been extended in [8, 9] to allow fluid flow models given by a rate function.

2.2 Peakedness in discrete time

In order to use the peakedness measures in a B-ISDN framework, we now extend the peakedness concept for discrete time arrival streams.

We use the following notation: $w[i]$ is the number of arrivals at epoch i, where $i = \ldots -1, 0, 1, \ldots$. We assume the stationarity of $w[i]$. The first and second moments of $w[t]$ (independent of t) are denoted by m_1 and m_2. The covariance density of continuous time is replaced here by the autocovariance function $k[s] = \mathrm{Cov}\,\{w[i], w[i+s]\} = k[-s]$. (It is seen that $k[0] = m_2 - m_1^2$.)

The service time random variable T is also discrete and has the distribution $t[1], t[2], \ldots$ on positive integers. (It cannot take on zero value.) $\mu = 1/\mathrm{E}\,\{T\}$ is again the service rate, and it is easily shown that $1/\mu = \mathrm{E}\,\{T\} = \sum_{s=-\infty}^{\infty} F^c[s]$ where $F^c[x]$ is the complementary holding time distribution function: $F^c[x] = \sum_{u=x+1}^{\infty} t[u] = \mathrm{P}\,\{T > x\}$ if $x \geq 0$ and $F^c[x] = 0$ if $x < 0$. The autocorrelation function is now $\rho_{F^c}[x] = \sum_{s=-\infty}^{\infty} F^c[s] F^c[s+x]$. It is seen that $\rho_{F^c}[0] = \sum_{s=-\infty}^{\infty} (F^c)^2[s]$.

The traffic is offered to an infinite group of servers with independent identically distributed service times determined by $F^c[x]$. Each arrival takes a separate server. The peakedness of the arrival stream is defined as the variance to mean ratio of the number of busy servers in the infinite server group:

$$z\{F^c\} = \frac{\mathrm{Var}\,\{L[t]\}}{\mathrm{E}\,\{L[t]\}} \tag{4}$$

where $L[t]$ is the number of busy servers at time epoch t.

An important modification of the definition is to let the service time depend on the arrival epoch only (have a common service time for all $w[t]$ arrivals at epoch t). We call (in accordance with [9]) the peakedness value defined in this way the *modified* peakedness $\tilde{z}\{F^c\}$. As we have shown [10],

$$\tilde{z}\{F^c\} - z\{F^c\} = \left(\frac{m_2}{m_1} - 1\right)(1 - \rho_{F^c}[0]\mu). \tag{5}$$

that is, their difference is constant (cf. (35) in [9]). The first factor in the difference is zero if and only if the arrival stream has no simultaneous arrivals, the second factor is zero if and only if the holding time distribution is deterministic.

The importance of this modified definition lies in the fact that it gives a way to handle a whole batch of arrivals together, which can save a lot of computational effort in the case of measuring the peakedness for a general holding time distribution. However, in the case of geometric service times, the original definition of peakedness is easier to measure as shown in section 3.1. We will use the original definition of peakedness (eq. (4)) below.

We can express peakedness in terms of the autocovariance function $k[s]$ similarly to eq. (2) as

$$z\{F^c\} = 1 + \frac{\mu}{m_1} \sum_{s=-\infty}^{\infty} \rho_{F^c}[s](k[s] - m_1\delta[s]). \tag{6}$$

The most important case in discrete time is the case of geometrically distributed holding times: $t[i] = \mu(1-\mu)^{i-1}$, $0 < \mu < 1$ (with $\mathrm{E}\{T\} = 1/\mu$ which justifies the notation).

In order to simplify the formulas, let us introduce the notation

$$K[s] = \begin{cases} \frac{2}{m_1}k[s] & \text{if } s > 0 \\ \frac{1}{m_1}k[0] & \text{if } s = 0 \end{cases}$$

and let its z-transform be $K^*(\omega) = \sum_{s=0}^{\infty} K[s]\omega^s$.

The peakedness function of the arrival stream with respect to geometric holding time distribution, as we derived in [10], is given by

$$z_{\text{geo}}(\mu) = 1 + \frac{K^*(1-\mu) - 1}{2-\mu} \tag{7}$$

2.3 Peakedness and IDC

The widely used measure to characterize the variability of an arrival stream on different time scales is the index of dispersion for counts (IDC). It is defined as $I[t] = \frac{V[t]}{E[t]} = \frac{V[t]}{m_1 t}$ where $E[t]$ and $V[t]$ are the mean and variance of the number of arrivals in t consecutive epochs ($t = 1, 2, \ldots$).

The connection of IDC and peakedness for geometric holding times is, as we have shown [10]

$$z_{\text{geo}}(\mu) = 1 + \frac{\mu^2 \frac{d}{d\omega} I^*(\omega)|_{\omega=1-\mu} - 1}{2-\mu} \tag{8}$$

where $I^*(\omega)$ is the z-transform of $I[t]$.

We can use eq. (8) to get asymptotic results which connect them [10]:

$$z_{\text{geo}}(0) = \frac{\lim_{s\to\infty} I[s] + 1}{2}, \qquad z_{\text{geo}}(1) = I[1] = \frac{\mathrm{Var}\{w[i]\}}{\mathrm{E}\{w[i]\}} \tag{9}$$

2.4 Peakedness of traffic models

Next, we present the peakedness results for important traffic models. We consider discrete time models for the number of arrivals in consecutive epochs.

(a) Batch Bernoulli process

A very simple type of arrival stream model is the model with the number of arrivals in a time epoch be independent identically and generally distributed with mean m_1 and second moment m_2.

In this case, $k[i] = 0$ for all $i > 0$. Thus, $K^*(1-\mu) = K[0] = \frac{\mathrm{Var}\{w[i]\}}{\mathrm{E}\{w[i]\}}$ and $z_{\text{geo}}(\mu) = 1 + \frac{\frac{\mathrm{Var}\{w[i]\}}{\mathrm{E}\{w[i]\}} - 1}{2-\mu}$ For the special case of Poisson batch arrivals, the distribution of arrivals in an epoch is Poissonian, thus $\frac{\mathrm{Var}\{w[i]\}}{\mathrm{E}\{w[i]\}} = 1$ which gives $z_{\text{geo}}(\mu) = 1$.

The Poisson process can be considered as a reference process with respect to peakedness characterization. Batch arrival processes that are more bursty than the Poisson process have higher peakedness values, smoother processes have lower peakedness. (In the case of deterministic traffic, $z_{\text{geo}}(\mu) = 1 - \frac{1}{2-\mu}$.)

(b) Markov modulated batch Bernoulli process

A very general Markovian process is the Markov modulated batch Bernoulli process (MMBBP). In this model, we have a discrete time Markov process as a modulating process. In each state of the modulating Markov-process, batch arrivals are generated according to a general distribution corresponding to the state.

Let $\mathbf{P}$ and $\mathbf{D}$ denote the transition probability matrix and the steady-state distribution vector of the modulating Markov process, respectively ($\mathbf{DP}=\mathbf{D}$). Let $\mathbf{M_1}$ and $\mathbf{M_2}$ be diagonal matrices corresponding to the first and second moments of the number of arrivals in the corresponding states. Let $\mathbf{e}$ be a vector of all ones and let $\mathbf{I}$ be the identity matrix.

We can express the mean number of arrivals as $m_1 = \mathbf{DM_1e}$ and the second moment as $m_2 = \mathbf{DM_2e}$. The autocovariance function of the arrival process is given by $k(i) = \mathbf{DM_1P}^i\mathbf{M_1e} - m_1^2$.

Using eq. (7) we have derived the peakedness function as [10]

$$z_{\text{geo}}(\mu) = 1 + \frac{1}{2-\mu}\left(\frac{2(1-\mu)\mathbf{DM_1P}(\mathbf{I}-(1-\mu)\mathbf{P})^{-1}\mathbf{M_1e} + m_2}{m_1} - 1\right) - \frac{m_1}{\mu} \tag{10}$$

A very important case of MMBBP is the Markov modulated Bernoulli process (MMBP); its peakedness curve is the special case of eq. (10).

(c) Switched batch Bernoulli process

Another important special case of MMBBP is the 2-state MMBBP (SBBP, switched batch Bernoulli process). Let us use the following notation: the transition matrix is $\mathbf{P} = \begin{bmatrix} 1-\alpha_1 & \alpha_1 \\ \alpha_2 & 1-\alpha_2 \end{bmatrix}$ and the steady state distribution is thus $\mathbf{D} = \frac{1}{\alpha_1+\alpha_2}(\alpha_2\ \alpha_1)$.

Denote $\gamma = 1 - \alpha_1 - \alpha_2$. In state 1, the first and second moments of the number of arrivals are $m_{1,(1)}$ and $m_{1,(2)}$, respectively; in state 2, the moments are $m_{2,(1)}$ and $m_{2,(2)}$.

The first and second moments of the number of arrivals are given by $m_1 = \frac{1}{\alpha_1+\alpha_2}(\alpha_2 m_{1,(1)} + \alpha_1 m_{2,(1)})$, $m_2 = \frac{1}{\alpha_1+\alpha_2}(\alpha_2 m_{1,(2)} + \alpha_1 m_{2,(2)})$. Let us also introduce the notation $m_* = \frac{1}{\alpha_1+\alpha_2}(\alpha_2 m_{1,(1)}^2 + \alpha_1 m_{2,(1)}^2)$. Note that if the distribution of the batch size in a given state is deterministic, or if it is geometric or Bernoulli, we have $m_{i,(1)}^2 = m_{i,(2)}$ $(i = 1, 2)$ and thus $m_* = m_2$. If the batch distribution is Poisson, we have $m_* + m_1 = m_2$.

Using eq. (10) and the possibility to explicitly compute the inverse of $\mathbf{I} - (1-\mu)\mathbf{P}$ in the 2-state case, we get

$$z_{\text{geo}}(\mu) = 1 + \frac{1}{2-\mu}\left(\frac{2}{m_1}\frac{(1-\mu)}{\mu}\left[m_* - \frac{(m_* - m_1^2)(1-\gamma)}{1-\gamma(1-\mu)}\right] + \frac{m_2}{m_1} - 1\right) - \frac{m_1}{\mu} \tag{11}$$

and by eq. (7) we get the peakedness curve.

It is interesting and important to note that the peakedness curve depends on the SBBP parameters only through m_1, m_2, m_*, γ. Therefore, we can get identical peakedness values for different SBBPs if these four parameters coincide.

(d) Batch renewal process

The batch renewal process is important to consider because of its ability to model the correlation structure of traffic [6]. The discrete time batch renewal process is made up of batches of arrivals, where the intervals between batches are independent and identically distributed random numbers, and the batch sizes are also independent and identically distributed, furthermore, the batch sizes are independent from the intervals between batches.

We use the following notation for the discrete time batch renewal process: a and b are the mean length of intervals between batches and the mean batch size, respectively. The first and second moments of the number of arrivals in an epoch is given by $m_1 = b/a$, and $m_2 = m_1 b(C_b^2+1)$ where C_b^2 is the squared coefficient of variation (variance to mean square ratio) of the batch size. The probability generating function of the distribution of time between batches is denoted by $A^*(\omega)$. ($A^*(\omega) = \sum_{s=1}^{\infty} a[s]\omega^s$ where $a[s]$ is the probability that the time between two consecutive batches is s.)

We have derived the peakedness for geometric holding times which is given by [10]

$$z_{\text{geo}}(\mu) = 1 + \frac{1}{2-\mu}\left(\frac{1+A^*(1-\mu)}{1-A^*(1-\mu)} - b + \frac{m_2}{m_1} - 1\right) - \frac{m_1}{\mu} \tag{12}$$

If the distribution of time between batches follows a shifted generalized geometric distribution [6], that is, $a[t] = 1-\sigma$ if $t = 1$ and $a[t] = \sigma\tau(1-\tau)^{t-2}$ if $t = 2, 3, \ldots$, then its probability generating function is: $A^*(\omega) = \omega\left(1 - \sigma + \frac{\sigma\tau\omega}{1-(1-\tau)\omega}\right)$ which makes the peakedness values easily computable.

2.5 Fitting traffic models to peakedness curves

The peakedness shows the variability of the arrival stream with respect to different service holding times. It is of interest to investigate whether we can fit traffic models to peakedness curves based on measurements.

We outline here a fitting procedure based on the mean rate m_1 of the arrival traffic, the peakedness value at $\mu = 1$ and at three other points, μ_1, μ_2, μ_3. The model we fit to the peakedness curve is an interrupted batch Bernoulli process (IBBP): in one state of the modulating Markov process, the arrival number has a general distribution, in the other state, there are no arrivals.

First, by $z(1) = m_2/m_1 - m_1$, we get m_2. Introducing $\omega = 1-\mu$, $\omega_i = 1-\mu_i$ and using the notations of section c, we can compute (using the values $K^*(\omega_i) = (z_{\text{geo}}(\mu_i) - 1)(\omega_i + 1) + 1$)

$$Y_i = Y(\omega_i) = m_1 \frac{1-\omega_i}{2\omega_i}\left(K^*(\omega_i) + m_1\frac{1+\omega_i}{1-\omega_i} - \frac{m_2}{m_1}\right) \tag{13}$$

Using eq. (11), $Y(\omega) = m_* - \frac{(m_* - m_1^2)(1-\gamma)}{1-\gamma\omega}$

Let us denote $\tilde{Y} = \frac{Y_1 - Y_2}{Y_2 - Y_3}$ which evaluates to $\tilde{Y} = \left(\frac{\omega_2 - \omega_1}{\omega_3 - \omega_2}\right)\left(\frac{1-\gamma\omega_3}{1-\gamma\omega_1}\right)$ and we get $\gamma = \frac{\tilde{Y}\frac{\omega_3-\omega_2}{\omega_2-\omega_1} - 1}{\tilde{Y}\frac{\omega_3-\omega_2}{\omega_2-\omega_1}\omega_1 - \omega_3}$ Once we have γ, we can obtain an estimation for m_* as $m_* = \frac{1}{3}\sum_{i=1}^{3} \frac{Y_i - \frac{m_1^2(1-\gamma)}{1-\gamma\omega_i}}{1 - \frac{1-\gamma}{1-\gamma\omega_i}}$ where we have on the right hand size an average for the known values ω_i, Y_i.

Then it is possible to fit an IBBP (no arrivals in state 2) as follows: $m_{1,(1)} = \frac{m_*}{m_1}, \alpha_2 = \frac{m_1(1-\gamma)}{m_{1,(1)}}, \alpha_1 = 1 - \gamma - \alpha_2, m_{1,(2)} = m_2\frac{\alpha_1+\alpha_2}{\alpha_2}$. Given the first and second moments of the number of arrivals in state 1, we can use for example a generalized geometric distribution for modeling the batch size distribution.

In this case, there are no arrivals with probability $1-\varphi$, and there is a batch of arrivals with geometrically distributed size of parameter ψ. The moments are given by $m_{1,(1)} = \varphi/\psi$, $m_{1,(2)} = \varphi/\psi^2$ by which we can get φ, ψ for the model.

If it is possible to exactly fit an IBBP to the μ_i, $z_{\text{geo}}(\mu_i)$ pairs, the values that are summed in the equation for m_* are identical. If there is no IBBP that exactly fits the given peakedness values, m_* gives an estimation and the peakedness curve of the fitted IBBP model approximates the μ_i, $z_{\text{geo}}(\mu_i)$ pairs.

3 GENERALIZED PEAKEDNESS OF REAL TRAFFIC

3.1 Measuring peakedness

To measure the generalized peakedness of a traffic with a given holding time distribution, one can simulate the infinite server group. In discrete time, one can keep track of the first and second moment of the number of busy servers and compute the variance to mean ratio from them. The following points should be made about the estimation.

- We should take care of the initial phase of the simulation. If we have no prior knowledge about the traffic, we do not know what the mean number of busy servers will be. In this case, we can start from an empty system. The initial transient in the number of busy servers should be excluded from measurements.
- According to the definition, we should assign a server to each arrival, that is, assign a random holding time variable to every arrival in an epoch, which could involve a huge amount of computational effort. However, using the modified definition of peakedness and eq. (5), we can reduce the computational effort by assigning only one random service time variable to all arrivals in an epoch.
- When the service time is geometric, we can minimize the computational effort by making use of the memoryless property. If at epoch t we have $L[t]$ busy servers, then at the next epoch we have $L[t+1] = L[t] + w[t+1] - D[t]$ where $D[t]$ is the number of departures from the service system at epoch t. The distribution of $D[t]$ is known to be binomial with parameters $L[t]$ and μ because each of the $L[t]$ servers finish service with probability μ. Therefore, in the measurement, it is enough to keep track of $L[t]$ together with the first and second moments of the previous $L[i], i \leq t$ values.
  This gives us the following procedure for computing the peakedness value for geometric holding time distribution with parameter μ:

  1. Reset $L_1 = 0$, $L_2 = 0$, L_{old} =initial value (see comments below);

2. Set $L_{new} = L_{old} + w_{new} - d$ where d is a random number with distribution $binom(L_{old}, \mu)$ and w_{new} is the number of new arrivals in the next epoch;
3. Set $L_1 = L_1 + L_{new}$, $L_2 = L_2 + L_{new}^2$;
4. Set $L_{old} = L_{new}$ and loop back to 2. unless the measurement is over;
5. Compute $l_1 = L_1/T, l_2 = L_2/T, z = l_2/l_1 - l_1$ where T is the length of the total measurement time.

The setting of the initial value of L_{old} depends on the amount of a priori information that we have about the traffic. If we know the mean rate, we can set the initial L_{old} to its mean value determined by Little formula as m_1/μ. If we do not know the mean rate, we have to start from an empty system (initial $L_{old} = 0$) and simulate the service system without actually measuring (executing step 3.) until the initial transient is over.

- An important advantage of using peakedness characterization is that we can measure peakedness by going through the traffic trace in only one sequence. This gives us the possibility of measuring peakedness for real-time traffic on the fly.

  Computing peakedness for one value of μ involves N cycles of the above procedure (where N is the total length of the measured traffic); if we want to measure peakedness at several μ values, we can easily implement the parallel execution of the procedure. In each cycle, we only have to compute a small number of additions and multiplications, and generate one binomially distributed random variable. Therefore, the complexity of the measurement is $O(N)$. The most time-consuming step in the measurement is the generation of the binomially distributed random number. We can reduce the computational cost of the measurement tremendously by approximating it with a normally distributed random number, for which pre-computed look-up tables can be used.
- The advantage of our approach compared to Eckberg's method for estimating peakedness for exponential holding times (cf. [3, 9]) is that our method does not neglect a lot of arrivals in the computation due to the selection of an arbitrary arrival.

3.2 Peakedness of video traffic

Video traffic is a very important example of variable rate traffic. We investigated the application of peakedness measure for the characterization of variability of MPEG video traces [17]. The MPEG sequences that we considered had a GOP (Group of Pictures) length of 12 frames, a GOP pattern of IBBPBBPBBPBB, and frames capture frequency of 25 frames per second.

Figure 1 shows the peakedness curve of an an MPEG video trace of a movie (MrBean) as a function of the service rate μ. The mean service time of a server is therefore $1/\mu$ time epochs, where one time epoch is now 40ms. The solid curve is the peakedness function for the frame sequence (one frame corresponds to one epoch), whereas the dashed curve is the peakedness function for the GOP sequence (one GOP corresponds to 12 epoch so that is has the same time-length as the frame sequence) The scaling in the vertical axis is such that one arrival corresponds to one bit.

By decreasing the service rate, the service times become longer, and the number of busy servers in the infinite server group depends on the traffic properties on longer time scales. In this way, the peakedness curves show the variability of the traffic on different time scales, i.e. on the time scale of $1/\mu$.

Figure 1 shows that on short time scales, the variability of the frame sequence is much greater compared to the GOP sequence. But as we go to longer and longer time scales, the variability of the two sequences converge. What we can learn from this is that on longer time scales (for example, when dimensioning larger buffers), the statistical characteristics of GOP structure is less significant, and it is enough to consider the GOP sequence.

Figure 2 shows the peakedness curves for geometric service time distributions for five MPEG video GOP size traces. It gives us a relative comparison of the variability of different kinds of video sequences. (In this figure, one time epoch is set to one GOP which introduces a scaling compared to Figure 1.) The highest values of peakedness are exhibited by the MTV sequence, which is known to have lots of scene changes. Movie sequences show lower peakedness compared to the MTV sequence. The peakedness of a video conference sequence is found to be the smallest by orders of magnitude.

Figure 3 shows an IBBP fitted to an MPEG movie trace (MrBean, [17]). The solid line is the peakedness curve of the GOP sequence, the dashed line shows the peakedness curve of the fitted model. The circles show the peakedness values where the fitting was made. The points were chosen to represent the variability of the traffic on a long time scale (corresponding to the time scale of 1/0.01=100 epoch, here one epoch corresponds to 0.48 sec). As we can see, the model is able to capture the variability of the arrival stream on the investigated time scales.

3.3 Peakedness of aggregated ATM traffic

We analysed the peakedness curve of an aggregated ATM traffic trace taken from the Finnish University and Research ATM WAN network (FUNET) [14]. The trace was approximately one hour long and consisted of the number of cell arrivals in each second. Figure 4 shows the peakedness curve of the measurement and two IBBPs fitted to it. The IBBP that was fitted at short time scale fits the measured peakedness curve well for shorter time scales, but it gives lower peakedness values for time scales longer than $1/0.05 = 20$sec. The other IBBP was fitted at a longer time scale; this model gives lower peakedness values for time scales shorter than 20sec.

3.4 Peakedness of Ethernet traffic

Figure 5 and Figure 6 show the peakedness curve of an Ethernet traffic taken from the Bellcore measurements [7]. The measurement covers 1 million arrivals (approx. one hour). Figure 5 depicts peakedness on a lin-lin plot, Figure 6 is a log-log plot. We can investigate 5 different time scales in Figure 6. The interesting finding is that the peakedness increases linearly on the log-log plot as we decrease the rate (go to long time scales). Due to eq. (9) and knowing that $\lim_{s\to\infty} I[s] = \infty$ if there is long range dependence (LRD) in the traffic, the peakedness diverges as the rate goes to zero. This observation of monotonicity in Figure 6 supports the presence of LRD assuming that the traffic stationarity assumption holds. It is important to note that *the peakedness curve can be used as an indicator of LRD.*

At different time scales we fitted simple Markovian models (IBBPs) to capture the peakedness curves in Figure 6. We can see that the burstiness scaling property of these models are not appropriate i.e. these models can cover a shorter range of time scales in burstiness than it would be necessary to follow the burstiness of the real traffic over all the investigated time scales.

Our investigations of the aggregated ATM and Ethernet traffic indicate that simple Markovian models are not able to capture the burstiness characteristic of traffic over many time scales. For this case fractal traffic models seem to be more appropriate [7, 14]. However, for several practical cases we do not need to focus on *all* time scales but only on our working time scales (e.g. time scales of queueing) which can be efficiently modeled by Markovian models, too.

4 CONCLUSION

We have shown that peakedness can be used to characterize the bursty nature of traffic. Peakedness curves show the variability of traffic on different time scales and can be efficiently computed for real time traffic. We have extended the peakedness theory to discrete time and applied the peakedness characterization to variable rate video traffic, Ethernet traffic and aggregated ATM traffic as well as to the most important traffic models. We have shown that generalized peakedness can also be used for detecting long range dependence. We have also presented a new model fitting technique based on the concept of peakedness.

The basic idea of peakedness characterization is that we characterize traffic by its interactions with the service system. Its generality is shown by the observation that peakedness gives a complete second order characterization, i.e. it contains all information about the correlation structure of the traffic.

The further development of peakedness theory including its extension to characterize non-stationary traffic are the topics of our future research.

REFERENCES

[1] D. R. Cox and P. A. W. Lewis. *The Statistical Analysis of Series of Events.* Methuen & Co Ltd, 1966.

[2] A. E. Eckberg, Jr. Generalized peakedness of teletraffic processes. In *ITC-10*, Montreal, 1983.

[3] A. E. Eckberg, Jr. Approximations for bursty (and smoothed) arrival queueing delays based on generalized peakedness. In *ITC-11*, Kyoto, Japan, 1985.

[4] R. Gusella. Characterizing the variability of arrival processes with indexes of dispersion. *IEEE Journal on Selected Areas in Communications*, 9(2), February 1991.

[5] H. Heffes and J. M. Holtzman. Peakedness of traffic carried by a finite trunk group with renewal input. *The Bell System Technical Journal*, 52(9):1617–1642, November 1973.

[6] D. Kouvatsos and R. Fretwell. Batch renewal process: Exact model of traffic correlation. In *High Speed Networking for Multimedia Application*, pages 285–304. Kluwer Academic Press, 1996.

[7] W. E. Leland, M. S. Taqqu, W. Willinger, and D. Wilson. On the self-similar nature of Ethernet traffic (extended version). *IEEE/ACM Transactions on Networking*, 2(1), February 1994.

[8] B. L. Mark, D. L. Jagerman, and G. Ramamurthy. Application of peakedness measures to resource allocation in high-speed networks. In *Proceedings of ITC-15, Washington D.C., USA*, June 1997.

[9] B. L. Mark, D. L. Jagerman, and G. Ramamurthy. Peakedness measures for traffic characterization in high-speed networks. In *Proceedings of IEEE INFOCOM'97*, 1997.

[10] Gy. Miklós. Peakedness measures. Technical report, High Speed Networks Lab, Department of Telecommunications and Telematics, Technical University of Budapest, 1997.

[11] S. Molnár. *Evaluation of Quality of Service and Network Performance in ATM Networks.* PhD thesis, Technical University of Budapest, Department of Telecommunications and Telematics, 1995.

[12] S. Molnár, I. Cselényi, and N. Björkman. ATM traffic characterization and modeling based on the leaky bucket algorithm. In *IEEE Singapore International Conference on Communication Systems*, Singapore, November 1996.

[13] S. Molnár and Gy. Miklós. On burst and correlation structure of teletraffic models. In D. D. Kouvatsos, editor, *5th IFIP Workshop on Performance Modelling and Evalution of ATM Networks*, Ilkley, U.K., July 1997.

[14] S. Molnár and A. Vidács. On modeling and shaping self-similar ATM traffic. In *Proceedings of ITC-15, Washington D.C., USA*, June 1997.

[15] R. O. Onvural. *Asynchronous Transfer Mode Networks, Performance Issues.* Artech House, Boston, London, 1994.

[16] V. Paxson and S. Floyd. Wide area traffic: The failure of Poisson modeling. *IEEE/ACM Transactions on Networking*, 3(3):226–244, 1995.

[17] O. Rose. Statistical properties of MPEG video traffic and their impact on traffic modeling in ATM systems. In *Proceedings of the 20th Annual Conference on Local Computer Networks*, pages 397–406, Minneapolis, MN, 1995. ftp://ftp-info3.informatik.uni-wuerzburg.de/pub/MPEG/.

[18] K. Sriram and W. Whitt. Characterizing superposition arrival processes in packet multiplexers fo voice and data. *IEEE Journal on Selected Areas in Communications*, 4(6), September 1986.

[19] G. D. Stamoulis, M. E. Anagnostou, and A. D. Georgantas. Traffic source models for ATM networks: a survey. *Computer Communications*, 17(6), 1994.

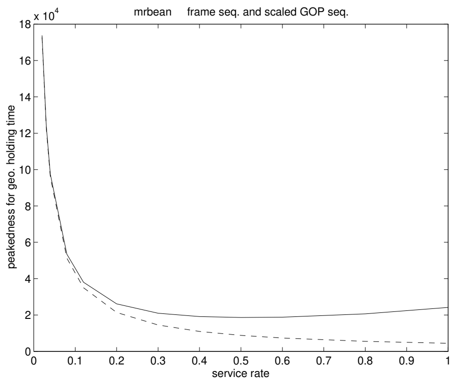

Figure 1: Peakedness of the frame (solid) and GOP (dashed) sequence of MPEG video trace (MrBean).

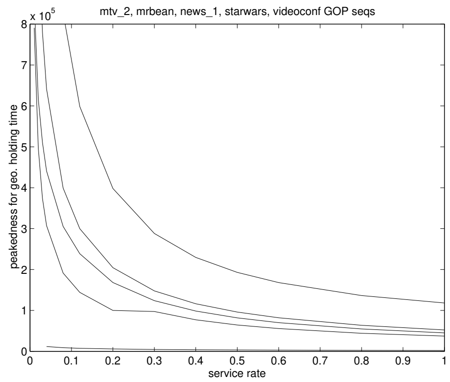

Figure 2: Peakedness of MPEG GOP video sequences. From the uppermost downwards, the sequences are from: TV (MTV), movie (MrBean), TV (News), movie (StarWars), video conference.

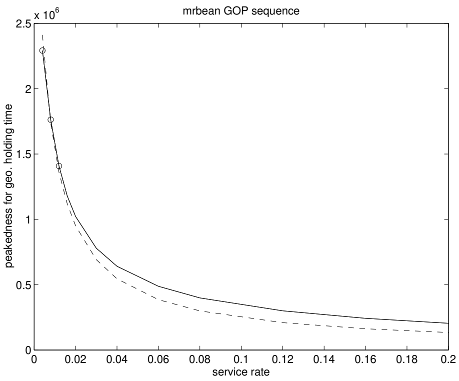

Figure 3: Peakedness curves of MPEG GOP movie trace (MrBean, solid) and its IBBP model (dashed).

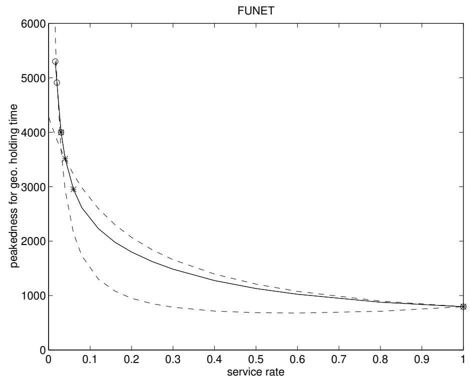

Figure 4: Peakedness of aggregated ATM traffic (solid) and IBBP models (dotted) fitted to it. The two IBBPs are fitted at short (stars) and long (circles) time scales.

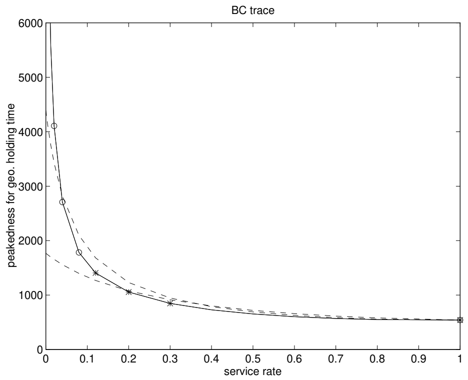

Figure 5: Peakedness of Ethernet trace (solid) and IBBP models (dotted) fitted to it. The two IBBPs are fitted at short (stars) and long (circles) time scales.

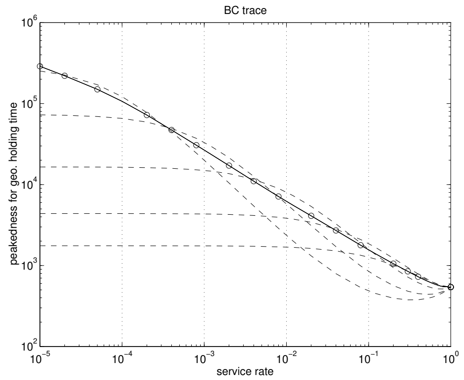

Figure 6: Peakedness of Ethernet trace (solid) in log-log plot. On five time scales (separated by vertical lines) IBBP models are fitted (dashed).