Gaussian Approximation of CDN Call Level Traffic

Andrzej Bak and Piotr Gajowniczek
Institute of Telecommunications
Warsaw University of Technology
Nowowiejska 15/19, 00-665 Warsaw, Poland
Email: bak@tele.pw.edu.pl

Marcin Pilarski
Orange Labs, Telekomunikacja Polska S.A.
Obrzezna 7, 02-679 Warsaw, Poland
Faculty of Mathematics and Information Science
Warsaw University of Technology
pl. Politechniki 1, 00-661 Warsaw, Poland

***Abstract*—The Content Distribution Networks (CDN) are based on one of the most fruitful ideas in nowadays Internet. There is an increasing competition between CDN providers that place their Web servers in various locations, closer to end users. The future Internet solutions have to assure however that users are not only able to observe the live events and download different content via caching systems, but also that the CDN providers can ensure the quality adequate to the observed demand. This requires knowledge about the characteristics of traffic incoming to the system. In this paper we analyzed the call-level CDN traffic on the base of observations and measurements taken from the Polish Telecom CDN network.**

I. Introduction

Content Distribution Networks (CDN) [1] play a very important role in current Internet infrastructure. A CDN deploys caching servers in multiple locations and provides algorithms to move the content requested by end users in such a way that the overall user experience is optimized. CDNs are deployed globally or regionally by specialized companies, such as Akamai [2], however, recently, the CDN infrastructures are also deployed by large Internet Service Providers (ISP), to optimize distribution of content within their networks.

Traditional CDN services are related to caching static Web pages and large file downloads. Today's CDNs often provide other services, such as delivering dynamic content, supporting Web 2.0 and streaming applications. There is much work done in the area of CDN optimization (server location etc.) and measurements related to the infrastructure of the well-known CDN networks (see for example [3]). However, there is very little work that investigates the properties of traffic (streams of requests) arriving to CDN servers, which is important for proper dimensioning of such systems. In this paper we analyze such streams, looking for self-similarity and long-range dependency properties on the base of measurements collected in a working ISP CDN system. We also investigate the applicability of Gaussian modeling to the analyzed data.

II. Measurement environment

The data collected for analysis comes from measurements undertaken in the CDN network of Polish Telecom (TP) that consists of several nodes (server farms) located in main cities of Poland, see Fig. 1.

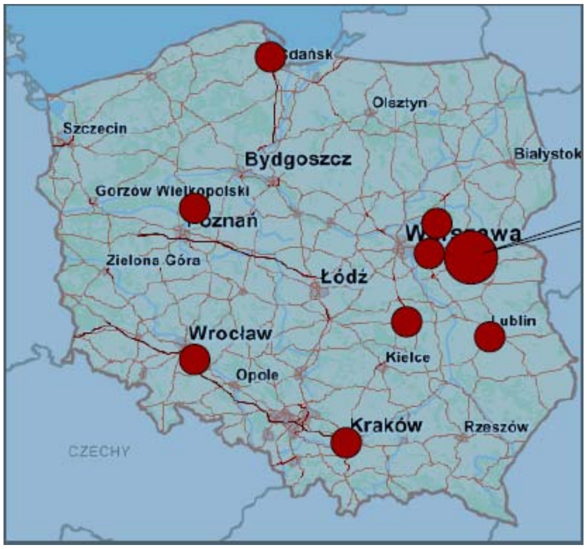


Fig. 1. TP CDN environment

The measurements come from two real-life cases. In the first one, the upgrade of popular PC game called "The Witcher" (www.witcher.phx.pl) was distributed in TP CDN environment and significantly contributed to the overall CDN load. The measurements were taken during several days when the upgrade was accessible (this data is hereafter called *the Witcher* data). In the second case we investigated the stream of requests to the CDN streaming server recorded during the time of live Internet transmission of events related to the Polish presidential plane crash in Smolensk, Russia (hereafter called *the livestream* data).

In both cases the number of requests was aggregated in 1 second intervals, forming traffic processes that were then checked for self-similarity and long-range dependency properties. We have also investigated the applicability of Gaussian processes (namely the well-known Fractional Gaussian Noise) to modeling the stream of requests incoming to the CDN servers.

III. Self-similarity

The term *self-similarity* is commonly used for describing the scaling invariance of the traffic process X_i, $i \in \mathbb{Z}$, created from measurements by collecting the numbers of bits or calls observed within the consecutive periods of fixed length (called *the base aggregation period*). The m-level aggregated process

978-0-9836283-0-9 © 2011 ITC

is then defined by Eqn. 1:

$$X^{(m)}(i) = \frac{1}{m} \sum_{t=m(i-1)+1}^{mi} X(t). \qquad (1)$$

Let $V = \sigma^2$ and r_k denote the variance and the autocorrelation function (ACF) of the original process and $V^{(m)}$, $r_k^{(m)}$ - the variance and ACF of the aggregated process. Assuming that X_i is second-order stationary, it is called *exactly second-order self-similar* if the original process and its aggregates have the same correlation structure, i.e.:

$$r_k^{(m)} = r_k, \quad m \in Z_+ \qquad (2)$$

The process is said to be *asymptotically second-order self similar* when:

$$\lim_{m\to\infty} r_k^{(m)} = r_k, \quad k \in Z_+ \qquad (3)$$

The second-order self-similarity of X_i implies that:

$$r_k \sim ck^{-\beta}, \quad \beta \in (0,1). \qquad (4)$$

In Eqn. 4, c is some positive constant, $\sim$ denotes asymptotically proportional to as k approaches infinity and $\beta = 2-2H$, where $H \in (\frac{1}{2}, 1)$ is commonly known as the Hurst parameter. The property defined by Eqn. 4 is called *long range dependence* (LRD), as the decay of the autocorrelation function of the process is slower than exponential. LRD implies that the decay of the variance of the aggregate process with increasing aggregation level is also slower than exponential, i.e.:

$$V^{(m)} \sim cm^{-\beta} \qquad (5)$$

The above definitions show that the Hurst parameter H can be interpreted as a measure of the scaling invariance of the process. Indeed, it had already become a classic measure of process self-similarity. However, the reliable estimation of its value for observed traffic streams is difficult. There are many different estimation methods that can yield diverse results, even if applied to the same data. Some examples of different approaches to estimating H are: the R/S approach [4], the VTP method [4] and the wavelet-based algorithms [5]. These methods are briefly described in the following subsections.

1) R/S method: From the traffic process X_i we create partial sums: $Y_n = X_1 + X_2 + \ + X_n$, $n \in Z$ and the so-called *adjusted range* R_n, defined as:

$$\begin{aligned} R_n = max\left(Y_i - \frac{iY_n}{n}; 1 \le i \le n\right) \\ -min\left(Y_i - \frac{iY_n}{n}; 1 \le i \le n\right) \end{aligned} \qquad (6)$$

where $\frac{Y_n}{n}$ is a sample mean for a given time range n.

The R/S method is a heuristics based on the fact that R_n is a measure of variability of X_n in relation to the mean value for various time ranges. The determination of the distribution of R_n and its statistical properties (even the mean value) is complex. With the sample standard deviation of $X_1 \dots X_n$ denoted as S_n, it can be however observed that for self similar processes the asymptotic behavior of the expected value of $\frac{R_n}{S_n}$ follows the power law:

$$E\left[\frac{R_n}{S_n}\right] \sim cn^H \qquad (7)$$

where c is constant and H is a Hurst parameter. Eqn. 7 allows estimating the value of H by calculating the slope of the linear regression line applied to the plot of $log\left[\frac{R_n}{S_n}\right]$ versus $log\ n$. The R/S method has inherent uncertainty implied by Eqn. 7 and can result in the values of H greater than 1, which is theoretically incorrect.

2) VTP method: VTP (*Variance-Time Plot*) approach is based on the concept of variance on multiple time scales and its specific behavior for self similar processes. According to [6], if the sample of the self-similar process is aggregated by a factor of m (corresponding to Eqn. 1), then, asymptotically, the variance of the aggregated process decreases by the same factor. The Hurst parameter can be then estimated by plotting the $log\ V^{(m)}$ versus $log\ m$ plot (called *the variance-time plot*), which indicates what is the change in the variability of the process when viewed over increasing time scales (increasing aggregation periods). From VTP one can calculate β as a slope of the linear regression line and then H using the relationship: $H = 1 - \frac{\beta}{2}$.

3) Wavelet-based approach: The method is based on the discrete wavelet transform (DWT), which for a given function f is defined by the set of coefficients $d_{j,k}(f)$, corresponding to the expansion of f into the following sum:

$$f(s) = \sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} d_{j,k}(f)\phi_{j,k}(s). \qquad (8)$$

where $\phi_{j,k}(s)$ is a set of linearly independent functions resulting from rescaling and time shifting of some square-integrable function $\phi(s)$ called the *mother wavelet*, such that:

$$\phi_{j,k}(s) := 2^{-j/2}\phi(2^{-j}s - k), \quad j,k \in \mathbb{Z}. \qquad (9)$$

The index j is commonly referred to as a *scale* and k as a *space*. Discrete wavelet transform conveys information about time/frequency characteristics of the signal, because each coefficient $d_{j,k}(f)$ characterizes the behavior of f at a given time scale (about $2^j k$) and frequency (about 2^j). Mother wavelets can be constructed for example by methodology described in [7]. The estimation of Hurst parameter is possible by analyzing the scaling behavior of the wavelet energy spectrum defined as $E\left[d_(j,k)^2\right]$. It can be shown that for LRD process:

$$log_2\mathbb{E}d_{j,k}^2 \cong j(2H-1) + c, \qquad (10)$$

where c is constant.

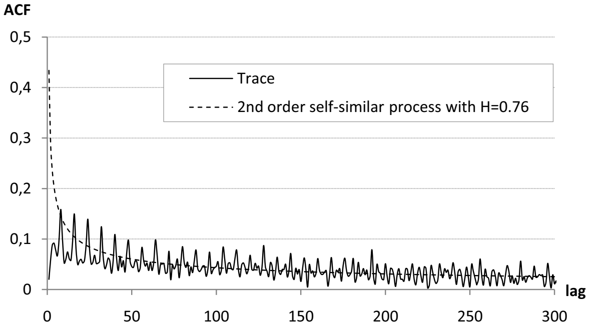


Fig. 2. ACF of *the Witcher* traffic process

For a discrete-time traffic process X_i the approximation of wavelet coefficients leads to the finite number of non-zero coefficients N_j for each scale j. Then, the so-called *wavelet spectrum* of X_i can be computed as:

$$S_j(X_i) = log_2 \left(\frac{1}{N_j} \sum_{k=1}^{N_j} d_{j,k}^2 \right). \tag{11}$$

From Eqn. 10 and 11 one can observe that the value of H can be estimated by plotting the graph of the wavelet spectrum S_j for the traffic process and using a linear regression over the selected range of scales j.

IV. Analysis of measured data

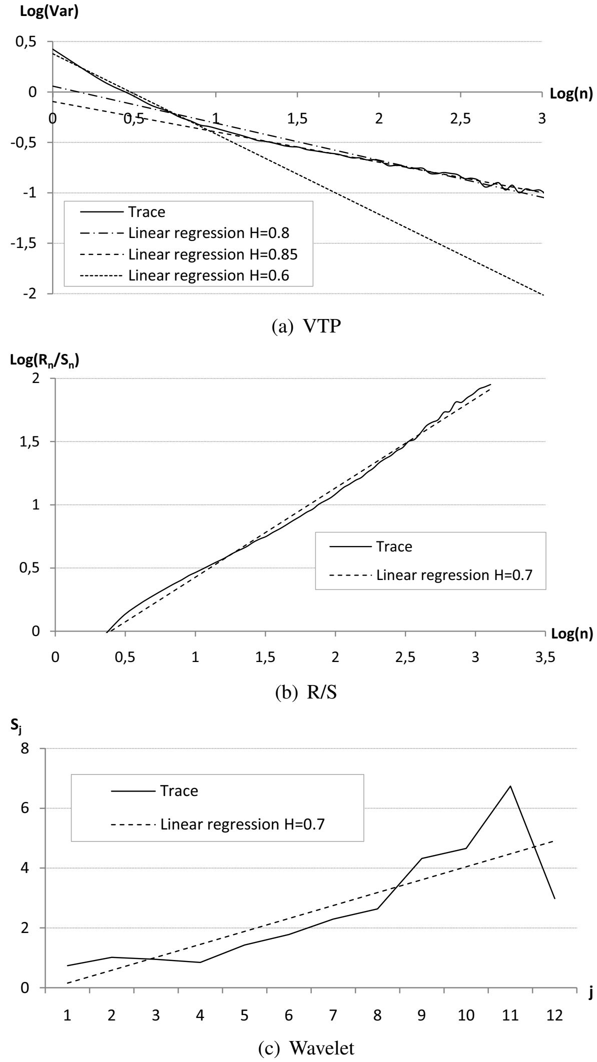
In this section we analyze the stream of requests arriving to the CDN servers on the base of measurements from two cases, described previously in Section II.

A. Self-similarity

Self-similarity of the traffic is important for studying the performance and for proper dimensioning of the network elements. In this section we analyze the self-similarity of the streams of calls arriving to the CDN servers. We are mainly interested in the statistical characteristics of the traffic, such as the shape of the autocorrelation function ACF (indicating the long range dependence) and the estimated values of the Hurst parameters, indicating the degree of self-similarity in the observed traffic streams. Traffic traces used for investigating the self-similarity properties (and further to validate the FGN modeling approach) for both *the Witcher* and *livestream* data were selected as examples of typical traffic after analyzing and searching the full request arrival processes registered during the large measurement period (several days) for samples characterized by most heavy but stationary traffic.

In Fig. 2 the ACF function of the traffic process for *the Witcher* data is presented together with the best-fit (least mean square error) theoretical ACF of the second-order self-similar process with the resulting Hurst parameter value equal to 0.76.

In Fig. 3 the plots related to the VTP, R/S and wavelet method described briefly in Section III are presented, together with the corresponding linear regression lines that allow estimating the values of Hurst parameter from these plots. The wavelet transform used here is based on the Daubechies mother wavelets [7]. In Fig. 3(a) there are two additional regression lines, corresponding to the two areas in the plot where the traffic process variance scales differently with increasing aggregation level, which leads to significant differences in an estimated value of H.



(a) VTP

(b) R/S

(c) Wavelet

Fig. 3. Self-similarity analysis of *the Witcher* traffic process

The obtained results vary from $H = 0.7$ to 0.85, however even the most conservative value obtained using the wavelet approach leads to the conclusion that the observed stream of requests incoming to the CDN server in case of *the Witcher* data is self-similar and therefore requires proper approach to modeling, that takes such characteristics into account.

Similar set of results is presented for the traffic process obtained from *the livestream* data in Fig. 4 and 5. In this case the value of H varies from 0.63 to 0.77, also indicating the presence of self-similarity in the investigated traffic process.

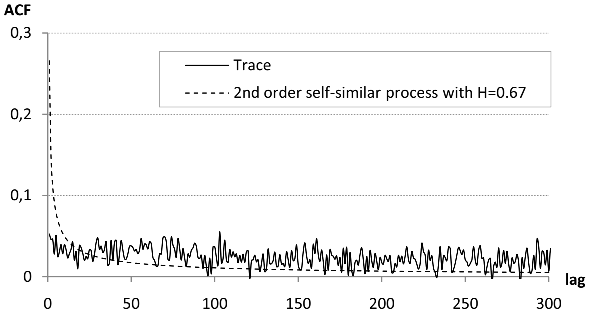
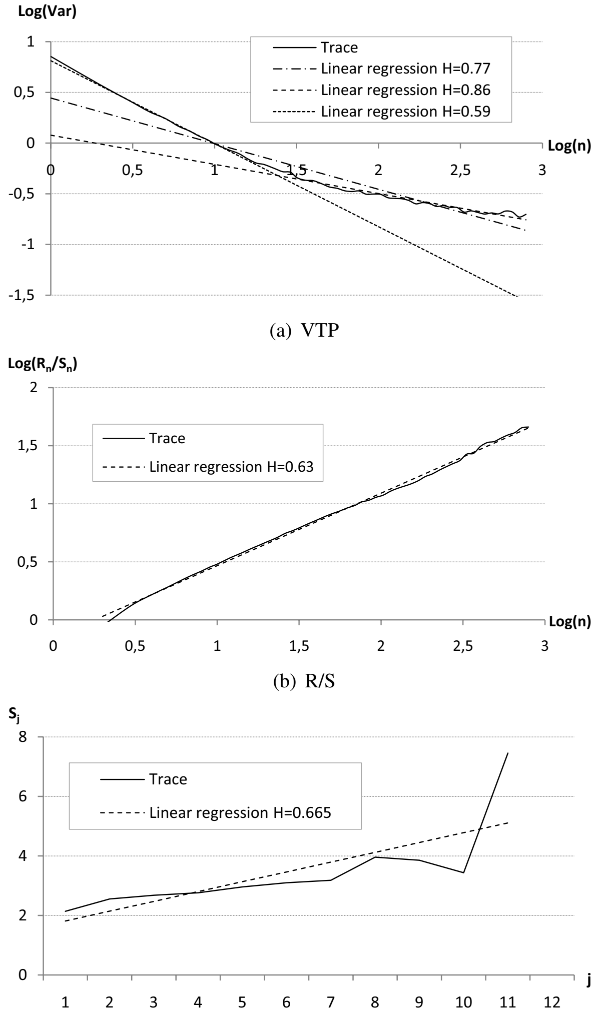


Fig. 4. ACF of *the livestream* traffic process



(a) VTP

(b) R/S

(c) Wavelet

Fig. 5. Self-similarity analysis of *the livestream* traffic process

B. Gaussian modeling

As it was shown in the previous section, the request arrival processes observed at the investigated CDN servers are self-similar, so their modeling using the memoryless processes is not proper. The Gaussian models are well-known as having the ability to recreate such properties of the original data. The term Gaussian means that all marginal distributions of traffic increments are approximately normal. Gaussian models are useful for modeling self-similar streams because their correlation structure is determined only by the variance function, so the whole spectrum of self-similar processes can be modeled. One of the best understood LRD Gaussian processes is called the Fractional Gaussian Noise (FGN) [8]. The FGN process is actually an incremental process:

$$Y_k = B_k - B_{k-1}, \quad k \in \mathbb{Z} \tag{12}$$

with mean μ, variance σ^2 and the ACF of the form:

$$r_k = \frac{1}{2}\left(|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H}\right), \quad k \in \mathbb{Z}. \tag{13}$$

B_k in Eqn. 12 is a Fractional Brownian Motion (FBM) [8] random process with Hurst parameter H. FBM process is a continuous-time, zero-mean Gaussian process with independent, stationary increments and autocorrelation function of the form:

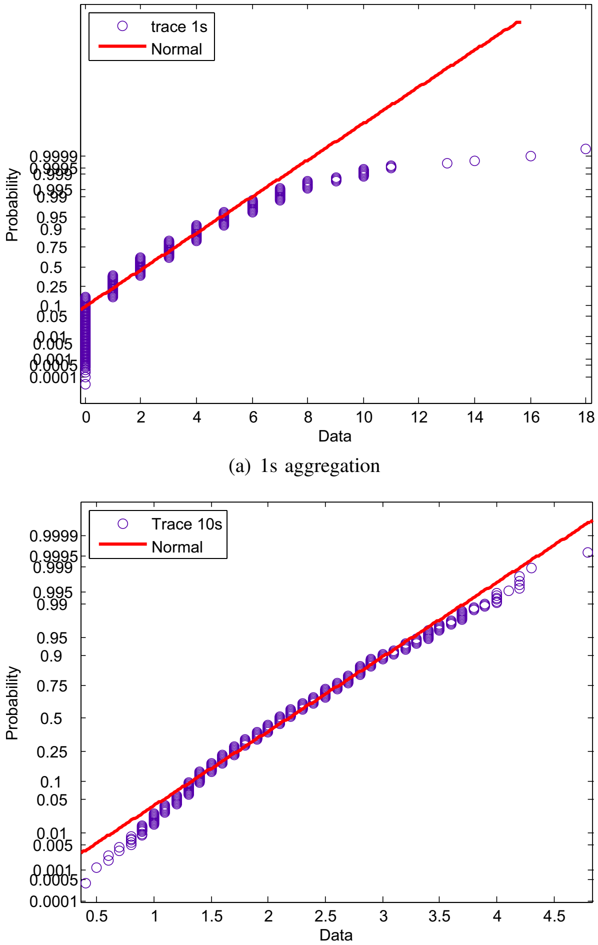
$$r_{(s,t)} = \frac{1}{2}\left(s^{2H} + t^{2H} - |s-t|^{2H}\right), \quad s,t \in \mathbb{R}_+. \tag{14}$$

The use of the FBM-based traffic model has the advantage, that the single server queue with unlimited buffer space and self-similar input process defined using the FBM process is analytically tractable (see [9]). This leads to the closed-form expression for estimating the equivalent bandwidth a notion related to the capacity of the server required to handle an incoming traffic with given Quality of Service (QoS) parameters.

In this section we use the FGN random process to model the empirical traffic data stream and examine the effectiveness of this approach. To accomplish this we have generated the artificial FGN traces at the 1s time scale using the method described in [10], which is based on creating the FGN power spectrum for a given number of samples and H and performing the inverse Discrete Time Fourier Transform (DTFT) to get the FGN time samples. This method is very fast due to the use of the Fast Fourier Transform (FFT) algorithm and is based on just three parameters: H, σ^2 and μ, estimated from the measured traffic streams aggregated in the same period.

The limitation of the FGN model is that with some probability it may produce negative values. The probability of yielding negative values depends on the ratio of the standard diviation to the mean value of the process. Whenever the standard deviation increases, the probability of having negative values also increases, which limits the usefulness of the FGN process for modeling real traffic with high variability (in relation to its mean). Also, for the processes with complex correlation structure, using the H as the sole parameter describing it may not be sufficient. In addition, for certain (usually small) time scales, the sampled process may not be well approximated by the Gaussian process, so the FGN model may not apply.

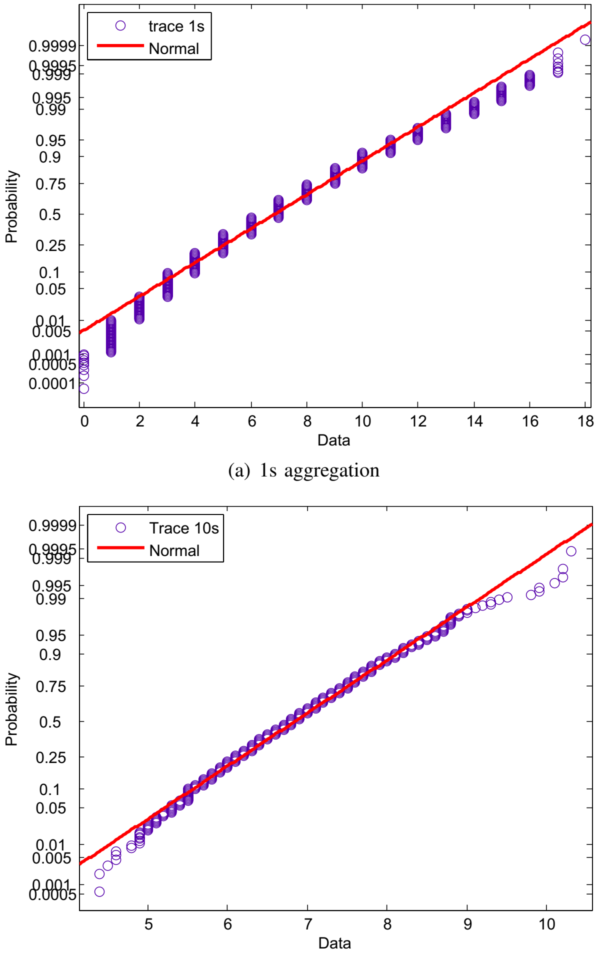
Therefore, we have analyzed the applicability of the FGN model to the measured traffic streams. In Fig. 6 the Q-Q (quantile-quantile) plots of the traces aggregated in sampling



(a) 1s aggregation

(b) 10s aggregation

Fig. 6. Q-Q plots for *the Witcher* data



(a) 1s aggregation

(b) 10s aggregation

Fig. 7. Q-Q plots for *the livestream* data

intervals of 1 and 10s are shown for *the Witcher* data; analogous plots for *the livestream* data are presented in Fig. 7. The Q-Q plots compare the quantiles of the traffic process distribution with quantiles of the best-fit normal distribution; the linear shape of the plot suggests that the distributions are the same.

The analysis of the Q-Q plots shows that for sampling interval of 1s there is not enough aggregation in time for Gaussian properties to show, especially in *the Witcher* data, but with increasing sampling interval the distributions get closer to Gaussian. It's worth to note that the traffic volume in the analyzed data (both *the Witcher* and *the livestream*) was relatively small. To get an insight how such traffic would look like in case of larger CDN server, we have artificially created an aggregated stream, joining several traffic streams from different days of *the livestream* measurements into a single process with much higher traffic intensity. Such traffic stream starts to exhibit good Gaussian properties even for small aggregation periods such as 2s (see Fig. 8), so one can assume that the validity of Gaussian modeling for CDN request arrival streams will increase together with the increasing server capacities.

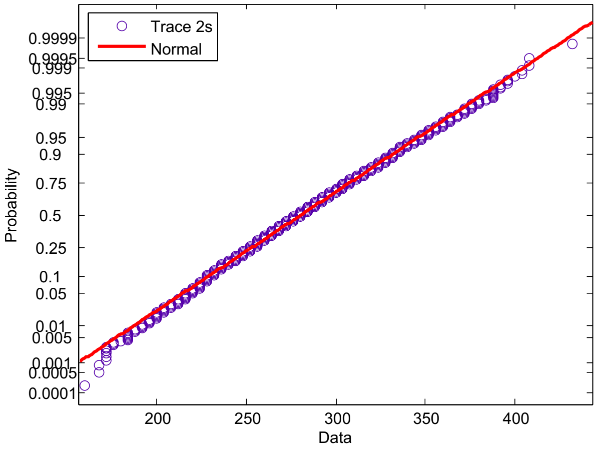
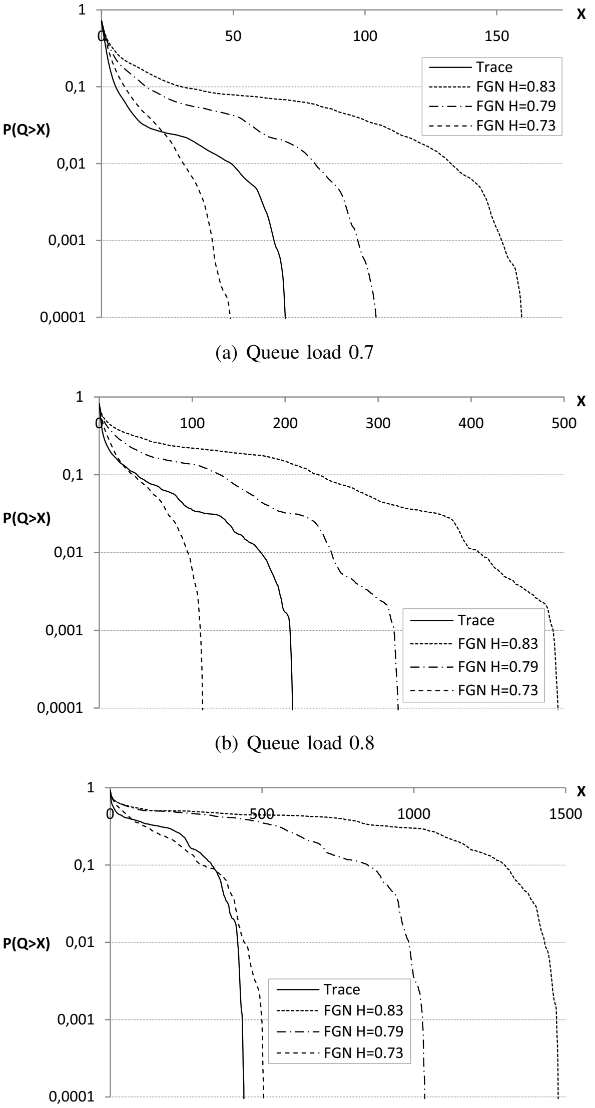


Fig. 8. Q-Q plot for combined *livestream* data, 2s aggregation

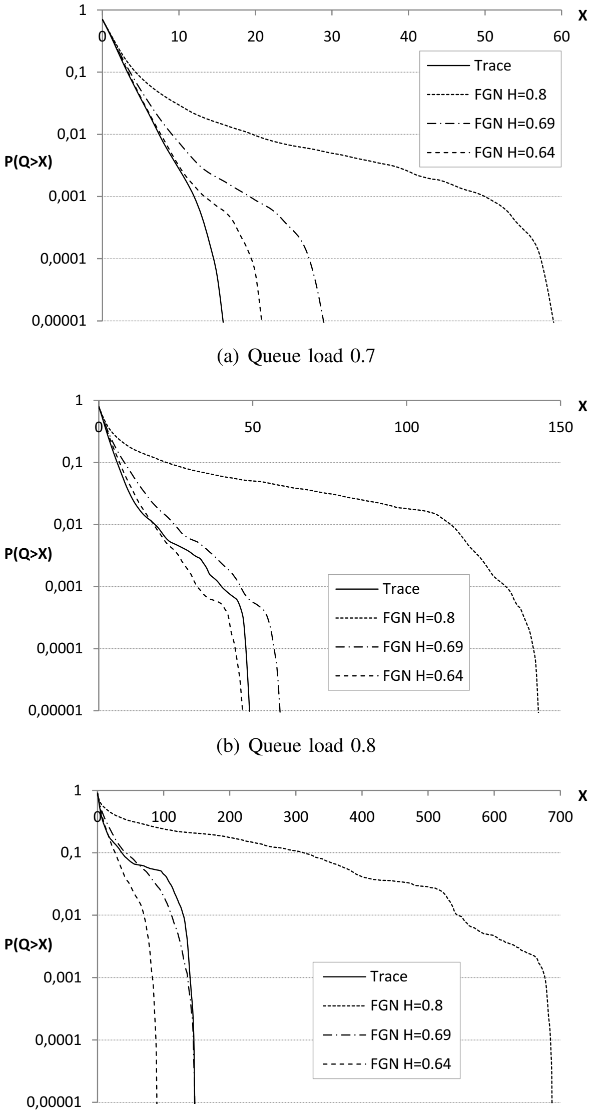


(a) Queue load 0.7

(b) Queue load 0.8

(c) Queue load 0.9

Fig. 9. Queue simulation for *the Witcher* data



(a) Queue load 0.7

(b) Queue load 0.8

(c) Queue load 0.9

Fig. 10. Queue simulation for *the livestream* data

C. Queueing analysis

To further validate the use of Gaussian models, we have simulated the FIFO fluid queue with single server and infinite buffer, fed with the actual traffic processes from both *the Witcher* and *the livestream* data and the artificial traces generated from the FGN model with corresponding Hurst parameter values, estimated on the base of the real data.

We did not intend to model any real elements of the CDN network with the above queueing system. It was used only to compare the arrival processes (so the service time was arbitrarily set to 1) by investigating their impact on the performance of the queueing system. However, such system can be viewed as a rough approximation of the request processing element of the CDN controller.

The complementary CDF of the buffer occupancy was compared for queue load factors ranging from of 0.7 to 0.9. The results are presented in Fig. 9 and 10. For each load factor we compared the measured trace with the FGN processes generated with different H parameters (using values estimated with the methods described in Section III.

First, from Fig. 9 it can be noticed that the *the Witcher* trace does not have enough Gaussian properties to be modeled by the FGN process. This trace starts to behave like Gaussian process for aggregation intervals larger that 10s (see Fig. 6). Better results were obtained for the *livestream* trace (Fig. 10). The *livestream* trace exhibits better Gaussian properties even for smaller aggregation intervals (starting from aggregation intervals larger than 1s).

The second observation that can be drawn from the queuing analysis is that the VTP method seems to overestimate the H parameter. This method is sensitive to trends and other instabilities in the measured data, and in practice it is very difficult to completely avoid such effects in case of real traces, so the usefulness of VTP to estimating the value of H is limited. The wavelet method in both cases gives the best estimation of the H parameter with the "classic" R/S method being the close second.

V. Conclusion

In the paper we have shown that the CDN call level traffic can exhibit self-similar properties, so often reported in case of measurements done in packet networks. The presence of self-similarity and LRD properties makes traditional call level modeling based on the assumption of Poisson arrivals inaccurate. We have investigated the usefulness of the Gaussian modeling for the CDN traffic. In particular we have analyzed the approximation accuracy of such traffic with the use of FGN process. Although the obtained results do not fully justify the use of FGN process to model the analyzed traces we believe that this is caused by relatively small level of traffic aggregation present in the measured data.

The Gaussian modeling requires sufficient level of traffic aggregation, both in vertical dimension (where large number of independent sources contribute to the aggregated stream) and in horizontal dimension (where the time scale is sufficiently large). The convergence to Gaussian process may not be present at certain level of aggregation (e.g. if the number of independent sources contributing to the measured traffic is too small). The analyzed traces had relatively small traffic intensities which suggests that there was either too little independent users contributing to the traffic or a single user generated very small traffic. In both cases it means that larger horizontal aggregation interval is required for Gaussian approximation and explains the differences obtained for *the Witcher* and *the livestream* in Section IV. It can be expected that with the increase of the traffic levels in the measured data the Gaussian modeling will be more accurate.

References

[1] M. Pathan, R. Buyya, and A. Vakali, *CDNs: state of the art, insights, and imperatives*, Content Delivery Networks, R. Buyya et al. (Eds.), Vol. 9, Springer-Verlag, 2008

[2] http://www.akamai.com/html/technology/index.html

[3] C. Huang, A. Wang, J. Li, and K. Ross, *Measuring and evaluating large-scale CDNs*, Proceedings of the 8th ACM SIGCOMM conference on Internet measurement, pp. 15–29, 2008

[4] R. G. Clegg, *A practical guide to measuring the Hurst parameter*, International Journal of Simulation: Systems, Science & Technology 7(2) 2006

[5] P. Abry and D. Veitch, *Wavelet analysis of long range dependent traffic*, IEEE Transactions on Information Theory, (44 (1)):2–15, 1998

[6] W.E. Leland, M.S. Taqqu, W. Willinger, and D.V. Wilson, *On the Self-Similar Nature of Ethernet Traffic (Extended Version)*, IEEE Transactions on Networking, Vol. 2, pp. 1–15, 1994

[7] I. Daubechies, *Ten Lectures on Wavelets*, Volume 61 of *CBMS-NSF Regional Conference Series in Applied Mathematics*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1992

[8] J. Beran, *Statistics For Long-Memory Processes*, Chapman and Hall, 1984

[9] I. Norros, *On the use of Fractional Brownian Motion in the theory of connectionless networks*, IEEE Journal on Selected Areas in Communications, vol. 13, no. 6, pp. 953–962, 1995

[10] V. Paxson, *Fast, Approximate Synthesis of Fractional Gaussian Noise for Generating Self-Similar Network Traffic*, Computer Communications Review, vol. 27, no. 5, pp. 5–18, 1997