12th GI/ITG CONFERENCE ON MEASURING, MODELLING AND EVALUATION OF COMPUTER AND COMMUNICATION SYSTEMS

3rd POLISH-GERMAN TELETRAFFIC SYMPOSIUM

IMPACT OF ACCESS BANDWIDTH ON AGGREGATED TRAFFIC BEHAVIOR AND QUEUEING PERFORMANCE

Guoqiang Hu
Institute of Communication Networks and Computer Engineering,
University of Stuttgart, Pfaffenwaldring 47, D-70569 Stuttgart, Germany
E-mail: hu@ikr.uni-stuttgart.de

Abstract

Recent traffic measurement of IP backbone networks discovered that aggregated IP traffic can be either uncorrelated or strongly correlated at small time scales, although at large time scales it exhibits long range dependence (LRD). Based on the infinite source Poisson traffic model, we show in this paper that the lack of correlation is an intrinsic small time scale property of LRD traffic due to the multiplexing of a large number of independent traffic flows. Particularly, the access bandwidth of users is a critical factor determining the boundary between the two ranges of time scales in which the uncorrelation and correlation property dominate the traffic behaviour respectively. A higher access bandwidth makes this boundary located at a smaller time scale. We refer to this time scale as boundary time scale and argue that the existence of a very small boundary time scale can explain the reported strong correlation at small time scales. Furthermore, the traffic behavior at different time scales leads to different queueing behaviors with respect to different queueing lengths. It is shown a large boundary time scale results in a large degree of buffer efficiency and thus brings substantial performance improvement in spite of the existence of LRD at large time scales.

Keywords
Internet traffic, long range dependence, access bandwidth, traffic aggregation, time scale, simulation, queueing performance

1. INTRODUCTION

The long range dependence (LRD) phenomenon, which describes the existence of strong temporal correlation over a large time span, was discovered as an ubiquitous large time scale phenomenon of IP traffic [3][5][9][11][13]. LRD traffic has

the hallmark of slow decay of traffic rate variability with the increasing measurement window. Explicitly,

$$\mathrm{VAR}[X_t] \sim ct^{2H-2} \text{ for } t \to \infty \tag{1}$$

where X_t denotes the traffic rate in byte/s observed in a time interval of length t, c is constant and $0.5 < H < 1$ is a measure of the degree of LRD called Hurst parameter. Larger values of Hurst parameter indicate higher degrees of LRD and an uncorrelated process like Poisson process has a Hurst parameter of 0.5. For the slow decay of variance, LRD traffic is regarded as bursty and generally has a significant detrimental impact on the queueing performance, for instance, high traffic loss for small buffer and large queueing delay in case of large buffer.

For the small time scale traffic behavior, recent measurement of the backbone IP traffic [15] discovered some very interesting features. In the time scale range of about 1ms~100ms, traffic fluctuation is either uncorrelated or strongly correlated depending on the composing traffic flows. If the aggregated traffic mainly consists of *sparse flows*, i.e., the flows with large packet interarrival time, it shows uncorrelation or only slight correlation in the observed time scale range. On the other hand, if there are a large number of *dense flows*, characterized by small packet interarrival time, strong correlation arises.

This paper aims to theoretically explain the reported small time scale behavior and inspect its indication for queueing performance. On the basis of infinite source Poisson traffic model [14] it is shown that the lack of correlation is an intrinsic small time scale property due to the multiplexing of independent sources. The uncorrelation characteristic is constrained within a time scale range, the upper bound of which is determined by the access bandwidth of individual users. If this bound is located at a very small time scale, the uncorrelation structure cannot be captured in practical measurement due to the limitation in the time granularity, so only the correlation property is recorded. Furthermore, it is demonstrated by simulation that the time-scale-dependent traffic behavior leads to queueing-length-dependent queueing behavior and the uncorrelation structure at small time scales can increase the buffer efficiency considerably.

The remainder of the paper is structured as follows. In Section 2 the characteristic of synthetic traffic is presented and analysed. The indication of small time scale traffic behavior for queueing performance is inspected in Section 3. The paper is concluded in Section 4.

2. UNCORRELATION BEHAVIOR AND ACCESS BANDWIDTHS

In this section, by simulation we demonstrate the existence of the uncorrelation structure in the aggregated LRD traffic and figure out the time scale bounding the

uncorrelation and strong correlation behavior. The role of the access bandwidth is identified.

2.1 Simulation scenario

The simulation scenario is illustrated in Fig. 1. Here n independent identical LRD traffic flows are aggregated on to one 622Mbps backbone link. Servers s1~sn are identical and used to model access links of limited capacity C_a. Server B is to model the backbone link with capacity C_b=622Mbps. Unbounded FIFO queues are applied so that there are no packet losses. The aggregated traffic rate X_T is measured at point P with a time interval of $T = 100\mu s$.

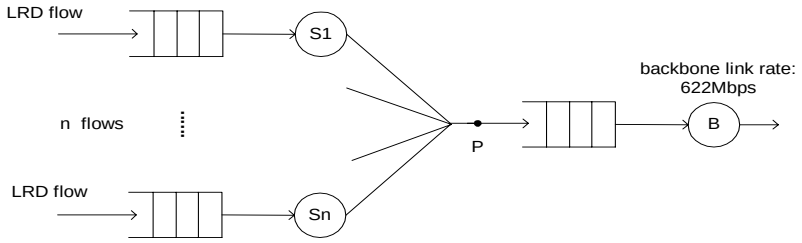

Figure 1: *Simulation scenario*

For each individual LRD flow the infinite source Poisson traffic model [14] (also known as $M/G/\infty$ model) is applied. Data sessions are generated according to Poisson process. The size of a session follows the Pareto distribution with a mean value of 10 Kbytes and parameter α=1.6. The resulting LRD flow thus has a Hurst parameter equal to 0.7. Data sessions are segmented into packets for transmission with maximal packet size of 1500 bytes.

2.2 Simulation results

To observe the possible influence of different system parameters on the traffic characteristic, different access bandwidths, backbone and access link utilization are taken into consideration. The simulation results are drawn in variance-time plot. The y-axis represents $\log_2\sigma_j^2$ where σ_j^2 denotes the variance of traffic rate X_t observed in a time interval of $t = 2^j T$, $j = 0, 1, 2, \ldots$. The x-axis corresponds the parameter j. For ideal LRD traffic it can be derived from Equation (1) that $\log_2\sigma_j^2$ has a linear relation to j and the Hurst parameter H can be estimated from the slope of the fitted line.

Fig. 2 shows the impact of the different access link utilization ρ_a on the traffic characteristic. ρ_a is set to 0.2, 0.5, 0.8 and $\rho_a \to 0$ respectively. Note $\rho_a \to 0$ means each flow contains only one data session and $n \to \infty$ is required to reach the given backbone link utilization. In this case the aggregated traffic degrades to traffic

generated by a single infinite source Poisson model. The access bandwidth C_a is fixed to 10Mbps and the backbone link utilization ρ_b is 0.8. For comparison, the variance-time relation of a single packet flow arriving according to Poisson process and having constant packet size of 1500 bytes is also plotted.

It can be seen that the curve of each aggregated LRD traffic has a knee area around Scale 3 and 4. At scales lower than and equal to 3, the variance of the traffic rate decreases as fast as, or in case of large access link utilization even faster than that of the Poisson traffic, which precludes the existence of strong correlation in this time scale range. At scales larger than 4, LRD traffic curves deviate from the Poisson traffic curve drastically and the variance decreases slowly with the increasing time scale, which is a sign of LRD as indicated in Equation (1). In general, the LRD traffic behaves differently in two time scale ranges and the boundary between these two ranges, i.e., the location of the knee, is independent of the access link utilization. This will be referred to as *boundary time scale* t_b in this paper. The curve corresponding to larger access link utilization has however lower variance everywhere since the traffic flows experience more intensive "shaping" on the access link and also because less number of flows is required to reach the given backbone link utilization.

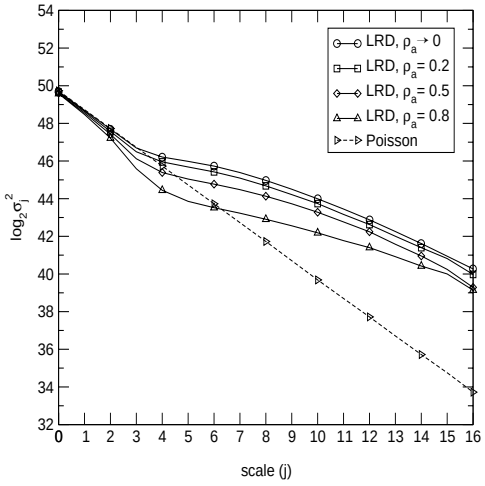

Figure 2: *Variance-time plot wrt. different access link utilizations*

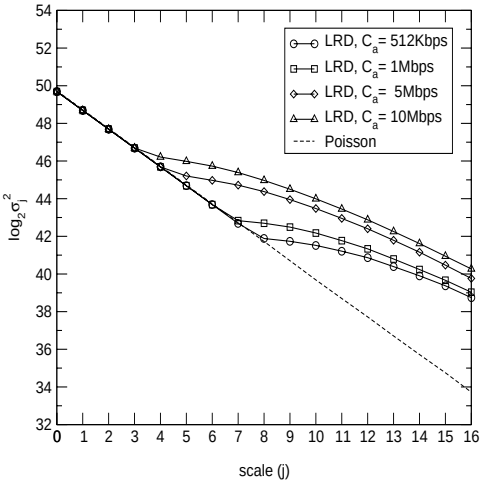

Figure 3: *Variance-time plot wrt. different access bandwidths*

In Fig. 3, the influence of different access bandwidths C_a: 512Kbps, 1Mbps, 5Mbps and 10Mbps is illustrated. The backbone link utilization is again 0.8 and only the case $\rho_a \to 0$ is shown for clarity. The boundary time scale is now located at different time scales depending on access bandwidths. The values of the boundary time scale t_b as well as the correspondent absolute time are tabulated in Table 1.

It is seen that the boundary time scale t_b is always close to the maximal packet transmission time T_{pkt} on the access link:

$$T_{\text{pkt}} = \frac{8 \text{ bits/byte} * 1500 \text{ bytes}}{C_a}. \quad (2)$$

Table 1: boundary time scale vs. access bandwidth

C_a	t_b (absolute time)	T_{pkt}
512Kbps	7~8 (12.8~25.6 ms)	23.4 ms
1Mbps	6~7 (6.4~12.8 ms)	12 ms
5Mbps	4~5 (1.6~3.2 ms)	2.4 ms
10Mbps	3~4 (0.8~1.6 ms)	1.2 ms

2.3 Explanation

To explain the simulation results, first note that more than 90% of the synthetic traffic is composed of packets of 1500 bytes due to the segmentation mechanism. So approximately the packet length can be regarded as constant. The number of packets transmitted on one access link within time interval $T' < T_{\text{pkt}}$ is either 0 or 1, which can be described with a Bernoulli model. Let the probability of one packet being transmitted is p, then the probability of no packet being sent is $1 - p$. p equals the mean number of transmitted packet within T', and is proportional to T' and access link utilization. The number of aggregated packet arrival $Y_{T'}$ from total n flows within $T' < T_{\text{pkt}}$ follows a Binomial distribution:

$$P\{Y_{T'}=m\} = \binom{n}{m} p^m (1-p)^{n-m}. \quad (3)$$

Then the mean and variance of $Y_{T'}$ are $\mathrm{E}[Y_{T'}] = np$ and $\mathrm{VAR}[Y_{T'}] = np(1-p)$. The packet arrival process of the aggregated traffic within $T' < T_{\text{pkt}}$ is independent because of the independence between n flows. So the uncorrelation property is a natural consequence of the statistical multiplexing.

When the backbone link utilization is fixed, $\mathrm{E}[Y_{T'}] = np$ keeps constant. With very small access link utilization there are $p \to 0$ and $n \to \infty$ so that the counting process $\{Y_{T'}\}$ becomes Poisson process (Poisson theorem) and $\mathrm{VAR}[Y_{T'}] = \mathrm{E}[Y_{T'}] = np$. This is reflected from the overlapping of LRD traffic curves with Poisson traffic curve at the small time scales in Fig. 3. Larger access link utilization leads to smaller value of $1 - p$ and $\mathrm{VAR}[Y_{T'}]$. However, if T' is tiny, p becomes so small that $\mathrm{VAR}[Y_{T'}] \approx np$, again similar to Poisson process.

This explains the dips of those LRD traffic curves with access link utilization greater than 0 in Fig. 2.

When the time interval exceeds T_{pkt}, the temporal correlation of each flow must be taken into consideration and the Bernoulli model is not valid any more. The heavy tailedness of the session size distribution begins to affect and leads to LRD. Thus T_{pkt} acts as a boundary between uncorrelation and strong correlation characteristic of aggregated LRD traffic.

It is necessary to point out here that the study of the uncorrelation property of aggregated traffic with the Bernoulli model is not new. For example, it is applied in [8] for the assembled burst traffic in optical burst switching (OBS) networks and in [12] for the study of the variance-mean relation at small time scales of traffic traces. Also in [2] the Poisson characteristic due to the multiplexing is discussed for Internet traffic. Our contribution here is that we identify the boundary time scale explicitly for the aggregated LRD traffic and disclose its relation to the access bandwidth.

2.4 Indication for real IP traffic

In real IP networks, about half of the traffic volume is composed of packets of 1500 bytes [16]. It was also found that packet interarrival time of traffic flows is closely coupled with access bandwidth [1]. So, the dense/sparse flows characterized in [15] indeed represent traffic flows from access links of different link capacities, which is also indicated in [15]. The current network access link bandwidths cover a large spectrum, from 56Kbps modem to 10/100Mbps Ethernet link. The boundary time scale of the aggregated traffic therefore depends on the statistical distribution of the access bandwidth of terminals. If the aggregated traffic contains a large amount of high rate flows like 10Mbps, the correspondent boundary time scale is in the order of one millisecond. In this case, the uncorrelation structure is invisible due to the limitation of the time granularity in measurement and only the strong correlation is captured at small time scales. This illuminates the observation in [15] as mentioned in Section 1.

To further justify our argument quantitatively, we note that in the case of strong correlation in time scale of 1ms~100ms, the dense flows, most of the packet interarrival time of which is less than 2 ms, amount to 15~20% of total traffic according to [15]. We simulate the aggregated traffic by adopting heterogeneous access link capacities. In Fig. 4 the variance-time plot for synthetic traffic aggregated from access links of different rates (1Mbps and 10Mbps) is presented. It can be seen that in case the traffic from 10Mbps links takes a slight portion (2%), the location of the boundary time scale is still decided by the transmission time on the 1Mbps link. However, it is clearly shifted to the smaller time scale provided that the contribution

of high speed access links goes up to 20%, matching well with the measurement results.

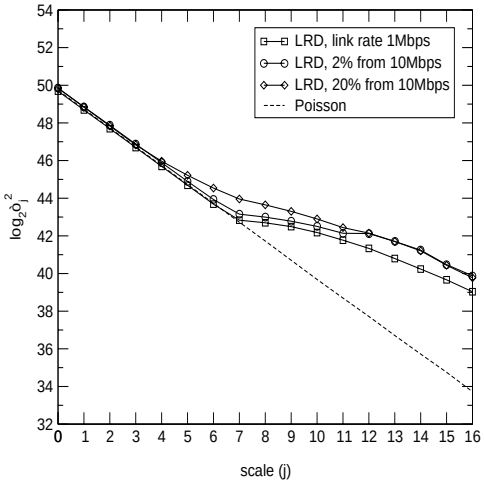

Figure 4: *Variance-time plot wrt. traffic of different combinations*

3. QUEUEING PERFORMANCE

The LRD characterizes the slow decay of variance with the increasing time scale. However, it alone does not decide the absolute value of the variance, which can be more important for queueing performance. It is pointed out in [10] that the buffer overflow probability is considerably affected by the absolute traffic variability on a so-called *relevant time scale* which denotes the time scale most relevant for the formation of the concerned queueing length. In [7] it is shown that for finite buffer only the correlation within a specific time scale (*cutoff lag*) influences the queueing performance. In these papers, it is also derived that the relevant time scale (or cutoff lag) is proportional to the queueing length (or buffer size). However, the traffic model studied in [10] is the classical fractional brownian motion (FBM) only for the large time scale LRD characteristic. In [7] uncorrelation at small time scales is inspected, but the adopted traffic model behaves like on-off source model after specialization, which is not suitable for the aggregated traffic. In this section the impact of the uncorrelation structure of $M/G/\infty$ model will be studied. Fig. 3 already illustrates that smaller access bandwidth makes the variance of traffic rate smaller at time scales beyond the boundary time scale. It is natural to expect some performance gain from it.

The same simulation scenario in Fig. 1 is applied and the performance of the buffer in front of Server B (the backbone link) is looked at. $\rho_a \to 0$ is taken since it

leads to the largest traffic variability in comparison to $\rho_a > 0$ (cf. Fig. 2) so represents the worst performance case. In Fig. 5 the mean waiting time is plotted with respect to ρ_b. The complementary cumulative distribution function (CCDF) of queueing length under the condition $\rho_b = 0.9$ is shown in Fig. 6.

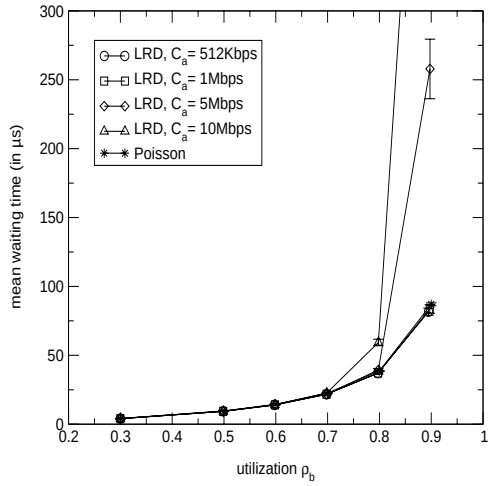

Figure 5: *Mean packet waiting time vs. utilization, wrt. LRD traffic of different boundary time scales*

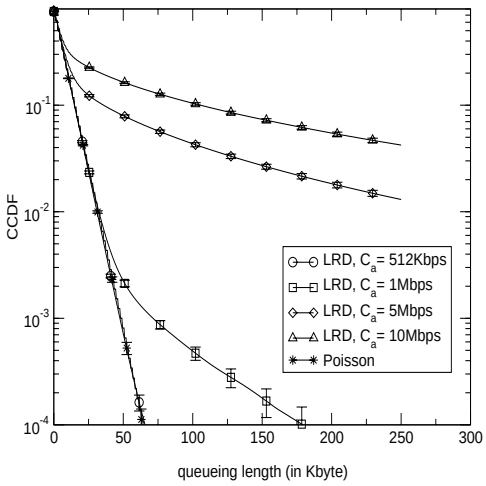

Figure 6: *CCDF of queueing length wrt. LRD traffic of different boundary time scales*

In both diagrams the performance gain due to lower access bandwidth can be figured out. Especially, with the access bandwidth of 512 Kbps the performance curves overlap with those of the Poisson packet traffic. In Fig. 6, particularly, breakpoint can be identified in CCDF curve for the aggregated LRD traffic, corresponding to the observed time-scale-dependent traffic behavior in Section 2. At small queueing lengths, the CCDF is analogous to that of Poisson traffic and decreases exponentially fast with the increase of queueing length. When the queueing length gets further larger, the CCDF begins to decline slowly, consistent with the known queueing behavior of LRD traffic [4]. The location of breakpoint, which corresponds to the efficient queueing area, again depends on the access bandwidth. Small access bandwidth results in the occurrence of breakpoint at a large queueing length.

This phenomenon can be well explained by the concept of relevant time scale (or cutoff lag). For small queueing length whose correspondent relevant time scale is smaller than the boundary time scale, the queueing performance is mainly determined by the traffic characteristic below this time scale, i.e., the uncorrelation traffic behavior. That leads to the exponential decay of the CCDF curve. For large queueing length, the relevant time scale of which lies above the boundary time scale, the LRD traffic characteristic plays a crucial role in affecting the queueing performance and the buffer efficiency decreases significantly. As a result, the

queueing length at which the breakpoint is located increases with the boundary time scale.

Since large boundary time scale corresponds to the small access bandwidth, the performance gain presented above can be also thought of as a kind of multiplexing gain, in the sense that with given backbone bandwidth it is more beneficial to aggregate more users and for each user allocate a relative small access bandwidth. It actually highlights that by keeping a sufficient multiplexing degree the negative impact from the LRD property on the queueing performance is limited, although it is known that the LRD characteristic itself is not alleviated by the multiplexing. This is a significant indication for practical traffic engineering, which is also implied in [2][4], however, from other perspectives.

4. CONCLUSION

Motivated by the new measurement results of Internet traffic, we review the well-known infinite source Poisson traffic model and inspect the cause of the uncorrelation structure of backbone IP traffic observed at the small time scales. It is shown by simulation and analysis that the uncorrelation is directly related to the multiplexing of a large number of independent traffic flows and it dominates in a time scale range upper-bounded by the maximal packet transmission time on the access link.

Corresponding to the boundary time scale distinguishing the different traffic behaviors over different ranges of time scales, the resulting queueing behavior also turns out to be different with respect to different queueing length. A large boundary time scale, or equivalently small access bandwidth, leads to a larger degree of buffer efficiency which brings substantial performance improvements.

Our future work will aim to build explicit mathematical relation between the boundary time scale and the location of the breakpoint (cf. Fig. 6) of queueing performance with respect to CCDF, which will be quite instructive for the resource allocation of network devices.

BIBLIOGRAPHY

[1] A. Broido, R. King, E, Nemeth: “Radon spectroscopy of packet delay.” *Proceedings of ITC 18*, Berlin, Sep. 2003, pp. 419-428.

[2] J. Cao, W. S. Cleveland, D. Lin, D. X. Sun: “Internet traffic tends to Poisson and independent as the load increases.” Tech. Rep., Bell Labs, 2001.

[3] M. E. CROVELLA, A. BESTAVROS: “Self-similarity in World Wide Web traffic: evidence and possible causes.” *IEEE/ACM Transactions on Networking*, Vol. 5, No. 6, Dec. 1997, pp. 835-846.

[4] A. ERRAMILLI, O. NARAYAN, W. WILLINGER, "Experimental queueing analysis with long-range dependent packet traffic," *IEEE/ACM Transactions on Networking*, Vol. 4, 209-223, 1996.

[5] A. Feldmann, A. C. Gilbert, W. Willinger, T. G. Kurtz: "The changing nature of network traffic: scaling phenomena." *Computer Communication Review*, Vol. 28, No. 2, April 1998.

[6] A. FELDMANN, A.C. GILBERT, W. WILLINGER: "Data networks as cascades: investigating the multifractal nature of Internet WAN traffic." *ACM Computer Communication Review*, Vol. 28, Sept. 1998, pp. 42-55.

[7] M. Grossglauser, J. Bolot: "On the relevance of long-range dependence in network traffic." *IEEE/ACM Transactions on Networking*, Vol. 7, No. 5, Oct 1999, pp. 629-640

[8] M. Izal, J. Aracil: "On the influence of self-similarity on optical burst switching traffic." *Proceedings of IEEE GLOBECOM 2002,* Taipei, Nov. 2002

[9] W. E. LELAND, M. S. TAQQU, W. WILLINGER, D. V. WILSON: "On the self-similar nature of Ethernet traffic (extended version)." *IEEE/ACM Transactions on Networking*, Vol. 2, No. 1, Feb. 1994, pp. 1-15.

[10] A.L. NEIDHARDT, J.L. WANG: "The concept of relevant time scales and its application to queueing analysis of self-similar traffic." *Proceedings of SIGMETRICS '98/PERFORMANCE '98*, 1998, pp. 222-232.

[11] V. PAXSON, S. FLOYD: "Wide-Area Traffic: The Failure of Poisson Modeling." *Computer Communication Review*, Vol. 24, No. 4, 1994, pp. 257-268.

[12] X. TIAN, J. WU, C. JI: "A unified framework for understanding network traffic using independent wavelet models." *Proceedings of IEEE INFOCOM 2002*, New York City, June 2002.

[13] W. WILLINGER, M. S. TAQQU, R. SHERMAN, D. V. WILSON: "Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level." *IEEE/ACM Transactions on Networking*, Vol. 5, No. 1, Feb. 1997, pp. 71-86.

[14] W. WILLINGER, V. PAXSON, R. H. RIEDI, M. S. TAQQU: "Long-range dependence and data network traffic." *Long-range dependence: theory and applications*, P. Doukhan, G. Oppenheim, M. S. Taqqu (Eds.), Birkhauser, Sep. 2002.

[15] Z. ZHANG, V. J. RIBEIRO, S. MOON, C. DIOT: "Small-time scaling behaviors of Internet backbone traffic: an empirical study." *Proceedings of IEEE INFOCOM 2003*, San Francisco, March 2003.

[16] http://www.caida.org/analysis/AIX/plen_hist/