# Long-Range Dependence of Internet Traffic Aggregates

Solange Lima, Magda Silva, Paulo Carvalho, Alexandre Santos, and Vasco Freitas

Universidade do Minho, Departamento de Informatica,
4710-059 Braga, Portugal
{solange, paulo, alex, vf}@uminho.pt

**Abstract.** This paper studies and discusses the presence of LRD in network traffic after classifying flows into traffic aggregates. Following DiffServ architecture principles, generic QoS application requirements and the transport protocol in use, a classification criterion of Internet traffic is established. Using fractal theory, the resulting traffic classes are analysed. The Hurst parameter is estimated and used as a measure of traffic burstiness and LRD in each traffic class. The traffic volume per class and per interface is also measured. The study uses real traffic traces collected at a University of Minho major backbone router in different periods of network activity.

## 1 Introduction

The diversity of quality of service (QoS) requirements of the actual and emergent services will force the network to differentiate traffic so that an adequate QoS level is offered. One of the most promising solutions proposed by the Internet Engineering Task Force (IETF) is the Differentiated Services architecture (DiffServ) [1], which aggregates traffic in a limited number of classes of service according to QoS objectives. This new network traffic paradigm poses renewed interest and challenge to network traffic analysis and characterisation. Although, several other studies focus on general Internet traffic characterisation, the effects of aggregating traffic in classes are still unclear. Will a particular traffic class be responsible for the behaviour expressed in [2]? Does aggregation affect burstiness at network nodes and links? The major objective of our work is to study fractal properties such as the long-range dependence (LRD) in Internet traffic aggregates.

Netflow[3] traffic samples collected at different time periods of network activity in a backbone router at the University of Minho were used. After establishing a traffic classification criterion based on a DiffServ model multi-field approach, all the samples are analysed applying that criterion. The time characteristics of each traffic class are studied resorting to the Mathematica software.

## 2 The DiffServ model

In the DiffServ model, network traffic is classified and marked using the DS-field [11]. This identifier determines the treatment or Per-Hop-Behaviour (PHB) [4, 5] traffic will receive in each network node. The IETF has proposed the Expedited Forwarding PHB [4] and the Assured Forwarding PHB Group [5] (EF and AF PHBs), besides best-effort (BE PHB). The EF PHB can be used to build services requiring low loss, reduced delay and jitter, and assured bandwidth. The AF PHB group, consisting of four classes, can be used to build services with minimum assured bandwidth and different tolerance levels to delay and loss.

## 3 Network traffic characterisation

The knowledge of the network traffic characteristics as a whole and, in particular, of traffic aggregates is relevant to allow a proper network resources allocation and management, to help traffic engineering, traffic and congestion control, and to specify services realistically. In our study, the analysis is based on the fractal time series theory since recent studies related to the characterisation and modelling of network traffic point to the presence of self-similarity and LRD. This last property may directly affect the items highlighted above, with strong impact on queuing and on the nature of congestion [6].

### 3.1 Fractal traffic properties

Self-similarity expresses the invariance of a data structure independently of the scale that data is analysed. From a network traffic perspective, self-similarity expresses a new notion of burstiness, i.e. there is no natural length for a burst and bursty structure of traffic is maintained over several time scales.

As an example of processes which exhibit self-similarity and LRD, one may consider $X(t)$ , an **asymptotically second order self-similar** stochastic process, with Hurst parameter $\frac{1}{2} < H < 1$ , i.e., $\lim_{m\to\infty} \gamma(k) = ((k+1)^{2H} - 2k^{2H} + (k-1)^{2H})\frac{\sigma^2}{2}$. $X(t)$ has the following properties: *long-range dependence* - the autocorrelation function $\rho(k)$ decays hyperbolically ($\rho(k)$ is non-summable) $\lim_{k\to\infty} \frac{\rho(k)}{ck^{-\beta}} = 1$; *slowly decaying variances* - the variance of the aggregated series processes $X^{(m)}$, $X_k^{(m)} = \frac{1}{m}\sum_{i=km-m+1}^{km} X_i(k = 1, 2, ...$ and $m = 1, 2, ...)$, is expressed by $var(X^{(m)}) \sim var(X)m^{-\beta}$, with $c > 0$ constant, $\beta = 2 - 2H$, $0 < \beta < 1$.

H is commonly used to measure LRD, and a valuable indicator of traffic burstiness (burstiness increases with H). If $\frac{1}{2}$ <$H$< 1 then an infinite persistence (indicating LRD) can be noticed. If 0 <$H$< $\frac{1}{2}$ then no-persistent behaviour occurs, whereas if $H$= $\frac{1}{2}$ the variables are independent.

There are several methods to estimate the H parameter [7, 2]. While methods such as the test of variances, the R/S statistic or the periodogram are based on graph analysis, the Whittle's estimator provides an analytical method to estimate H. Although the limitation of working with finite data samples and the error probability associated with graph based methods, they are widely used. The test of variances, which was used in this study, is based on the slowly decaying variance property, and H is obtained by $H = 1 - \frac{\beta}{2}$ resorting to a log-log plot of $(var(X^m), var(X)m^{-\beta})$.

### 3.2 Collecting and preparing traffic samples

Traffic samples were collected from a major backbone router located at the Department of Informatics in University of Minho, using Cisco NetFlow tool [3]. NetFlow considers a flow as a unidirectional stream of packets from a source to a destination and records in each entry timing information, fields such as the source and destination IP addresses, port numbers, the protocol identifier, the input and output interfaces, and the number of packets and bytes sent.

The collection of traffic was carried out along different time periods and several days. These time periods were chosen reflecting typical levels of network activity (low: from 2 a.m. to 3 a.m.; medium: from 1 to 2 p.m. and from 10 to 11 p.m.; high: from 10 to 11 a.m. and from 3 to 4 p.m.). Each one-hour traffic sample is filtered by output interface according to the classification criterion presented in section 3.3 for different time intervals (100ms, 500ms, 1s and 10s). This process resulted in around 150 sample sets for analysis.

### 3.3 Traffic Classification Criterion

Due to economical and technical reasons the definition of a traffic classification criterion is a subjective task. For instance, for identical traffic types, a client may be willing to pay more than other to obtain a better service quality. Moreover, when a criterion is based on TCP/UDP/IP packet headers both packet fragmentation[1], packet encryption and the use of negotiated or unregistered application ports difficult classification[2]. Therefore, a classification criterion should be generic enough to be easily adopted and implemented. Most of the criteria suggest distinct classes for UDP and TCP traffic so that non-reactive and reactive applications do not compete for the same resources. Some go further suggesting that the duration of flows, the transmission rate and packet size characteristics should also be considered [9]. A classification method based on QoS application requirements such as delay or loss sensitivity is also proposed [10].

Considering the aspects above and the Type of Service (ToS) proposed for classical applications [10], our classification criterion is oriented to traffic aggregation which can easily be mapped to a class-based QoS architecture. As a first

[1] While fragmentation of UDP traffic is increasing, TCP traffic (around 85% of Internet traffic [8]) is virtually not fragmented due to the widespread use of MTU path discovery techniques and relatively small default packet sizes.

[2] The use of a modified IP Encapsulating Security Payload (which leaves protocol ports unchanged), the analysis of traffic at the control channels (which uses well-known ports) and/or the applications' usual range of ports or addresses might be possible solutions.

approach, the classification process distinguishes TCP from UDP traffic, and then, the generic applications requirements are taken into account. A filtering process based on more detailed rules to differentiate specific or proprietary applications (e.g. NetMeeting, Cisco IP/TV, and many other unicast or multicast applications) was left for further study. The resulting traffic classes are:

- **Class 1** - delay sensitive TCP traffic, resulting from interactive protocol applications such as Telnet, SSH or FTP control;
- **Class 2** - loss and throughput sensitive TCP traffic, resulting from bulk transfer protocol applications such as FTP data, DNS zone transfers, SMTP, POP, IMAP, NNTP;
- **Class 3** - essentially, HTTP traffic. Other TCP traffic not included in classes 1, 2 and 4 (a reduced volume) is mapped to this class;
- **Class 4** - priority traffic e.g. routing or management protocols (TCP/UDP);
- **Class 5** - generic UDP traffic e.g. TFTP, DNS, POP, IMAP or HTTP/UDP;
- **Class 6** - traffic from applications using transient ports (not allowing their classification) or UDP ports not covered by classes 4 and 5[3].

There is not a direct mapping between the defined classes and the DiffServ PHBs. Such mapping would depend on the administrative and contractual service policies. However, a possible match could be: classes 1 and 3 supported by high priority AF PHBs; class 4 by EF PHB; classes 2 and 5 by BE or low priority AF PHBs; class 6, could be either EF or AF depending on the relevance given to the diversity of applications.

## 4 Statistical data analysis

### 4.1 Traffic volumes

| class | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| bytes | 2.31 % | 26.40 % | 67.04 % | 0.01 % | 1.67 % | 2.67 % |
| packets | 2.37 % | 18.95 % | 68.41 % | 0.04 % | 5.94 % | 4.27 % |

**Table 1.** Traffic volume per class

Table 1 presents the percentage of traffic each class contributes for the total load in the router. The results show that the only class which contents is not clearly identified (Class 6) represents a small amount of traffic[4], which shows

[3] According to [8], traffic from Real Audio and online game applications is likely the most significant one can consider in this class. Other real-time unicast/multicast traffic is also included here.

[4] For this reason, the statistical analysis of this class will be considered in a later stage of our study.

the broadness of the classification criterion proposed. As expected, classes 2 and 3, including mainly bulk and web traffic respectively, contribute heavily to the global router load. The results also show a correlation between the percentages of packets and bytes for all classes, although, classes 2 and 5 denote the presence of large and small packet sizes, respectively.

### 4.2 Testing long-range dependence

In order to analyse statistically whether a particular traffic class exhibits LRD tests of variance and autocorrelation analysis were carried out. For most of the samples, these tests illustrate similar results both for the analysis of the time series of packets and bytes.

| | Perc. of Samples | | | | | Traffic Volume | | | |
|---|---|---|---|---|---|---|---|---|---|
| classes | a) | b) | c) | d) | | a) | b) | c) | d) |
| 1 | 76.5% | 0.0% | 17.6% | 5.9% | | 99.3% | 0.0% | 0.6% | 0.1% |
| 2 | 50.0% | 6.7% | 30.0% | 13.3% | | 37.5% | 3.3% | 43.7% | 15.5% |
| 3 | 17.2% | 6.9% | 20.7% | 55.2% | | 4.9% | 6.6% | 7.5% | 81.0% |
| 4 | 91.7% | 8.3% | 0.0% | 0.0% | | 98.9% | 1.2% | 0% | 0.0% |
| 5 | 40.9% | 18.2% | 22.7% | 18.2% | | 37.0% | 30.4% | 19.0% | 13.7% |

**Table 2.** Percentage of samples and traffic volume with: a) $H < 0.45$, b) $0.45 \leq H \leq 0.5$, c) $0.5 < H < 0.7$, d) $H \geq 0.7$.

For each class, Table 2 shows the percentage of samples and the corresponding traffic volume for H within specific intervals when submitted to the variance analysis. The analysis of these tables demonstrates that distinct classes can behave very differently. Note that most of the class 3 traffic volume is included in these samples. Class 3 (HTTP traffic) is clearly the one showing the higher degree of burstiness. Most of the samples (76%) exhibit an $H > 0.5$, and 55% of them have an $H > 0.7$. Table 3 extends this analysis to the activity periods defined in section 3.2. Similar analysis was also carried out for the other classes.

| | Perc. of Samples | | | Traffic Volume | |
|---|---|---|---|---|---|
| Network activity | $0.5 < H < 0.7$ | $H \geq 0.7$ | | $0.5 < H < 0.7$ | $H \geq 0.7$ |
| High | 14.3% | 78.6% | | 4.0% | 95.5% |
| Medium | 33.3% | 66.7% | | 1.6% | 98.4% |
| Low | 22.2% | 11.1% | | 25.6% | 14.3% |

**Table 3.** Class 3: Percentage of samples and traffic volume with an estimate $H > 0.5$.

It is notorious that H increases with network activity, which is consistent with [7]. Although the above tables do not differentiate the results by interface, the analysis of classified traffic per interface shows the same tendency. Excluding class 2, the relation between H and the traffic volumes is not clear for the remaining classes which may indicate an application type dependence. In fact, class 1 behaves in opposite way, and class 4 does not show evidence of burstiness for the different activity periods. This can be due to the regular nature of traffic it emcompasses (e.g. routing traffic). As regards the autocorrelation, almost all the samples with autocorrelation functions decaying slowly to zero (which suggests LRD) had an H above 0.5 in the variance time plots, which is consistent.

## 5 Conclusions

In this study, the Hurst parameter was used to measure LRD in real traffic samples classified according to a proposed criterion. The values for H were determined for the defined traffic classes and for periods of different network activity.

The results show that classes 2 and 3 (bulk transfer and HTTP traffic) play a major role in the total load per interface. In particular, Class 3 is clearly the one showing a higher evidence of burstiness, which increases with traffic load. While Class 2 has similar characteristics, Class 1 behaves in opposite way. Class 4 presents an estimated H below 0.5 independently of the network activity period.

For most of the samples, the tests illustrate similar results either analysing the time series of packets or bytes.

Currently, a larger set of samples are being analysed to consolidate these results. Obtaining complementary statistics is also a matter of concern.

## References

1. S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss. An Architecture for Differentiated Services. Technical report, IETF RFC 2475, 1998.
2. M.S. Taqqu, W. Willinger, W.E. Leland, and D.V. Wilson. On the Self-Similar Nature of Ethernet Traffic. *SIGCOMM'93*, 1993.
3. Cisco Systems. NetFlow. http://www.iagu.on.net/software/netflow-collector.
4. V. Jacobson, K. Nichols, and K. Poduri. An Expedited Forwarding PHB. Technical report, IETF RFC 2598, 1999.
5. J. Heinanen, F. Baker, W. Weiss, and J. Wroclawski. Assured Forwarding PHB Group. Technical report, IETF RFC 2597, 1999.
6. A. Erramilli and W. Willinger. Experimental Queueing Analysis with Long-Range Dependent Packet Traffic. *IEEE/ACM Trans. on Networking*, 4(2), April 1996.
7. M.S. Taqqu, V. Teverovsky, and W. Willinger. Estimators for Long-Range Dependence: An Empirical Study. *Fractals*, 3:785...788, 1995.
8. Trends in Wide Area IP Traffic Patterns. http://www.caida.org/outreach/papers/.
9. A. Bak, W. Burakowski, F. Ricciato, S. Salsano, and H. Tarasiuk. Traffic Handling in AQUILA QoS IP Networks. *QoFIS2001*, page 243...260, September 2001.
10. A. Croll and E. Packman. *Managing Bandwidth: Deploying QoS in Enterprise Networks*. Prentice Hall, 2000.