

A Statistical Analysis of Network Parameters for the Self-management of Lambda-Connections

Tiago Fioreze¹, Lisandro Granville², Ramin Sadre¹, and Aiko Pras¹

¹ University of Twente

Design and Analysis of Communication Systems (DACs)

Enschede, The Netherlands

{t.fioreze,r.sadre,a.pras}@utwente.nl

² Federal University of Rio Grande do Sul,

Institute of Informatics

Porto Alegre, Brazil

granville@inf.ufrgs.br

Abstract. Network monitoring plays an important role in network management. Through the analysis of network parameters (*e.g.*, flow throughput), managers can observe network behavior and make decisions based on them. The choice of network parameters although should be relevant for each specific objective. In this paper, we focus on the analysis of network parameters that are relevant for our self-management of lambda-connections proposal. This proposal consists of an automatic decision process to offload large IP flows onto lambda-connections. This paper aims at statistically analyzing a list of potential network parameters as relevant estimators for flow volume. The main contribution of this work is the introduction of a statistical methodology to validate that some few network parameters can be considered as good predictors for flow volume. These predictors are therefore of great interest to be used in our automatic decision process.

1 Introduction

Recently, a clear change in the set of core technologies that form the Internet is being observed. Internet backbones that once relied solely on IPv4 routing to deliver end-to-end communications are moving towards hybrid solutions. One example is the rising of hybrid optical switching and packet forwarding networks. These hybrid networks are composed of intermediate devices that are both switches at the optical level and traditional routers at the network level. In such an environment, network flows can traverse a hybrid network through either an optical path or a chain of routing decisions. In this hybrid environment, moving large flows from the IP level to the optical level presents the following advantages: (a) flows experience faster and more reliable transmissions with optical switching than with traditional IP routing; (b) remaining flows at the IP level also experience better services because the network layer is less congested after offloading these large flows.

The decision today to select which flows will be placed in which level (optical or network) is still up to network human operators. However, this is not a trivial decision to be made. Operators have to cope with several network parameters (*e.g.*, flow throughput,

flow duration) to come up with a balanced plan. The choice of a proper set of network parameters is therefore crucial for this task. Besides, network operators have to select flows in a timely manner, since it is far cheaper to send traffic at the optical level than at the IP level, as pointed out by Cees de Laat's research [1].

A new approach for the management of hybrid networks has been investigated by us [2] to speed up the process of flow selection. Our approach, called self-management of hybrid networks, consists of monitoring a network of interest to automatically decide which IP flows should be transported at the optical level and which other flows should remain at the network level. The main targets of our approach are flows that despite being small in number are responsible for most of the IP traffic, *i.e.*, the so called *elephant flows* [3] [4]. However, to properly select those flows, our self-management approach needs to employ a set of parameters while analyzing network data.

The goal of this paper is to present a study on selecting a proper set of potential parameters to be used as good predictors for the traffic volume generated by *elephant flows*. Volume prediction is employed in our self-management approach to reduce the time (and consequently cost) of a high-volume flow staying at the IP level consuming network resources. The faster a high-volume flow is detected, the lesser the network resources it consumes. In this context, the main research questions that motivate the investigation presented in this paper are: *Which network parameters could be used to predict the volume of flows?*, and *How can these parameters be evaluated?*

In order to answer these questions, we have first carried out a study on the literature about the parameters (*e.g.*, throughput, duration) that can be observed using current management technologies (*e.g.*, NetFlow, SNMP). To find out how accurate each parameter is, we collected and analyzed network traces from the University of Twente (UT) network. From the collected traces, those flows that have been classified as elephant ones have been filtered and analyzed. We then propose the employment of statistical techniques, such as correlation and tree classification, as a methodology to evaluate which network parameters are more relevant to predict flow volume. Such a methodology is therefore the main contribution of this paper.

The remainder of paper is structured as follows. In Section 2 we review the current research work on traffic characterization. In Section 3 we describe our methodology by presenting the decisions and steps taken to make our statistical analysis. In Section 4 we present the selected parameters. Finally, we close this paper in Section 5, where we draw our conclusions and future work.

2 Related Work

Investigations about flow characteristics have been carried out by researchers for some years already. Thompson *et al.* [5] present a flow-based characterization of network usage and workloads on a commercial backbone. The authors analyzed traffic data collected at one observation point on different time scales. Recently, Kim *et al.* [6] presented a detailed analysis of flow-based traffic characteristics. The metrics that have been considered were packets, bytes, and port distribution. It is interesting to mention that these two papers did not present an extensive reasoning about the chosen metrics.

More recently, Ribeiro *et al.* [7] used packet sampling as the measurement technique while mainly observing its effect on the flow size distribution. They also observed the

effects on the packet counts, SYN information, and sequence number information. They concluded that TCP sequence numbers are essential for accurate flow size estimation, but no conclusions have been drawn about the best estimators.

We believe these works, both historical analysis of traces and comparison of diverse observation points, even if suitable for highly detailed studies, are missing one important dimension of analysis: they focus on estimating precise flow size distribution or packet distribution, but none of them focus on the best parameters to predict flow size.

3 Methodology

This section first presents our approach to create the list of potential network parameters. After that, the measurement set up to collect network traces is introduced. Finally, we explain the statistical techniques we used to evaluate the potential parameters.

3.1 List of Potential Network Parameters

Network parameters provide valuable information about the status of network traffic and devices (*e.g.*, routers, switches). In the context of our self-management approach, network parameters are important to predict the volume of each flow. This prediction is used to decide which flows should be moved to the optical level and which others should stay at the network level.

To define a set of relevant parameters for our autonomic decision process, a research on the literature has been carried out to list candidates parameters. Such candidates parameters have been taken from the following sources: (a) MIB modules MIB-II [8], RMON-2 [9], and SMON [10]; (b) the information model for the IP Flow Information eXport (IPFIX) protocol [11]. Since these sources also deal with information other than related to flows (*e.g.*, MAC to IP address translation in MIB-II), we have intuitively selected the subset of information that are helpful in predicting the volume of flows. The outcome of this research is the classification of network parameters divided in two main groups: flow identifier parameters, and flow behavior parameters.

Flow Identifier is a set of fields that defines groups of packets (flows) that share some common fields. Since the number of fields in a flow can be extensive [11], we limited the fields to those relevant to our approach:

1. *TCP/UDP port numbers* represent the communication end points that network applications use to exchange data via transport protocols (*i.e.*, TCP and UDP). Ports are important because some applications can generate more traffic than others. For example, a Telnet session is expected to generate far less data than an FTP session.
2. *IP addresses* identify network devices belonging to a certain network. Following the same reasoning above, some devices can generate more traffic than others. For example, a file server is expected to generate more traffic than an ordinary desktop.
3. *Network segments* are portions of computer networks that vary in size from small networks (*e.g.*, LAN) to large ones (*e.g.*, WAN). Some network segments can generate more traffic than others. For example, the University of Groningen network receives lots of data from the LOFAR [12] sensor network segment. The parameters that identify network segments are:

- *Subnet*: Continuous bits in an IP address prefix used to identify a subnet;
 - *Autonomous System*: A collection of IP routing prefixes.
4. *Protocols* allow the communication between end points on top of IP. TCP and UDP are the most important Internet protocols. Several other protocols exist, but few of them have a representative significance in terms of traffic [13]. The value of the protocol number in the IP header is the parameter we consider.
 5. *Type of Service (ToS)* is a 8 bits portion of the IP header that is reserved to define a service level request. Even though the support for ToS is not widely employed in current routers and therefore not widely used, it can represent a potential network parameter to track flows that generate considerable amount of traffic.

Flow Behavior is a set of network parameters used to characterize the behavior of flows. The potential flow behavior parameters that are relevant to our research are:

1. *Duration*: literature [14] [15] [16] has shown that some large flows may also be long in duration, normally presenting a heavy-tail distribution. Duration can therefore be a potential input parameter for our automatic decision approach.
2. The number of *packets* of a flow can give a good indicative about a flow's behavior.
3. The number of *bytes* of a flow is naturally an important parameter to be considered, since the automatic decision module aims at offloading high volume flows.
4. *Throughput* is the average rate of a communication. It is usually expressed in bytes per seconds (Bps), but it can also be measured in packets per second (Pps). Those two throughput units of measurement will be considered.

It is worth mentioning that analysis on flow identifiers has been previously carried out by us [17] [18] [19], and for the sake of space it will not be included in this paper. We thus focus on the analysis of the flow behavior group only.

3.2 Collecting Network Traces

This subsection shows how network traces have been collected in order to evaluate the set of parameters to describe flow behaviors. The collection process consisted in collecting NetFlow data from the UT NetFlow-enabled router running NetFlow version 9. This router exported NetFlow records to a flow collector hosted in our department. Traces from the UT network have been collected over a period of one day (Sep 18th, 2008). Once the traces were collected, they needed to be combined for our analysis. Since NetFlow reports long-lived flows in different records, one needs to combine the NetFlow records in order to closely compute the original flow duration, number of packets, and octets. The throughputs (in Bps and Pps), however, are calculated as an average throughput over the entire duration of the flow.

In order to combine NetFlow records, there is a need to determine the maximum gap that separates two consecutive flow records of the same flow. We have deliberately chosen a gap of 30 seconds, which is a common value for the TCP TIME-WAIT state. We then decided that all NetFlow records of the same flow whose gap was smaller or equal to 30 seconds are grouped into the same flow. Our whole analysis has been made then over combined flows rather than using the original NetFlow records.

Once the NetFlow records have been combined into flows, we stored the flows in a format suitable for our analysis. In our case, we imported the flows into a MySQL database. MySQL has been chosen due to the familiarity of the paper's authors with such a tool. The amount of collected NetFlow records stored in MySQL accounted for 30.51 GB, plus 37.28 GB of indexes to speed up our analysis. Once combined, flows accounted for 26.08 GB plus 37.24 GB of indexes.

3.3 The Steps of Our Statistical Analysis

The statistical analysis of the potential network parameters could have been performed in different ways. Simulation tools, for example, could be employed to reproduce a network being measured. A controlled environment of a lab network could be used too. However, none of these methods can 100% capture the real behavior of flows. Considering that, we believe that statistically analyzing data collected from real networks would provide more significant and relevant conclusions.

The initial number of flows considered in our statistical analysis was 378,363,608, which generated a volume of 18.11 TB on Sept 18th, 2008. We started our statistical analysis by defining the set of flows we are focused on. This set of flows is based on the target our self-management of lambda-connections aims at, *i.e.*, the *elephant flows*. We focus our analysis therefore on flows that have the following characteristics: (a) few in number, (b) persistent in time, and (c) represent most of the traffic.

Out of the total number of collected flows (378,363,608), a very small percentage of the flows (0.82%) accounted for the biggest percentage (97%) of the total traffic. This percentage of flows (3,092,885 in numbers), from now on addressed as big flows, matches the characteristics (a) and (c) previously mentioned. Besides, it also confirms the long tail distribution on the network traffic, showing that few flows are responsible for most of the traffic.

Our next step was to check, out of the big flows, the ones persistent in time. Persistent means that they do not have a short duration, as it is the case of bursty flows [6]. Table 1 shows some statistics about the duration of our big flows. It also shows that most of these flows (75%) have a considerable short duration, *i.e.*, a duration shorter than 57 seconds. This allows us to conclude that most of the flows are short-lived.

Conditional Probability

Statistics about flow duration do not say much about the persistence in time of the other 25% of the flows. To have a better insight about the persistence of these flows,

Table 1. Statistics of duration of the big flows

Flows	3,092,885
Mean	274 sec
Median	19 sec
Minimum	0 sec
Maximum	86,507 sec
Percentiles (25%)	12 sec
Percentiles (50%)	19 sec
Percentiles (75%)	57 sec

more specifically about when the majority of them tend to get stable regarding to their duration, we used conditional probability [20]. Conditional probability is the probability of some event A happening given the occurrence of some other event B ($P(A | B)$).

In our analysis, we are interested in knowing the probability of flows being persistent in time. For that, we observe the conditional probability of flow duration as follows. Given that the duration (D) of a flow has already lasted at least certain amount of time B ($D \geq B$), what is the probability this flow will last for at least one more minute ($D \geq B + 60 \text{ sec}$)?

Table 2 shows the conditional probability of duration of our big flows expressed in seconds. It shows that there is a small probability (24%) that a flow will last at least one more minute, given the fact it has just started. However, there is a considerable improvement (67%) in this probability when flows have elapsed at least one minute. This probability gets more stable the bigger the minimum flow duration is. This allows us to conclude that the longer a flow has already elapsed, the smaller is its probability of ending in the next time period.

Table 2. Conditional probability of duration of the big flows

$P(D \geq B + 60 \text{ sec} D \geq B)$	Percentage
$P(D \geq 60 \text{ sec} D \geq 0 \text{ sec})$	24%
$P(D \geq 120 \text{ sec} D \geq 60 \text{ sec})$	67%
$P(D \geq 180 \text{ sec} D \geq 120 \text{ sec})$	78%
$P(D \geq 240 \text{ sec} D \geq 180 \text{ sec})$	83%
$P(D \geq 300 \text{ sec} D \geq 240 \text{ sec})$	86%
$P(D \geq 360 \text{ sec} D \geq 300 \text{ sec})$	89%
$P(D \geq 420 \text{ sec} D \geq 360 \text{ sec})$	90%
$P(D \geq 480 \text{ sec} D \geq 420 \text{ sec})$	91%
$P(D \geq 540 \text{ sec} D \geq 480 \text{ sec})$	92%
$P(D \geq 600 \text{ sec} D \geq 540 \text{ sec})$	92%

Applying this finding in our self-management approach results in a trade-off between flow duration and decision time. A decision about selecting a flow to be offloaded to the optical level cannot be taken too soon because there is a high probability that the flow is going to last short. In contrast, this decision cannot either be postponed too long because the flow will be consuming resources at the IP level while the decision is not taken. Moreover, there is a slight chance that the flow may end when the decision is finally made.

We choose therefore an elapsed duration of 5 minutes to define a flow as being persistent in time. The reason for that comes from the fact the percentage of flows lasting for at least another minute gets relatively stable (around 90%) when flows reach a minimum duration of 5 minutes. Moreover, in our analysis, flows with duration below 5 minutes did not represent a considerable amount of traffic (26% of the total traffic), whereas flows with duration above or equal to 5 minutes represented 74% of the total traffic.

Based on the outcome of the conditional probability applied on our traces, another filtering was done in the 3,092,885 flows in order to remove flows with a duration

shorter than 5 minutes. This resulted in the selection of 283,783 flows (0.07% of the total number of collected flows) having duration above or equal to 5 minutes. These are thus the flows whose metrics throughput (Pps and Bps), packets, and duration are observed in relation to the volume generated.

Correlation and Classification Tree Techniques

Once we defined our set of flows to be analyzed, we observe how those metrics relate with flow volume by drawing correlation charts. These charts provide us a visual impression of what the correlation is, although they are not very quantitative. In order to obtain quantitative correlation values regarding the considered metrics, we used Pearson's correlation method [21]. Pearson's correlation computes the pairwise associations for a set of variables and displays the results in a matrix. This is useful for determining the strength and direction of the association between two metrics. It can be used therefore to measure the linear association between two metrics. In our context, there is an unmistakable intuition that Bps and duration have the strongest correlation to tell about the volume of flows. We use Pearson's correlation, however, to see if this fairly obvious correlation may contain some other unsuspected correlations. Even though we suspected that there are some correlations, we did not know which are the strongest. Applying therefore correlation analysis on our data can lead to a better understanding.

Even though Pearson's correlation is a good method to find out the relationship between two metrics only, it does not consider the combination of more than two of them. It could be therefore that an interaction of more than two metrics could give a better refinement about the best predictors for a flow volume. In order to find that out, we used a classification tree technique widely used in data mining areas, called *CHI-squared Automatic Interaction Detector* (CHAID) method. CHAID is a method that divides a data set into exclusive and comprehensive partitions that differently relate with an observed dependent variable [22]. These partitions are defined by a tree structure and they are classified in descendent order of independent variables, called *predictors*. For each partition of predictors, CHAID assigns a probability of response. All probabilities are subsequently used to rank the partitions with the strongest relation with the dependent variable. It is worth mentioning that the partitions of each predictor are merged if they are not significantly (significance level of 0.05%) different in regard to the dependent variable.

In the case of our analysis, CHAID calculates which independent metrics – duration, packets, Pps, and Bps (the predictors) – have the strongest relation with flow volume (the dependent variable). The CHAID outcome and the results of our correlation analysis are presented in the next section.

4 Results

The potential flow behavior parameters have been statistically evaluated by observing how they individually contribute to identify flows that generate large amounts of data. This section starts by presenting the results regarding our correlation analysis, followed then by the CHAID classification tree.

4.1 Volume vs. Considered Metrics

Figures 1, 2, 3, and 4 show the relation of packets, duration, Bps, and Pps with the number of octets (flow volume), respectively. All figures present both axes with logarithmic scale. The figures show that there is a linear relation between packets, Bps, and Pps with octets, but that does not hold in the case of duration. Figure 2 shows that a flow can have a long duration and a small amount of octets, a short duration and a great amount of octets, and all in between. On the other side, the metrics packets, Bps, and Pps walk along with octets in the sense that the bigger those metrics are, the bigger the flow volume is expected to be generated.

It is worth mentioning that there is a certain degree of linearity among those metrics. Figure 1 shows a strong linear relationship between the number of packets and flow volume. Figures 4 and 3, in turn, show a slightly bigger variability in the linearity when compared to packets. The reason for that comes from the fact that flow duration presents a strong variability, which affects Bps and Pps linearity when related to flow volume. For example, we have seen cases in which a lot of packets can be generated in a short amount of time (high Pps), but it is also possible that a small amount of packets can be generated in a longer period (low Pps).

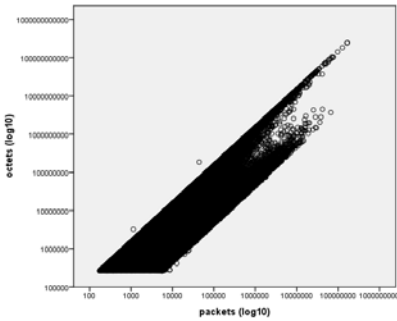


Fig. 1. Correlation between octets and packets

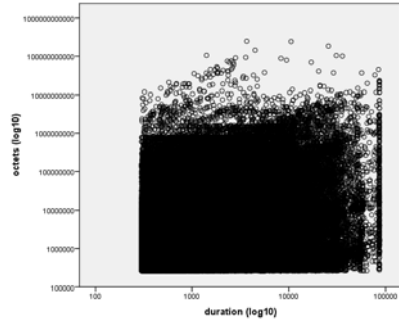


Fig. 2. Correlation between octets and duration

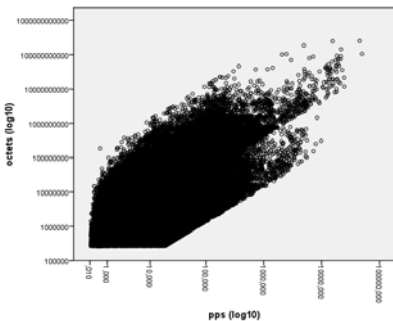


Fig. 3. Correlation between octets and Pps

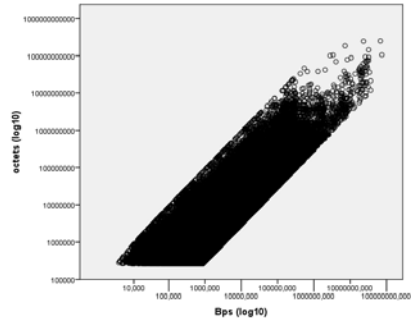


Fig. 4. Correlation between octets and Bps

In order to quantify this linearity, we used then Pearson's r correlation, which is defined as the sum of the products of the standard scores of the two measures divided by the degrees of freedom:

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{s_x} \right) \left(\frac{Y_i - \bar{Y}}{s_y} \right)$$

where $\left(\frac{X_i - \bar{X}}{s_x} \right)$, \bar{X} , and s_x are the standard score, sample mean, and sample standard deviation, respectively [23]. Table 3 shows the r correlation among our considered parameters.

Table 3. Pearson's correlation octets pairwise with other considered parameters

	Packets	Bps	Pps	Duration
Pearson's r correlation for octets	0.927	0.671	0.642	0.058

A correlation of 0 (zero) means that there is no linear relationship between the two variables. On the other hand, a correlation of 1 means that there is a strong positive linear relationship between the two variables. As table 3 shows, there is a strong linear relationship between packets (0.927), Bps (0.671), and Pps (0.642), with octets. On the contrary, duration does not present a strong linear relationship (0.058). This means that duration should not be exclusively focused on when trying to predict a flow volume.

Since Pearson's correlation only shows the relation between two metrics, we used CHAID in order to see the relation of more than two metrics with the flow volume. For that, we first divided flow volume into 4 categories according to their size. The 25% biggest flows are referred as *HIGH* volume, whereas the 25% smallest flows are named *LOW* volume. The medium flow sizes (*i.e.*, the remaining 50%) are then divided into the 25% biggest medium flows, referred as *MEDIUM HIGH* volume, whereas the 25% smallest medium flows are named *MEDIUM LOW* volume. Once the flow volume is categorized, we observe how the considered parameters relate to these 4 categories.

Figure 5 partially shows the result of our CHAID classification tree. Only *nodes 1* and *10* were expanded for the sake of space. Flow volume is the dependent variable and it is the *node 0* in the CHAID tree. *Node 0* contains the 4 aforementioned categories for the flow volume. The CHAID method then starts dividing the predictors (*i.e.*, Bps, Pps, duration and packets) into partitions (*nodes*) and cross-tabulating them against the dependent variable (*node 0*). The predictor that presents the smallest level of significance (*i.e.*, the most statistically significant relationship with the dependent variable) is placed at the first depth of the CHAID classification tree along with its partitions. After the CHAID method has decided about the first level predictor and its best merged partitions, CHAID begins to place other predictors beneath the initial predictor. The CHAID method continues this procedure until further sub-divisions cannot be performed.

CHAID chooses Bps as the best predictor for flow volume as it is ranked right below *node 0*. From *node 1* to *node 10* it is possible to see how each group of flows, classified by their Bps throughput, influences the flow volume. Flows with small Bps throughput tend to be small in flow size (*e.g.*, *node 1*). On the other extreme, flows with big Bps throughput tend to generate large flows (*e.g.*, *node 10*). The second best predictor

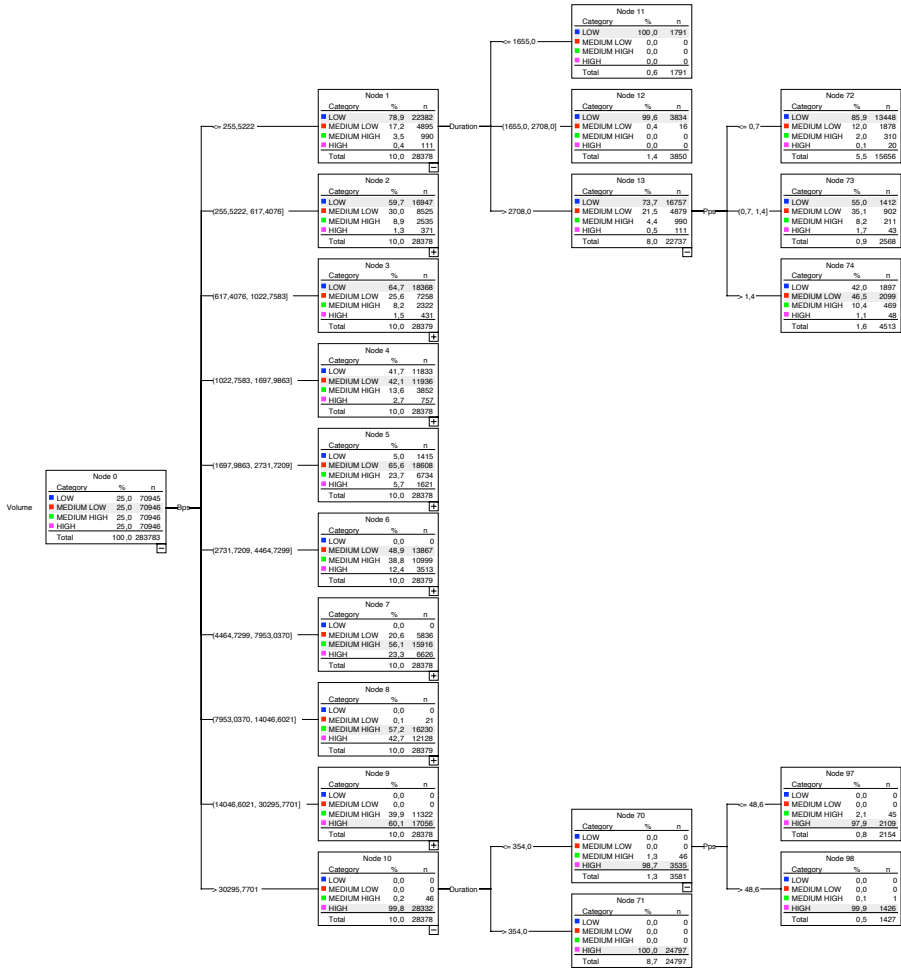


Fig. 5. CHAID classification tree

pointed out by the CHAID method is flow duration. CHAID classification tree shows that flows with high Bps and long durations are those that have the greater flow volume. Finally, for flows with long duration, those with high Pps are responsible for most of the volume generated, being therefore Pps the third best predictor. The number of packets did not show a significant prediction in our model, and it was ignored by the CHAID method.

CHAID also provides a measure of confidence that the classification model is correct as presented in Table 4. Table 4 shows the accuracy of our classification tree model for the 4 observed flow volume categories. It indicates that our classification model correctly classifies about 89% of the flows while misclassifying a flow volume only in about 11% of the cases. That high accuracy allows us to assume that our CHAID classification model correctly selects the proper predictors for the flow volume.

Table 4. The accuracy of our CHAID classification model

Observed	Predicted				Accuracy
	LOW	MEDIUM LOW	MEDIUM HIGH	HIGH	
LOW	65,499	5,446	0	0	92.3%
MEDIUM LOW	6,381	60,334	4,231	0	85.0%
MEDIUM HIGH	521	3,016	62,835	4,574	88.6%
HIGH	63	48	7,257	63,578	89.6%
<i>Overall percentage</i>	<i>25.5%</i>	<i>24.3%</i>	<i>26.2%</i>	<i>24.0%</i>	88.9%

CHAID statistically confirms the intuition that Bps and duration are the metrics to be considered when observing flow volume. In order of importance, Bps, duration, and Pps (as an optional refinement) are the best predictors. These metrics have more impact on the flow volume than others. Our self-management of hybrid networks should therefore take them into account, rather than other metrics, when taking decisions on moving flows to the optical level.

5 Conclusions and Future Work

This paper presented a statistical evaluation of potential network parameters for our self-management of lambda-connections. Two research questions were risen:

Research question 1: *Which network parameters could be used to predict the volume of flows?* The number of potential network parameters is wide. The selection of the parameters shall be done depending on a defined objective. Since we focus on the prediction of the volume of flows, we narrowed down the selection of network parameters, dividing them into two main groups: flow identifiers parameters, and flow behavior parameters. The former was exhaustively researched in previous works of ours [17] [18] [19]. The latter group was the focus of this paper and resulted in the evaluation of the parameters: duration, packets, Bps, and Pps.

Research question 2: *How can these parameters be evaluated?* These metrics were evaluated by using statistics methods. We started with conditional probability to know when most of the flows gets stable regarding to their duration. We found out that flows with a minimum duration of 5 minutes have 89% of chance of continuing running for at least the next minute. Our next statistical step was to find out correlation among the considered parameters. Even though initially some metrics such as duration and Bps were intuitively expected to influence flow volume, little was known about how strong this influence could be. Moreover, we were not aware if there were any other unsuspected parameters (*e.g.*, Pps and packets) that could have significant influence to flow volume. To solve this uncertainty, we used first Pearson's r correlation. Pearson's correlation showed that packets ($r = 0.927$), Bps ($r = 0.671$), and Pps ($r = 0.642$) have a strong linear relationship with flow volume, while duration ($r = 0.058$) has not. Thus, since all parameters have certain influence on the flow volume, they should not be used alone, but in groups. We used then CHAID technique to analyze that. As evaluated by CHAID, Bps and duration are the best predictors for flow volume, followed by Pps.

Even though packets was considered by Pearson's correlation as the parameter with the strongest linear relationship with flow volume, the total number of packets can only be known after the end of a flow, being therefore inadequate to predict flow volume.

As future research, we aim at performing further investigations on Bps and duration parameters. We will investigate how to use them in a tuning process for our decision process to take automatic and online decisions to offload elephant flows.

Acknowledgments. This research work has been supported by the EC IST-EMANICS Network of Excellence (#26854). Special thanks to Roel Hoek (UT) and Anna Sperotto for their valuable contribution in the flow collection process.

References

1. de Laat, C., Radius, E., Wallace, S.: The Rationale of the Current Optical Networking Initiatives. *Future Gener. Comput. Syst.* 19(6), 999–1008 (2003)
2. Fioreze, T., van de Meent, R., Pras, A.: An Architecture for the Self-management of Lambda-Connections in Hybrid Networks. In: Pras, A., van Sinderen, M. (eds.) *EUNICE 2007. LNCS*, vol. 4606, pp. 141–148. Springer, Heidelberg (2007)
3. Mori, T., Uchida, M., Kawahara, R., Pan, J., Goto, S.: Identifying Elephant Flows Through Periodically Sampled Packets. In: *ACM SIGCOMM*, pp. 115–120 (2004)
4. Wallerich, J., Dreger, H., Feldmann, A., Krishnamurthy, B., Willinger, W.: A Methodology For Studying Persistency Aspects of Internet Flows. *Computer Communication Review* 35(2), 23–36 (2005)
5. Thompson, K., Miller, G.J., Wilder, R.: Wide-Area Internet Traffic Patterns and Characteristics. *IEEE Network* 11(6), 10–23 (1997)
6. Kim, M.S., Won, Y.J., Hong, J.W.: Characteristic Analysis of Internet Traffic from the Perspective of Flows. *Computer Communications* 29(10), 1639–1652 (2006)
7. Ribeiro, B., Towsley, D., Ye, T., Bolot, J.C.: Fisher information of sampled packets: an application to flow size estimation. In: *IMC 2006: Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*, pp. 15–26. ACM, New York (2006)
8. Raghunathan, R.: Management Information Base for the Transmission Control Protocol (TCP). RFC 4022 (Proposed Standard) (March 2005)
9. Waldbusser, S.: Remote Network Monitoring Management Information Base Version 2. RFC 4502 (Draft Standard) (May 2006)
10. Waterman, R., Lahaye, B., Romascanu, D., Waldbusser, S.: Remote Network Monitoring MIB Extensions for Switched Networks Version 1.0. RFC 2613 (Draft Standard) (June 1999)
11. Quittek, J., Bryant, S., Claise, B., Aitken, P., Meyer, J.: Information Model for IP Flow Information Export. RFC 5102 (Proposed Standard) (January 2008)
12. Schaaf, K., Broekema, C., Diepen, G., Meijeren, E.: The Lofar Central Processing Facility Architecture. *Experimental Astronomy* 17(1-3), 43–58 (2004)
13. Estan, C.: Internet Traffic Measurement: What's Going on in my Network? PhD thesis (2003)
14. Brownlee, N., Claffy, K.: Understanding Internet Traffic Streams: Dragonflies and Tortoises. *IEEE Communications Magazine* 40(10), 110–117 (2002)
15. Soule, A., Salamatia, K., Taft, N., Emilion, R., Papagiannaki, K.: Flow Classification by Histograms: or How to Go on Safari in the Internet. In: *SIGMETRICS 2004/Performance 2004: Proceedings of the joint international conference on Measurement and modeling of computer systems*, pp. 49–60. ACM, New York (2004)
16. Lan, K., Heidemann, J.: A Measurement Study of Correlations of Internet Flow Characteristics. *Computer Networks* 50(1), 46–62 (2006)

17. Fioreze, T., Wolbers, M.O., van de Meent, R., Pras, A.: Finding Elephant Flows for Optical Networks. In: Application session proceeding of the 10th IFIP/IEEE International Symposium on Integrated Network Management (IM 2007), Munich, Germany, Piscataway, pp. 627–640. IEEE Computer Society Press, Los Alamitos (2007)
18. Fioreze, T., Wolbers, M.O., van de Meent, R., Pras, A.: Offloading IP Flows onto Lambda-Connections. In: Clemm, A., Granville, L.Z., Stadler, R. (eds.) DSOM 2007. LNCS, vol. 4785, pp. 183–186. Springer, Heidelberg (2007)
19. Fioreze, T., Wolbers, M.O., van de Meent, R., Pras, A.: Characterization of IP Flows Eligible for Lambda-Connections in Optical Networks. In: Proceedings of the 11th IEEE/IFIP Network Operations & Management Symposium (NOMS 2008), Salvador, Bahia, Brazil, Piscataway, pp. 256–262. IEEE Computer Society Press, Los Alamitos (2008)
20. Montgomery, D.C., Runger, G.C.: Applied Statistics and Probability for Engineers, 4th edn. Wiley, Chichester (2006)
21. Pearson, K.: Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity and Panmixia. *Philosophical Transactions of the Royal Society A* 187, 253–318 (1896)
22. Kass, G.V.: An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics* 29(2), 119–127 (1980)
23. Moore, D.S.: The Basic Practice of Statistics, 4th edn. W. H. Freeman, New York (2006)