Verifying IP Meters from Sampled Measurements

Carlos Veciana-Nogués, Albert Cabellos-Aparicio, Jordi Domingo-Pascual, Josep Solé-Pareta
Centre de Comunicacions Avançades de Banda Ampla (CCABA)
Departament d'Arquitectura de Computadors
Universitat Politècnica de Catalunya (UPC)
Jordi Girona, 1-3, 08034 Barcelona, Catalunya (Spain)
Phone: + 34 934017182
Fax: + 34 934017055
{carlosv,acabello,jordid,pareta}@ac.upc.es

Abstract: Traffic measurement and analysis is nowadays a very critical task and requires complex and high-cost traffic analysis equipment and infrastructure. The difficulty increases when high bandwidth connections are to be monitored. Network topology, the network technology used, applications and user behavior influence the overall traffic on the network. For network planning and provisioning is crucial to know the trends of the traffic behavior, that is why the analysis of protocol application is very useful.

Most recent low-cost monitoring platforms for high-speed links rely on samples of the whole traffic. The purpose of the measurements range from monitoring and traffic characterization, to charging and billing. It is a must to know the statistical validity of the sampled data and the consequent detailed results obtained by classifying the traffic according to different categories. This fact is even more important when it is intended to be used for charging or cost sharing.

In this paper we propose an approach to validate the statistical capture of our system. We compare full traffic measurements which contain all the traffic with sampled traffic measurements, in order to know what is the threshold of a sampled traffic to estimate the real traffic with a certain degree of confidence. Results are presented using traffic from a real network environment and comparing our capture platform results with results derived from the real traffic.

Key words: passive IP traffic analysis, ATM, statistic validation

1 INTRODUCTION

New transmission technologies and new applications change the patterns of the traffic in the network. Moreover, new users with different profiles can change the patterns of the traffic. The knowledge of the Internet traffic characteristics is necessary to optimize resources, to plan the growth of the network, and to know the usage of the resources.

New trends in charging and billing proposals are based on traffic classification[1]. This classification could be made based on traffic volumes besides of the usage time or a fixed cost for renting links or access connections. The traffic classification for high-speed links is a high-cost and difficult management issue. The complexity of the traffic analysis for a network is related to the amount of traffic carried on, the topology of the network, the number of users, and the degree of detail desired to the reports.

Traffic characterization could be made at different levels. A first approach is to characterize the traffic based on network parameters (i.e. delay, losses, throughput, jitter, etc.) to study the network behavior. Another approach may characterize the traffic based on protocol parameters (i.e. number of flows, session duration, application types, etc.). Finally, any cross-related characterization or traffic analysis can be made based on any parameter at any level of the communication stack.

There are some projects working on IP traffic measurement on high-speed networks. CAIDA [2][3] consortium is making traffic analysis based on statistical measurements. SPRINT [4] labs makes traffic analysis based on the analysis of a complete (full) traffic capture during given periods. While the former can report very useful data about trends in the network behavior, the latter can report very detailed information about network parameters for a short period of time, including IP flows analysis and network QoS parameters across network nodes.

Several projects, CASTBA [5], MEHARI [6] and MIRA [7] projects, are our contribution to the traffic analysis in the Spanish NRN RedIRIS[8]. The main difference between our projects and the other traffic analysis initiatives has been the need of a full packet capture to perform traffic analysis at application header level. This means a lower capture ratio than if capturing only the IP and the TCP/UDP Transport headers.

We assume that the characteristics of the traffic change along the time, but they are more or less stable for short time periods (weeks or months) [5]. Our methodology consists on long-term traffic analysis based on samples using a low-cost hardware. The traffic is periodically validated with short-term full traffic analysis performed with high-cost equipment. As the load of the link under analysis increases, the sample corresponds to a smaller percentage of the total traffic. Then the main concern in this paper is how significant may be the conclusions derived from the sampled captures.

In the next Section we describe a generic architecture for the capture and analysis platform and our implemented system is presented. Section 3 describes the validation approach and Section 4 the application to the MIRA platform is presented. Section 5 discusses the results for MIRA analysis. Finally, Section 6 presents some conclusions and the ongoing work.

2 PLATFORM DESCRIPTION

A generic traffic analysis platform may be divided into two subsystems: the Traffic Capture Subsystem (TCS) and the Traffic Analysis Subsystem (TAS).

The TCS is a hardware platform that collects samples of the whole traffic in a high-speed link. The traffic capture is passive (usually with optical splitters), without interfering the performance of the network. Each packet of the network is labeled with a time stamp for further calculations and time estimations. The amount of traffic captured into each sample is bounded by the size of the buffers in the traffic capture equipment.

The Traffic Analysis Subsystem (TAS) is divided into several subsystems in order to improve the performance during the analysis (Figure 1). Some different modules may be defined in the TAS architecture: the preprocessing modules (PM), and the analysis modules (AM) -which can run in parallel on different machines-. The PM acts as a collector [9], reading all the samples, and extracting all the significant parameters of the traffic. This reduces the amount of data to be processed by the AM. The AM reorganize the data, add new attributes with queries to external databases, and produce reports about the captured traffic. The number of samples analyzed depends on the complexity and computation requirements of the PM and AM. An overflow control protocol is defined between PM and TCS to avoid data overflow. While PM is reading data, TCS is stopped and waits for a signal to continue capturing.

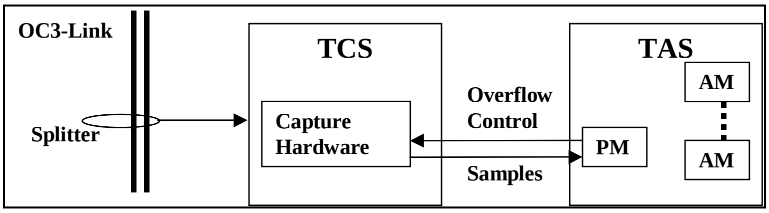

Figure1. Capture and Analysis Modular Architecture

As it has been explained, the TCS provides the TAS with samples. Then, the confidence of the final reports depends on the percentage of the whole traffic that the traffic samples represent.

The Spanish National Research Network [8] backbone was designed as a star. The core was settled in Madrid. There are 17 links, each one connecting a Comunidad Autonoma[1] (CA) to the central interconnection point. In the early stage of our project, about 1997, the bandwidth assigned to our CA was 10/8Mbps (in/out). Input traffic means traffic coming from the central interconnection point to our CA, and the output traffic is the reverse one. This allowed us to capture between the 10% and the 20% of the whole traffic, fair enough to make estimations about the whole traffic with a great level of confidence. Nowadays, the bandwidth assigned to the Catalunya link is 155Mbps (in/out) with a mean load about 64Mbps, while the size of the TCS buffers is the same. The percentage of the traffic captured is clearly lower now. In the near future, the link will be upgraded to 622Mbps. This fact places a real challenge for our traffic analysis platform because as the percentage of captured traffic diminishes we must validate the confidence of the results derived from it. The current low-cost traffic capture platforms have known limitations for full traffic capture at speeds over 622Mbps. Then a validation of a sampling method is need for the future short-term.

Figure 2 shows the current deployment of analysis points in the Spanish NRN. In this paper, we focus on the study of the TCS in one CA, the one located in the UPC (CCABA analysis point).

The first TCS we used was the HP75000 Broadband Series Traffic Capture System (BSTS) [10]. This is a high cost ATM traffic analyser, which main purpose is to perform accurate ATM traffic analysis. Moreover, the BSTS can be configured to capture ATM cells at full link-speed. The capture is bounded by the size of the hardware buffers (131,072 ATM cells ≅7Mbytes). The ATM cells can be reassembled to obtain AAL5 frames and IP packets.

In 1998, a second capture platform was setup (MEHARI). It is a modification of the OC3MON [11] software for the PCA200 Fore ATM card adapter. The OC3MON traffic analysis platform relies on periodic IP flow records, while our modified software relies on periodic full IP packet samples. The TCS based on PC is 50 times cheaper compared to the BSTS. As shown in Figure 2, several analysis platforms are deployed in the NRN. The platform deployed in the MIRA project uses the same TCS and adds more complex AM.

[1] “Comunidad Autonoma” stands for state or political designed administrative country area.

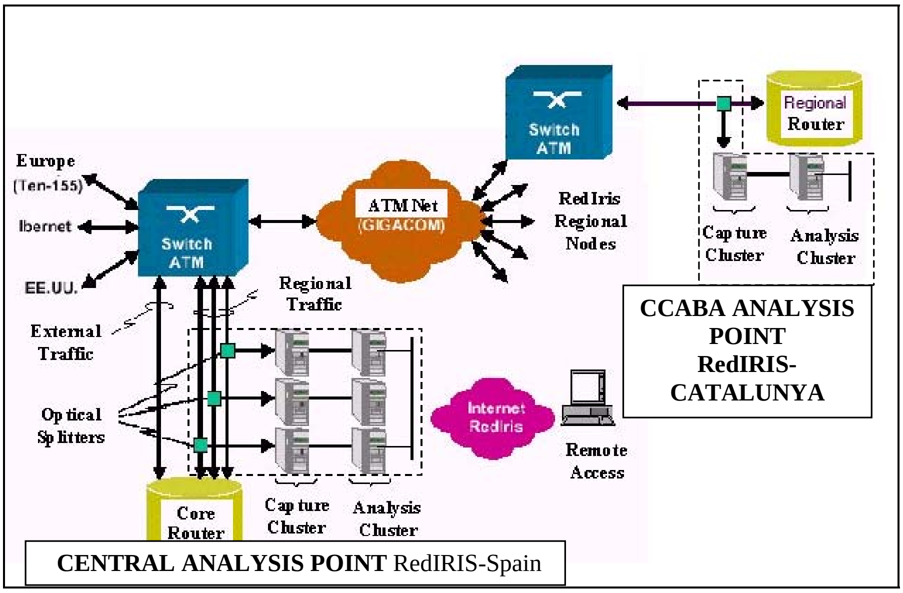

Figure 2 Traffic Capture And Analysis Points for CASTBA, MEHARI and MIRA projects.

Now we are planing to use traffic capture specific cards like DAG cards [12]. While these cards are able to capture at full speed link at 155Mbps.

3 VALIDATION APPROACH FOR STATISTICAL TRAFFIC CAPTURE

The high cost of the BSTS equipment is justified, among other capabilities, for the ability to perform real-time traffic statistics at full link speed. The drawback is that the available real-time statistics are far from IP traffic parameters. But it is possible to account more IP parameters using pattern matching over ATM cells. A set of counters may be configured to perform bit pattern matching inside the ATM cells so that some IP and TCP/UDP header fields may be accounted during the capture.

There are some cases where a simple pattern matching over cells cannot find the desired data of the IP headers. When a packet contains options or an IP packet is fragmented, the transport protocol information (i.e. source and destination ports) are not detected. Nevertheless, the amount of those packets is small enough to be ignored in the study [3][5].

The validation approach takes as a reference a long period of capture using the BSTS and accounting several parameters. As this is a full traffic capture all comparisons are related to it. In order to validate different traffic capture and analysis platforms we model them and simulate their captures. It must be noticed that it is very difficult to obtain a full traffic capture and a statistical capture for a long period of time and have them synchronized in order to perform the same analysis. Moreover, the different capture platforms are not available at the same time. That is the main reason to use what we call the simulation approach.

Our Statistical Traffic Validation Platform (STVP) is based on real time traffic statistics measured with the BSTS and a sampler program that simulates the effect of other statistical TCS. This allows us to compare the differences of mean values of such parameters, comparing the full traffic mean to the simulated sampled mean of the same traffic. The BSTS average measurements are accounted every second, assuming no cell loss and a perfect clock (nanosecond precision).

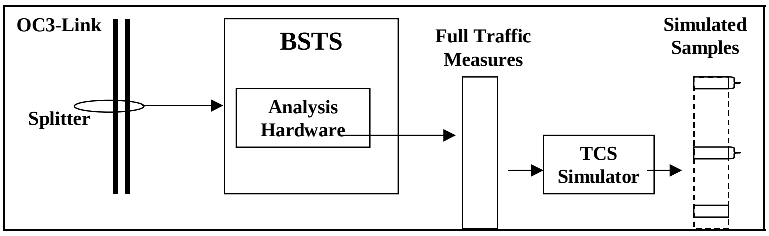

Figure3. Statistical traffic Validation Platform

The STVP simulator can be configured to produce samples of the whole traffic. The STVP can be configured to collect samples based on capture size of the capture buffers, and silence periods. The silence periods are related to the time required for dumping the buffers to the disk in the real TCS and for the preprocessing and analysis time between traffic captures. In particular, this is the model that fits our TCS. Other TCS where capture time does not depend on buffer size, but depends on time can be also simulated. In this case, the size of the capture is bounded by the capture time. This applies for all the systems that perform periodic sampling.

One important parameter is the integration period, that is the period of time that represents each sample. We defined an integration period higher than STVP Simulator sampling period for the mean calculation. Then, the mean of the full traffic measurements during such period, and the mean of the simulated samples during the same period, can be compared. We assume that the mean obtained with the BSTS is the real value of the traffic mean. Then, the statistical analysis via test for a zero mean for the difference of the means can be applied. Moreover, as we simulate the samples from a real trace, a Confidence Interval for the difference of paired observations can be used. A simple lineal regression (SLR), comparing both analysis methods, will give the measurement of the fairness of the statistic capture for any load of the link.

That is, while the confidence interval (C.I.) for the difference gives information about the statistic confidence about the sampling method, the SLR gives the same information but related to different loads of the link.

Figure 3 shows the architecture of the STVP. The BSTS collects real-time traffic parameters mean values. This is a continuous trace, one measurement per second. Then, a sampled trace based on configured TCS activity/silence periods and buffer size is given.

4 STATISTICAL TRAFFIC VALIDATION FOR MIRA PLATFORM

A long trace of real traffic has been obtained. The validation is performed considering two different cases. The first one takes into account the full traffic parameter estimation, and the second one is for classified traffic parameter estimation. The full traffic parameter validation includes the whole aggregated flows that compose the traffic. In our case of study they are the total number of bytes per second and packets per second. The classified traffic parameter validation stands for those parameters derived from some of the flows that compose the whole traffic. In our case, they are transport protocol packets per second (TCP and UDP), application protocol packets per second and network address packets per second. This latter validation is very important since it is very common to classify the traffic per flows, per protocol, per source or destination Autonomous System, per server, etc.

4.1 Description of the traffic traces

The full traffic trace[2] corresponds to the traffic in the CCABA (UPC) access point link during 37 days (from 20th July 2001 to 25th August 2001, 3196800 seconds) accounting cells per second, and AAL5 frames per second, in both directions (in/out). The number of cells multiplied by 53 (ATM bytes) is the value of IP bytes plus IP/ATM overhead. In [5][2] a study of the average IP/ATM overhead is about the 15%. However, all the ATM bytes can be considered for the study of an IP/ATM link. The average load of the link is 20,38Mbps/27,98Mbps (in/out). The number of AAL5

[2] The traces from the real traffic which this paper is based, can be found at: http://www.ccaba.upc.es/projects/mira/E_mira_local.html

frames fits exactly the number of IP packets. This measurement is useful for the study of routing devices that are limited more by the number of packet routed than the size of those packets.

The second trace is also obtained from the same link, and at the same time that the first trace. The accounted parameters are based on bit pattern matching over ATM cells. Patterns that define some types of IP headers have been configured. This allows us to study the effect of traffic classification over sampled measurements. We would like to know the effect of sampling on the full traffic and on each class of traffic, in function of the percentage each class represents. The study of traffic classification is performed only for the number of packets, since real time statistics cannot provide byte counters per packet matched, only packet counters.

Tables 1,2 show the traces of real traffic used in the validation. Table 2 also shows the percentage of each class of traffic.

Table 1. Full traffic measurements

Full Traffic(3196800 s)	In	Out
GigaBytes	8147,23	11184,27
IP packets *10E6	14190,59	16767,21
Mean IP Packets per second	4439	5245
Mean bytes per second	2548558	3498583

Table 2. Classified Traffic measurements

Traffic class 362400s	InIP Packets	OutIP Packets
TCP	10926,66*10E6	14037,14*10E6
Mean	3418 pps	4391 pps
Percentage	76.99%	83.71%
UDP	364,43*10E6	620,17*10E6
Mean	114 pps	194 pps
Percentage	2.56%	3.69%
Big Network	786,41 *10E6	549,84*10E6
Mean	246 pps	172 pps
Percentage	5.54%	3.27%
Small Network	63,93*10E6	47,95*10E6
Mean	20 pps	15 pps
Percentage	0.45%	0.28%

The traffic classes shown in Table 2 are not intended to be exhaustive but an example of the validation. For each traffic class the total amount of packets, the mean number of packets per second and the percentage respect to the total traffic are given, both for input and output traffic. TCP and UDP classes deserve no more comments. The important fact to take into account is that as more detailed is the classification less traffic is captured and the measurements may not be significant enough. The row labeled “Big Network” corresponds to the input and output traffic for a medium-size University network provider identified by a range of IP addresses. This traffic class represents a classification per AS, per peer traffic, per entity, etc. Finally, the row labeled “Small Network” represents a small department inside one university.

4.2 TCS Simulation for MIRA

The TCS simulations are the result of different sampling schemes over the full traffic measurements. In fact, the different configurations of the STVP can produce the same results than other real TCS.

The BSTS reports real-time traffic statistics with a precision of one second. In the simulation, we model the real TCS considering that the traffic within a second is constant. Then, the proportional traffic to the capture time inside that second is calculated. The MIRA TCS reads 25 buffers of 1MB between preprocessing periods. This means an average capture time of 0.43 seconds at average link load. The traffic capture periods in MIRA simulation is configured based on buffer sizes. Then, the variable capture time is derived from the cell accounting in the BSTS trace, until the byte counter reaches 1MB. This short capture time means that most of the simulated samples are based on the proportional value of the average within a second. There is a silence time of about 1.25±0.25 seconds related to the dumping of each 1MB buffer to the disk. In addition, there is a preprocessing time of about 100±10 seconds between each batch of 25 files. A fixed time (90 seconds) is related to the sequential access to each packet of the file, and the variable time (20 seconds) is related to different analysis time depending on the complexity of the traffic captured. In summary, 25 buffers of 1 MB are captured every 2 minutes approximately.

5 RESULTS AND VALIDATION

This section shows the results of the comparison between the mean values obtained from BSTS trace (real traffic), and the mean values obtained from the simulated sampling process over the same trace. In order to have comparable mean values, both means are calculated over the same period. Then, we have paired observations obtained from two different sources, the real measurement and the sampled measurement. The difference of the mean is computed for each pair, then a 95% Confidence Interval is computed [13]. If the C.I. contains the zero value, there is no statistical difference for the mean value between the real measurement and the sampled measurement. We call it the test of zero.

Moreover, a Simple Lineal Regression (SLR) [13] study is applied in order to verify the confidence of the sampled model against the real samples for all traffic loads.

5.1 Full traffic parameter estimation

Table 3 shows the 95% C.I. for the test of zero under MIRA TCS conditions. The integration time for the mean calculation is 2 minutes, higher than MIRA TCS time between samples. The column labeled Sampled Mean gives the mean value to be compared with the error values. The C.I. does not contain the zero value for these conditions, but it is two orders of magnitude lower than the mean of packets and three orders of magnitude lower than the mean of bytes.

Table 3 95% C.I. test of zero with 2 minutes integration period

Type	Sampled Mean	% Captured	C.I.
In IP packets/s	4476	14,59	36,041 ± 7,057
In Bytes/s	2561172	14,34	1910,173 ± 4238,463
Out IP packets/s	5298	8,17	0,387 ± 7,971
Out Bytes/s	3527468	8,15	2741,266 ± 4759,824

Table 4 shows the SLR analysis. The second column also shows the sampled measured mean value in order to have a reference for the errors. The third column shows the real percentage of traffic that has been captured in relation to the total traffic on the link. All straight lines are very close to the theoretic y=x (real average = sampled mean). In fact, there are small variations in the intercept and slope values compared to the mean values. Moreover, the intercept point represents a zero traffic load that almost never occurs. The slope is always below 1. This fact means that under these conditions the sampled measurement is always below the real value. The R^2 (Coefficient of determination) gives values above 90% in all cases. No differences are detected between in and out traffic, neither for packet average or bytes average comparison. Figure 4 shows an example for the UDP traffic, comparing the theoretic line x=y with the SLR analysis.

Table 4 SLR values for full traffic parameter estimation using with 2 minutes integration period

Type	Sampled Mean	% Captured	B0	B1	R²
In IP packets/s	4476	14,59	288,902 ± 12,691	0,943 ± 0,002	94,25
In Bytes/s	2561172	14,34	73001,832 ± 5530,704	0,975 ± 0,001	97,61
Out IP packets/s	5298	8,17	447,699 ± 16,557	0,924 ± 0,002	92,35
Out Bytes/s	3527468	8,15	389608,597 ± 11263,180	0,896 ± 0,003	91,24

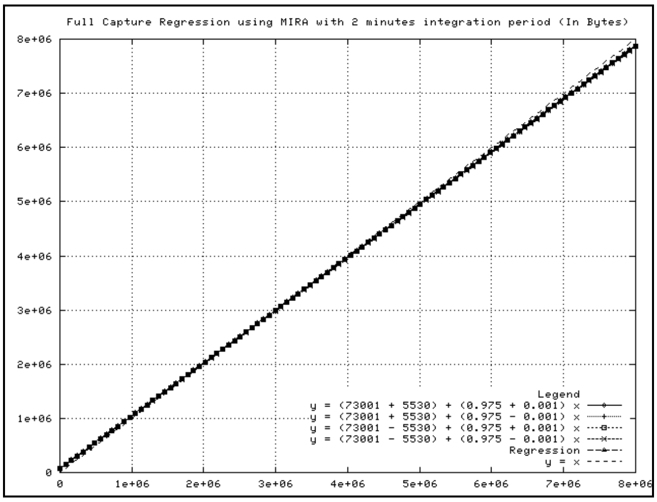

Figure4. SLR Straight lines for Input traffic in bytes, with two minures integration period.

Table 5 shows the results of the C.I for the test of zero, with an integration period of one hour. In fact, this is the integration period used in the MIRA traffic estimation in the current version. Now, the order of magnitude of this parameter is always two orders of magnitude lower than the mean average for the packets and three orders of magnitude lower for the bytes.

Table 5 95% C.I. test of zero with an hour integration period

Type	Sampled Mean	% Captured	C.I.
In IP packets/s	4361	14,59	-78,208 ± 13,536
In Bytes/s	2460101	14,34	-4145,024 ± 9192,585
Out IP packets/s	5182	8,17	0,01 ± 13,898
Out Bytes/s	3448180	8,15	3359,193 ± 6717,909

Table 6 shows the results of the SLR values for an integration period of one hour. It can be noticed that the increase of the integration period improves the validity of the b0, b1 and R2 values. Now, all R² values are better than 99. However, the slope values are still below one.

Table 6. SLR values for full traffic parameter estimation with an hour integration period

Type	Sampled Mean	% Captured	B0	B1	R²
In IP packets/s	4361	14,59	-14,968 ± 25,655	0,985 ± 0,005	99,23
In Bytes/s	2460101	14,34	-53871,675 ± 12011,120	0,986 ± 0,003	99,62
Out IP packets/s	5182	8,17	19,850 ± 32,258	0,984 ± 0,005	99,06
Out Bytes/s	3448180	8,15	-27710,349 ± 17412,907	0,993 ± 0,004	99,38

Table 7 and 8 shows the test of zero and SLR values for full traffic parameter estimation with 15 minutes integration period. This table has been obtained in order to compare MIRA TCS with most SNMP systems, which usually report counters every 15 minutes.

The 15 minutes integration time gives C.I. with an error similar than one hour integration time, but containing zero in most cases (except for the In IP packets). For the SLR analysis, the R² values are always better than 98%. Moreover, the slope values contain the zero slope and the intercept point is clearly better than 2 minutes integration time case.

Table 7 95% C.I. test of zero with 15 minutes integration period

Type	Sampled Mean	% Captured	C.I
In IP packets/s	4413	14,59	-26,29 ± 8,563
In Bytes/s	2509503	14,34	-1393,37 ± 5152,554
Out IP packets/s	5224	8,17	0,111 ± 7,742
Out Bytes/s	3477118	8,15	-1170,346 ± 3963,711

Table 8 SLR values for full traffic parameter estimation with 15 minutes integration period

Type	Sampled Mean	% Captured	B0	B1	R²
In IP packets/s	4413	14,59	108,673 ± 15,463	0,9696 ± 0,003	98,86
In Bytes/s	2509603	14,34	-199,723 ± 6662,097	0,984 ± 0,002	99,53
Out IP packets/s	5224	8,17	110,319 ± 16,771	0,974 ± 0,002	98,98
Out Bytes/s	3477118	8,15	69972,113 ± 9692,829	0,973 ± 0,002	99,2

5.2 Traffic Classification

In this section, we want to analyze the effect of the sampling over meters that account more specific traffic characteristics.

For the classified traffic traces (Table 2), a 95% C.I. and a test of zero has been made, under the same conditions explained in section 5.1. The traffic classifications proposed in this section are only examples.

The SLR values have the same characteristics as the values for full traffic. We only present the C.I. for one hour (Table 9) and 15 minutes (Table 10) integration periods, in order to reduce the data.

Table 9 shows the mean comparison analysis over the real trace and the sampled trace for the MIRA TCS. All straight lines are also very near the theoretic y=x. Figure 5 shows the regression lines with confidence interval for the class with less traffic (Small Network). Also in this case, a value of R^2 about 99%, and the slope value contains the 1 value (Figure 5).

Once again, confidence interval for the mean is two orders of magnitude smaller compared to mean values.

Table 9. 95% C.I. test of zero with 1 hour integration period

Type	Sampled Average	C.I for the average of the errors
In TCP packets/s	3364	53,85 ± 8,708
In UDP packets/s	111	3,698 ± 5,372
In Big Network packets/s	242	4,81 ± 1,137
In Small Network packets/s	19	-0,382 ± 0,167
Out TCP packets/s	4350	-40,964 ± 9,971
Out UDP packets/s	186	-7,957 ± 4,646
Out Big Network packets/s	170	-2,39 ± 1,168
Out Small Network packets/s	15	0,01 ± 0,097

Table 10. 95% C.I. test of zero with 15 minutes integration period

Type	Sampled Average	C.I for the average of the errors
In TCP packets/s	3402	-17,038 ± 5,643
In UDP packets/s	112	-2,189 ± 3,065
In Big Network packets/s	246	-0,972 ± 0,967
In Small Network packets/s	20	-0,153 ± 0,134
Out TCP packets/s	4379	-12,452 ± 5,187
Out UDP packets/s	190	-4,033 ± 3,187
Out Big Network packets/s	171	-0,92 ± 0,982
Out Small Network packets/s	15	0,111 ± 0,088

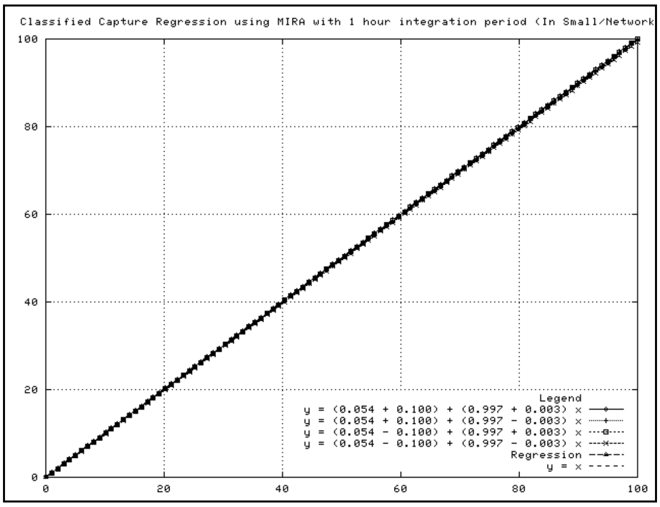

Figure5. SLR Straight lines for MIRA UDP classified traffic, with two minutes integration period.

5.3 MIRA traffic classification validation

Some of the traffic classifications proposed in the MIRA project fit into the examples of traffic classification shown in this paper. Then, they may be useful for charging or cost sharing under great level of confidence. Some sampling-based analysis tools, that currently report trends about Internet traffic, could be used for charging and network provisioning if a confidence interval for the measurements is given. Also, some commercial charging methods like UUNet "Burstable and Metered"[14], can be improved with more detailed traffic classifications for charging. The MIRA project proposes several traffic classifications over sampled traffic.

1. Network Address: classifies the traffic based on source and destination address, accounting bytes and packets per AS, Network, and Host.
2. Application port: classifies the traffic based on source and destination port, extracting the information from the IP header and accounting bytes and packets per application (group of ports).
3. Application protocol: classifies the traffic based on application protocol analysis, extracting the information from the IP payload, and accounting bytes and packets per application.
4. Payload pattern matching: classifies the traffic based on application payload, and accounting bytes and packets.

The third and fourth classifications, cannot be verified with our method because it is not possible to perform such complex analysis with our real time traffic analyzer. However, we are working on some alternatives which allow real-time accounting via a smarter method than simple pattern matching over IP headers, using specialized hardware [12].

The Network Address classification seems to be unaltered by the sampling method, like we saw in the section 5.2, but a further study is needed for more complex and different size aggregations. In our example, also very small networks, with few packet per second means, can be estimated with good C.I. Other address aggregations like Autonomous Systems or singular networks are under study.

The Application Port classification is under study. Some preliminary results show that HTTP traffic sampled means seem to be unaltered by the sampling process. Since the HTTP traffic is usually 80% of the whole traffic, its behavior is usually the same as the whole traffic.

Most traffic classes in MIRA traffic classification are over 5% of the total traffic. Then, the traffic estimation from sampled averages seems correct, whenever low traffic classes compounds will be like full traffic.

6 CONCLUSIONS AND ONGOING WORK

Statistical capture subsystems are good for real traffic estimations under some conditions. In this paper we present an approach to validate the accuracy of the sampling methods, and to determine the significance of the results derived from the traffic analysis. The Statistical Traffic Validation Platform (STVP) is a simple method that allows the simulation of the behavior of some Traffic Capture Subsystems (TCS). It must be noted that it is required a traffic capture equipment able process or store all the traffic at line speed, since this is the reference trace for the validations. But, once the cheap TCS platform has been validated for the classes of traffic under study, the expensive equipment is not needed any more. Then, the deployment of many non-expensive TCS may provide a very good knowledge about the trends of the traffic in the network. Finally, these TCS may be used to collect significant information to be used for approximated billing or cost sharing.

The statistical analysis via confidence interval for the difference of the means, even though does not fulfil the zero test, gives an error values two orders of magnitude lower than mean values. This seems a good approximation, taking into account the simplicity of the hardware equipment and accounting process.

Our preliminary results show that very low capture ratios (lower than 1% for some traffic classifications like small networks) can produce approximations of average measurements with an error lower than two orders of magnitude compared to mean values.

The current work consist on a systematic verification of different traffic classifications, specially those based on address aggregations (Autonomous Systems, Networks and subnetworks), since the classification by application seems to be difficult for new peer-to-peer applications, web based services and encrypted services.

ACKNOWLEDGEMENTS

This work is funded by the CICYT (Spanish Education Ministry), under contract 2FD97-2234-C03-02. Some of the ideas of this work came from the participation on the MIRA project, together with UPM, UC3M, Telefónica I+D, and RedIRIS.

REFERENCES

[1] A.M. Odlyzko, "The history of communications and its implications for the Internet". http://www.research.att.com/~amo/doc/history.communications0.ps. June 2000

[2] J. Aspirdof et al., "OC3MON: Flexible, Affordable, High Performance Statistics Collection", INET'97, Malasya, June 1997.

[3] C. McCreary and KC Claffy, "Trends in Wide Area IP Traffic Patterns: A View from Ames Internet eXchange". In Proc. of 13th ITC Specialist Seminar: IP Traffic Measurement, Modeling and Management. Monterey, USA, September 18-20,2000.

[4] C. Fraleigh, S. Moon, C. Diot, B. Lyles, F. Tobagi. "Architecture of a Passive Monitoring System for Backbone IP Networks". Sprint technical report TR-00-ATL-101801. October.2000.

[5] M.Alvarez et al., "CASTBA: Internet Traffic Measurements over the Spanish R&D ATM Network", HP-OVUA Workshop, Rennes (France), April 1998.

[6] P. Lizcano et al., "MEHARI: A System for Analyzing the Use of the Internet Services", Computer Networks 31 (1999), pp. 2293-2307.

[7] C Veciana-Nogués, J. Domingo-Pascual and J. Solé-Pareta, "Server Location & Verification Tool for Backbone Access Points" In Proc. of 13th ITC Specialist Seminar: IP Traffic Measurement, Modeling and Management. Monterey, USA, September 18-20,2000.

[8] http://www.rediris.es

[9] Nevil Brownlee et al. "Traffic flow measurement: Architecture" RFC 2063, January 1997.

[10] HP75000 BSTS http://advanced.comms.agilent.com/bsts/datasheets/e4200b.htm

[11] http://www.caida.org/tools/measurement/coralreef/

[12] http://dag.cs.waikato.ac.nz

[13] Raj Jain, "The art of computer systems performance analysis: techniques for experimental design, measurement, simulation, and modeling". Wiley professional Computing, 1991, ISBN 0-417-50336-3

[14] http://www.uunet.com/products/uudirect/uudirect