TAT: Traffic Analysis Tool for the Statistical Study of IP networks

Josemaria Malgosa-Sanahuja, Maria-Dolores Cano, Fernando Cerdan , Joan Garcia-Haro
Polytechnic University of Cartagena
Department of Information Technologies and Communications
Campus Muralla del Mar s/n (Ed. Hospital de Marina)
30202 Cartagena, Spain
Ph. +34 968 325953
Fax. +34 968 325338
{josem.malgosa, mdolores.cano, fernando.cerdan, joang.haro}@upct.es

Abstract

In this paper, we present a new software package, TAT, that allows the statistical study of the IP traffic transmitted trough an Internet node. This tool has been successfully tested under a work carried out in a subnet of the regional network of Murcia (Spain) called Ciez@net. Cieza is a village of 30,000 people where the pilot network Ciez@net intends to bring these people the opportunity to use the new emerging Internet applications in their daily life. The goal of the traffic analysis tool developed in this project has been to analyse the IP traffic transmitted in the regional network in order to figure out if the economic effort done during the network deployment and operation was worth enough regarding the use that people do of the network services and infrastructure. The success of this software tool lies on its flexibility, easy handling and interface independence.

1. Introduction

The project Ciez@net testbed [1] is the first pilot experience of a Digital City in the Autonomous Community of the Region of Murcia in Spain. The idea behind it is to ensure the fast introduction of the Information Society in a medium-size village where several types of inter-relationships are given amongst citizens, administrations and the business world. This approach may allow to easily extend the benefits of this pilot experience to the whole Region and even nationwide.

The testbed assumes the fact that the Information Society will only be fully reached when each citizen, institution or company can have access to advanced electronic information services, training to use them and offering prices according to the user willingness to pay.

In this context, monitoring when and how the network is being used is a key issue. To help on that task the TAT (Traffic Analysis Tool) was developed. The TAT is a software tool that makes easy a statistical study of the traffic transmitted over an IP network.

Frame capturing can be done by hardware or software mechanisms [2]. A software mechanism gets the network state information by means of polling techniques using network management protocols (e.g. SNMP) [3]. This clearly, in turn, influences or affects the traffic network itself. Given that we were interested in an accuracy monitoring of the Internet access and utilization of Ciez@net, we selected a hardware mechanism inserting a network analyzer to sniff and capture frames in the node under study. The TAT software is in charge of the frame and data of interest processing to help in the statistical IP traffic study.

The graphical environment of the TAT tool is based on the Tcl/Tk programming language [4]. The interactive nature of Tcl, combined with the fact that Tcl code require far fewer lines than other languages like C or C++, makes it a good selection for a fast application development. In addition Tcl/Tk is supported by many operating system including LINUX, MS Windows and Macinthosh.

In this paper, we present TAT as a software package to provide statistical analysis of IP traffic in a Internet node. As commented above the TAT tool is based on Tcl/Tk. Nevertheless the statistical functions were developed using the *awk* language. *awk* is a programming language supported in any UNIX system specially designed to work with structured data files. The interface between *awk* and Tcl/Tk is provided by system scripts. For this reason *awk* language was used as a first step, leaving scripts development for other platforms, as a second step in the proyect. The TAT makes a powerful analysis of the IP traffic in the node under study in terms of utilization, packet size distribution, service distribution (HTTP, FTP, IRC, etc.), top 10 site addresses, instantaneous number of connected users, etc., in both up and down stream directions. In summary, this work tries to help developers in the construction of statistical functions to manipulate TCP/IP network traces obtained by means of a hardware network analyzer.

0-7803-7080-5/01/$10.00 ©2001 IEEE

The rest of the paper is organized as follows: In section 2 we fully describe the TAT tool including the methodology for data processing, data format options and configuration files. Section 3 is devoted to explain in pseudocode the algorithms used to implement the statistical functions. Finally, we present the main conclusions derived from this work in section 4.

2. TAT description

Figure 1 shows the entire process from the capture of the data frames until to the TAT processing and presentation of the results. Raw data files include IP packets and additional fields as frame capture time, frame size, source and destination addresses as well as network transport and application protocols.

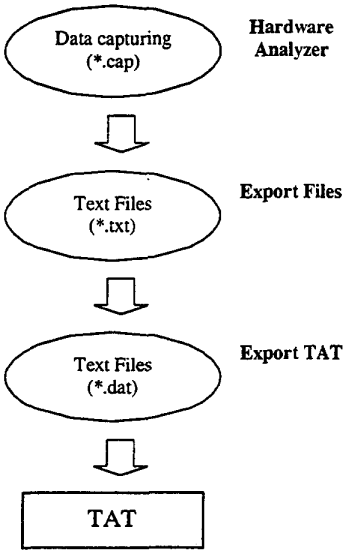

Figure 1: Data processing from capturing to analysis.

Raw data files are first exported to text format (skill provided by the analyzer) and finally to the especific format used by the TAT tool, whose characteristics are described below:

- Text files
- Field separator is the tab keystroke

The following script is an example to export the text files, where the file *convert.awk* contains the *awk* instructions to provide the above listed characteristics. That is, white spaces between fields are substituted by one tab and all fields are grouped in one word.

```
For i in *.txt; do
        fout =`echo $i | sed –e 's/.txt/.dat`
        echo "Recoding $i"
        awk –f convert.awk $i >fout
done
```

The main options for plotting results can be seen in figure 2 while in figure 3 we display the parameter configuration window.

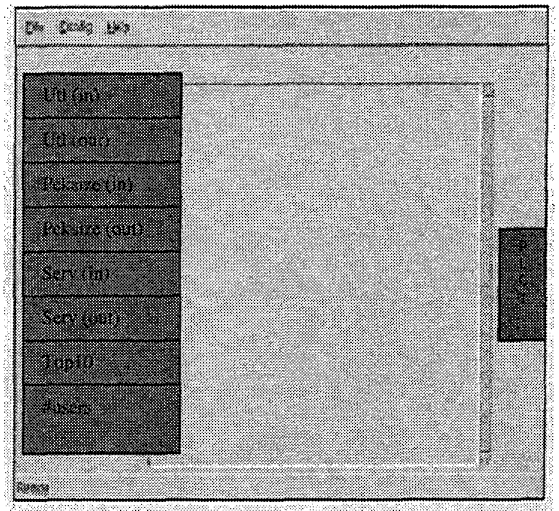

Figure 2: Main TAT window.

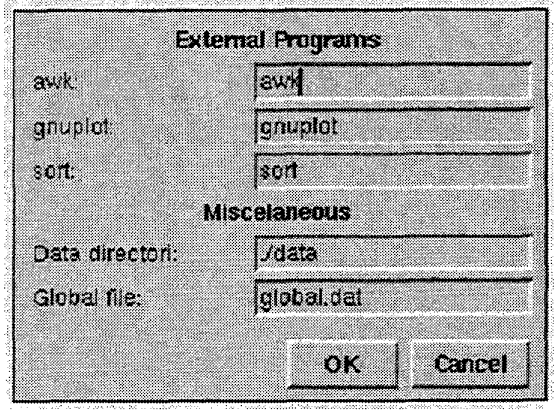

Figure 3: Configuration file (.tatrc.).

3. Functions for statistical analysis

In this section we describe how the TAT computes the different statistics of interest.

a) Utilization: This function calculates the link occupation time, that is, transmitting frames both up and downstream. For this metric the observation time interval is 5 minutes. See the pseudocode below and the result in figure 4. A graphical representation is plotted by clicking in the box *PLOT*.

The algorithm differentiates both up and downstream utilization.

```
For all registers
(
        First register ? --> init = frame capture time
        abs = frame capture time
        interval = f(init,abs(5 minutes multiple))
```

```
        destination address cieza@net user ? -->
utl_down[interval] += frame size
        source address cieza@net user ? --->
utl_up[interval] += frame size
}
    For all intervals
    {
    print -->      interval,utl_down[interval]/(link
bandwidth * interval)
    print --> interval,utl_up[interval]/(link bandwidth
* interval)
    }
```

b) Number of connected users. It may seem that this metric has no sense in a connectionless network like Internet. For this reason, we consider the number of different users that in an interval of 5 minutes use the network. This interval is long enough to detect a simple mouse click, and short enough to avoid low activity users not loading effectively the network.

```
For all registers
{
first register ?-->init=Frame capture time
abs= Frame capture time
interval =f{init, abs(multiple of 5)}

Interval change ?
        For all IPaddress--> {IPaddress[IP]=0}

It is the source a cieza@net user ?
        IP=destination address
        IPaddress[IP]++
        IPaddress[IP]== 1?-->usuarios[interval]++
}

For all intervals {print-->interval, users[interval]
```

c) Packet size distribution: This parameter measures the probability that the packet size be in a interval defined in multiples of 100 bytes. The algorithm includes the packet size distribution both, up and downstream

```
For all registers
{
interval = a multiple of 100 bytes
It is the destination a cieza@net user ?
        pksize_in [int{IP packetsize
        captured/interval}]++
        total_in +=pksize_in
It is the source a cieza@net user ?
        pksize_out [int{IP packetsize
        captured/interval}]++
        total_out +=pksize_out
}

For all interval
{
        print --> interval,
pksize_in[interval]/total_in
        print -->interval,
pksize_out[interval]/total_out
}
```

d) Service distribution: Here we calculate the probability that an Internet application like HTTP, FTP, IRC, etc., be used. We also distinguish between up and downstream traffic and we take care to isolate Frame relay control frames. This is because ciez@net users get Internet access through a 512 Kbps Frame Relay link.

```
For all registers
{
FR control frame @NET? --> serv_in[Control_FR]++
FR control frame @CPE? --> serv_out[Control_FR]++

It is the destination a cieza@net user ?
        service=application(http,ftp,etc)
        serv_in[service]++
It is the source a cieza@net user ?
        service=application(http,ftp,etc)
        serv_out[service]++
}
for all service
{
    total_in += serv_in[service]
    total_out += serv_out[service]
    print --> service, serv_in[service]/total_in,
    serv_out[service]/total_out
}
```

e) Top ten addresses: In this case we only consider upstream traffic. See results in figure 5.

```
For all registers
{
It is the source a cieza@net user ?
        destination=IP destination address
        top10[destination]++
}
for all service
{
print --> destination, top10[destination]
```

4. Conclusions

In this paper, we present the TAT (Traffic Analysis Tool). The TAT is a software package written in Tcl/Tk language mainly, although it also uses *awk* and system scripts algorithms to develop an efficient and friendly software tool from the user's point of view that helps in the handling and further study of previously IP captured traffic.

The tool is fed with the IP data captured by using an electronic device to avoid traffic interfering in

the network. This makes the measurement procedure more accurate than other software solutions that use polling techniques like SNMP.

Frame data processing and result displaying is performed by a friendly windows environment which is described in detail. From the user point of view, how static functions were developed may seem useless. Nevertheless, they were fully described because it may help other developers to make their own traffic analysis tools based on this model as it happens with the free agents software based on SNMP or RMON packets.

The main advantage of the methodology explained for the IP traffic analysis including the development of the TAT package described in this paper, is its independence on the network node technology under study, in addition it does not load the network, have an extremely easy and friendly window configuration environment and allows a data analysis more specific and wider than other free software tools do, with a lower cost than other commercial or freeware solutions.

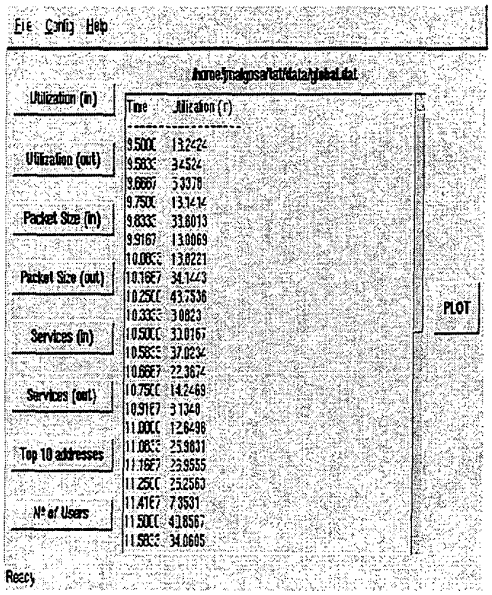

Figure 4: Utilization downstream.

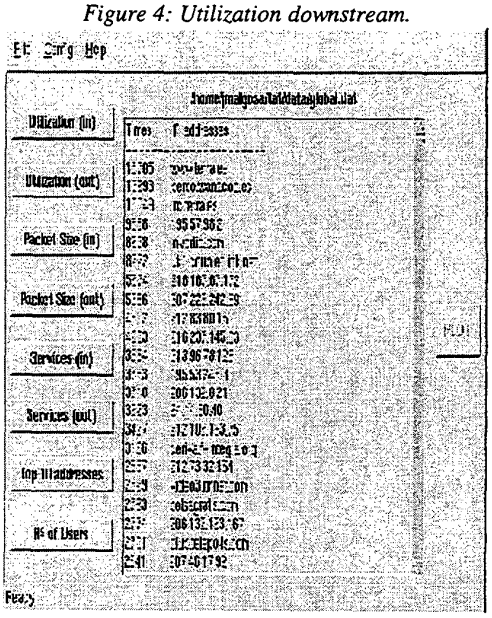

Figure 5: Top 10 addresses.

Acknowledgements

This work was partly supported by the Spanish Research Council under grant TIC2000-1734-C03-03 and by the Fundación Integra under project 0124.

References

[1] *www.cieza.net*

[2] A. Salles, G. Vargas, R. Carlesso, R. Piccoli "Test and Evaluation of Software Tools for Monitoring Computer Networks" Proceedings of CITEL'2000, La Habana (Cuba).

[3] J. Case, "A Simple Network Management Protocol (SNMP)" RFC 1157.

[4] Eric Foster-Johnson, "Graphical Applications with Tcl & Tk" 2d ed., M&T Books 1997. ISBN 1-55851-5690.