ELEC-E7130 Assignment 3. User traffic

Markus Peuhkuri Tran Thien Thi Weixuan Jiang
César Iván Olvera Espinosa Yu Fu

Prerequisites

1. For the second and third part you will need a root level access to Linux computer (or administrator access on Windows computer). If you do not have a computer suitable for that (e.g. if you only have a company laptop), please contact course staff and a loan computer can be arranged. A virtual computer will work for that purpose.
2. If you are not very familiar with network capture skills (TCPdump, Wireshark or tshark), you can
   a. Begin by watching TCPdump introductory video and Wireshark introduction video for network capture.
   b. View ELEC-E7130 Network capture tutorial to look through those commands and codes in detail.
   c. Take a look at some code snippets which may give you some help.

Learning outcomes

At the end of this assignment, students should be able to

1. Get to know the differences between flow data and packet data.
2. Capture traffic as a passive measurement.
3. Learn how to capture internet traffic.
4. Analyse the network captured traffic from different aspects.
5. Compare the memory and time used by different methods and know the more appropriate method for large data.

Introduction

This assignment contains three tasks to introduce in more detail the traffic data that can be analysed for different tools. Please read all instructions before starting because it is helpful to identify common work.

- Task 1: Introduction to the traffic data
- Task 2: Analyse flow data
- Task 3: Analyse packet capture (user traffic)

To use some of the course-specific tools, some environment settings are needed in Aalto servers. Depending on your login shell, you need to **run one of the following commands on school computer**. The first command is used if you have any Bourne Shell compatible (like the Aalto default zsh or bash).

> **Note:** You may type the command `kinit` before accessing the directory to avoid issues related to the permissions.

1. `source /work/courses/unix/T/ELEC/E7130/general/use.sh`
2. `source /work/courses/unix/T/ELEC/E7130/general/use.csh`

You need to **provide the tool's name and method (command line, if any) you have used to answer the above questions in your report file**. We recommend that you try to use at *least one command-line tool* for analysis because, in a final assignment, the data volume is much larger.

Task 1: Introduction to the traffic data

You must answer the following points appropriately:

1. What is the **passive measurement** in terms of network traffic? What kind of information does it provide, and what is its role or significance?
2. Please provide an explanation of the concepts of **packet capture and flow data**. What kind of information they can provide? Additionally, discuss the advantages, disadvantages, and importance of both packet capture and flow data in network analysis.
3. What is **hashing**? How does the hash algorithm work and what is the relation with the **memory management in the large data analysis**?

Task 2: Analyse flow data

In this task, **capture the traffic data from your computer**. In the case of using a virtual machine (VM), generate traffic within that virtual computer instead of the usual host because it acts as a separate computer.

Choose one of the packet-capturing tools available such as *dumpcap*, *Wireshark*, *tcpdump*, etc.; **to capture network traffic for one hour or more** while using the computer as your normally do (browse web, check e-mails, watch video, listen music, do assignments, and so on).

Once you have the pcap file, use a tool (CoralReef, NetMate, tstat or program of your choice) to **convert the pcap file into flows**.

Once with flow data, **answer the following points**.

1. Provide basic statistics of flow data, including

- total number of flows,
- minimum, median, mean and maximum flow sizes in bytes and packets

2. Plot the traffic volume (bytes) of the flow data file.

   Note: Getting traffic volume is more difficult from flow data files due to the known information are only start time, end time, and flow size (bytes) (as shown in the figure). For example, if the flow contains 100,000 bytes starting at 3.4 and ending at 7.8, we can calculate that about 20,000 bytes for each second. See more information in Network capture tutorial (*Traffic volume in certain interval*, pp. 14).

3. Please provide the top 5 most commonly used protocols, as well as the five most common source ports and five most common destination ports based on flows. Detail in a table for each one

   - the number of flows
   - the number of packets
   - the amount of data (bytes)
   - the application or usage

   Hint: The column 'pro' defines the protocol used.

4. Which are the top-ten host pairs based on

   - number of flows
   - number of bytes Are there the same pairs?

5. Plot the number of flows for the 100 most common pairs of hosts

   - Using linear scale
   - Using logarithmic scale

6. Repeat the previous plot (both linear and logarithmic scale) using this time fixed size (2^{16} slots) array approach (Network capture tutorial - *Large data analysis*, pp. 8 and *solution #2*, pp. 10). What can you say about the results?

7. Is there a more efficient approach in terms of running time and memory consumption to accomplish this task?

   Note: You can use `/bin/time` command to get resource consumption of a command, use `-v` for more verbose. It provided a more detailed output than shell built-in `time`.

Report, task 2

- Describe your analysis setup. Include code snippets.
- Provide answers for 7 questions above.

- Discussion on memory resources requirements.

Task 3: Analyse packet capture (user traffic)

Based on the traffic captured in Task 2, **utilize an appropriate tool to analyze the captured data** and provide answers to the following questions:

1. How many IPv4 hosts (and IPv6, if any) are communicating?
2. Top 5 host countries (e.g. GeoIP)
3. Top 15 hosts by byte counts.
4. Top 15 hosts by packet counts. Were there any differences between the top 15 hosts in terms of byte counts and packet counts?
5. Top 10 TCP and top 5 UDP port numbers (by packet count).
6. Top 10 fastest TCP connections
7. Bit and packet rate over time (e.g. *tcpstat*, *capinfos*)
8. How many hosts were tried to contact to, but communication failed for a reason or another? Can you identify different subclasses of failed communications?

   Note: Please **choose one of the mass analysis tools** to use such as shown in the Table 1. Mass analysis tools or another suitable tool (some packet-capturing software can also analyze for such a small amount of data, but it is better to practice the mass analyzer tool)

Report, task 3

- Describe the tool you chose and how you used it to complete the analysis.
- Answers to questions above.
- Based on the analysis above, do you have any other interesting observations to share?

Grading standard

To pass this course, you need to achieve at least 15 points in this assignment. And if you submit the assignment late, you can get a maximum of 15 points.

You can get up to 30 points for this assignment:

Task 1

- Explain the concepts requested related to traffic data. (4p)

Task 2

- Use the correct method to convert a pcap file into flows. (1p)
- Accurately answer the 7 questions raised in the task. (12p)

Task 3

- Accurately answer the 8 questions raised in the task. (11p)
- Summarize based on the answers to the questions you answered. (2p)

The quality of the report (bonus 2p)

The instruction of assignment

For the assignment, your submission must contain (Please don't contain original data in your submission):

- A zip file that includes your codes and scripts.
- A PDF file as your report.

Regarding the report, your report must have:

- A cover page indicating your name, student ID and your e-mail address.
- The report should include a description of measurements, a summary of the results and conclusions based on the results.
- An explanation of each problem, explain how you solved it and why you did it.

Annex

- **How to calculate traffic volume with bits per second?**

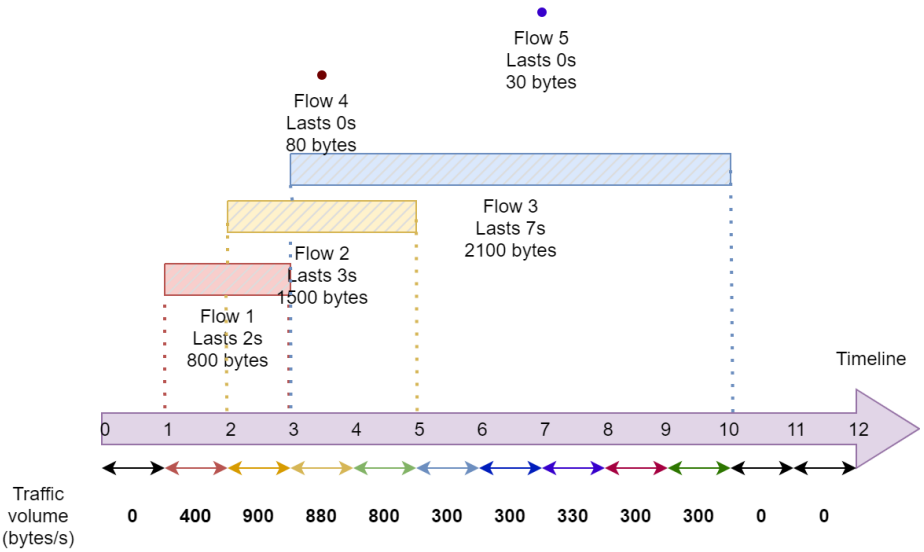

Figure 1: Calculation example of traffic volume

See more information in Network capture tutorial (*Traffic volume in certain interval*, pp. 14).