# Exercise 2 - Flows for S-38.3183 - Spring 2007

Mika Ilvesmäki
Helsinki University of Technology
Networking Laboratory
P.O. Box 3000
02015 TKK
Finland
*mika.ilvesmaki@netlab.tkk.fi*
*s383183-exercise@netlab.tkk.fi*

**Abstract**

Deadline for this exercise is 2.4.2007. To get your report into the grading process it must be submitted via email to *s383183-exercise@netlab.tkk.fi*. No returns in paper format or otherwise are allowed. The reports must be in pdf-format and sent as attachments to the email message. Do not forget your name and study book number from your report.

## I. Introduction

This second exercise introduces you to flow trace files and how to get information out of them. You are already familiar working with trace files and therefore the main learning goal of this exercise is getting better at basic level flow analysis.

## II. Problems

All your answers should also contain the scripts/command line with which you created your answers. Lengthy entries of code should be put to the appendix of your exercise report. Remember always to comment (discuss) the results. The grading is largely based on your discussion of the results.

### *A. Traces*

The traces are available from
http://www.netlab.tkk.fi/opetus/s383183/k07/exercises/traces/
For this exercise, choose one of the funet$x$-files, where $x$ is

$$x = 1 + \text{Your study book number} \mod 8 \tag{1}$$

and one of the dec$y$-files where $y$ is

$$y = 1 + \text{Your study book number} \mod 4 \tag{2}$$

Remember to read the readme.txt file. Note also, that the original files are big (and authentic) so you may need additional space in your work directory (use scratch- and tmp -directories). Catenate from the compressed files when possible.

## III. Exercise questions

### A. Flow data from packet data with flowID

As you may remember your dec$x$-trace contains information on packets. There is also predetermined flow identifier calculated for each packet. This flow identifier identifies the flow to which the packet belongs.

Your first task is is produce flow-data for every flow found in your dec$x$-trace. For every flow you need to record the srcIP, dstIP, protocol, and both the source port and destination port numbers. You will also have to determine the start and finish time of the flow, amount of packets seen in the flow and the amount of data seen in the flow. All this data is to be gathered into one row per flow for all flows.

Store your dec$x$- flow file. You will need it in the following tasks.

*1) Extra work:* If you want, you can shortly discuss how the flowID could be determined if none was given in the packet trace files. You may also suggest an implementation level solution.

### B. Basic flow analysis - spatial aggregation

For both (dec and funet) of your flow traces determine the following basic properties:

- Length (in seconds) of the trace, determine also the start time and the end time.
- Number of packets in the trace?
- Amount of data (in bytes) seen in the trace?
- How many flows are there in the trace?
- How many single packet flows (calculate also the percentage from all flows)?
- Number and shares of TCP and UDP packets?
- Number and shares of TCP and UDP flows?
- Amount and shares of TCP and UDP data?
- Number and shares of TCP flows that contain only 1 (one) packet? How much of the total data is in 1-pkt flows?
- What share of the bytes of all bytes / TCP bytes originate from TCP source port 80? What is this source port used for?
- What is the average length of a multi-packet flow (in seconds and bytes)?
- What is the average length of a multi-packet flow (in seconds and bytes) originating from TCP source port 80?
- What share of the bytes of all bytes / TCP bytes are destined to TCP destination port 80? What is this destination port used for?

- How does the amount of data sent *to and from* TCP port 80 divide?
- What are the 10 longest flow in length (time)? Describe five of these flows if possible.

Present your results in a table format. Use several tables if needed. Remember to dicuss your results.

### C. Spatio-Temporal aggregation

The intention of this part of the exercise is to further enhance the concept of spatio-temporal aggregation. The idea is to aggregate flow arrivals by observing these arrivals per certain time intervals (1 second, 10 second and 60 second intervals).

For both of your traces calculate the timeseries for flow arrivals for 1) all flows, 2) flows that originate from TCP source port 6881 and 3) flows that originate from TCP source port 80:

- What are the TCP ports 6881 and 80 used for?
- Determine the number of flow arrivals per 1, 10 and 60 seconds. *Tip: you should be able to build upon the code you produced in Exercise 1. Do include the modified code in the appendix of your exercise report.*
  Visualize your timeseries for flow arrivals in a format that shows the arrived number of flows per time interval. Discuss your results.
- Plot density histograms of the flow arrival data (arrivals per 1 second) with mean and variance/standard deviation of the data showing. Compare the data from different traces and different TCP source ports and discuss your results. Compare the histograms to two known distributions (normal, exponential etc.). Use p-p -plots, for instance. Discuss your results.

Store both your timeseries data and the scripts you used to produce the data for the duration of this course.

### D. Return of this exercises

Deadline of this exercise is April 2nd, 2007. To get your report into the grading process it must be submitted via email to *s383183-exercise@netlab.tkk.fi*. No returns in paper format or otherwise are allowed. The reports must be in pdf-format and sent as attachments to the email message. Do not forget your name and study book number from your report.

## IV. Acknowledgements

The author would like to thank CSC - the Center of Scientific Computing in Finland - for providing access to Funet network and for computing and archive resources and Lic. Sc. Markus Peuhkuri for his kind help in preprocessing the traces.