Aalto University

# Final Assignment

From measurements to conclusions

Markus Peuhkuri

2022-10-24

# Requirements

# Introduction

- Almost all the concepts covered in this course ranging from data measurements to driving the results and conclusion from the data sets.
- Three main tasks both with several subtasks and final conclusions:
  - Task 1: Capturing packets
  - Task 2: Flow data
  - Task 3: Analysing active measurements
  - Final conclusions
- Will take non-trivial time to complete, plan accordingly

# Completing

- Review sessions approximately week before deadline
  - Should have mostly ready document by then.
  - The sessions will follow format of weekly assignments i.e. discussion in groups and joint review and discussion about matter.
- The report will be returned via MyCourses by 5th December
  - A late submission will only get **grade 1 maximum**.

# The result: Report

1. Main document explaining results and findings without technical details. This is like information that would be given to the customer hired you to make analysis.
2. Appendix contains detailed explanations on what have been done supplemented by commands used to get a result or draw a figure, if appropriate. Plain commands, scripts, or codes without comments are not sufficient. This is like information you would hand out to your colleague who needs to do similar analysis for another customer.

   Also include samples of data sources, like 5-10 first relevant lines when appropriate. Do **not** include full datasets.

# Things to note

- Figures are as informative as possible
  - Select appropriate plot, units and scales (linear/log and range).
  - Note colours and symbols in plot if there are multiple variables or data sets.
  - Labels and captions informative but not redundant.
- Following processes for each data set:
  - Pre-Processing
  - Analysis
  - Conclusions
- Do not forget any question or section!

# Assessment

- Individual work.
- Cooperation is encouraged, but **all output should be produced by yourself**.
- The assignment grade is composed of:
  - Correct and insightful answers (weight 80%)
  - Readability, clarity and style of the report (weight 20%)

# Support

The assignment is meant to be individual work, but there are two kinds of support available for the students:

- Interactive exercise classes
- Review sessions late in November (schedule will be published a week in advance)
- Course Zulip *finalassignment* stream for questions to course staff and also peer support.

# Tools

- Set up the environment:
  1. `source /work/courses/unix/T/ELEC/E7130/general/use.sh`
  2. `source /work/courses/unix/T/ELEC/E7130/general/use.csh`
- Coral Reef
  - `crl_to_pcap`: for converting and anonymizing packet traces
  - `crl_flow`: for summarising packet data to flows
- Tstat
- GeoIP library: python2/3, perl modules and command line
- Tip: after preprosessing and analysis, script graph generation to produce figures with uniform look.

# Tasks

# Task 1: Capturing packets

- Acquiring packet capture data
  - use `dumpcap`, `tcpdump` or `wireshark` to collect *at least* two hours of *normal* computer / network usage
  - day-long trace is much better
  - remember to document how and where data was acquired (metadata!)
- PS1: packet data including desired columns
- PS2: convert packet data to flow data
- PS3: TCP connection statistics

# Data analysis: part 1

- Packet data PS1
  - 1.1 - 1.3: by port numbers, traffic volume, packet length
- Flow data DS2
  - 1.4 - 1.10: by port, country, OD-pairs, flow length, timeouts
- TCP connection data DS3
  - 1.11 -1.12: relation of retransmissions to RTT, traffic volume
- Conclusions

# Task 2: Flow data

- Data found `/work/courses/unix/T/ELEC/E7130/general/trace`
  - also in `$TRACE` variable if sourced `use.sh` script
- Directories contain following data:
  - `flow-continue`: output generated with crl_flow tool using 60 second timeout to expire flow. Time intervals are aligned as one hour.
  - `flow-expire`: same as above, but all flows are expired when reporting period (one hour) ends.
  - `tstat-log`: output generated with tstat tool.
- IP addresses are randomized: reverse name lookups, *whois* databases or *geoip* databases will leads to random results.
- Student number last digit tells what subnet to analyze

# Data analysis: part 2

- 2.1: repeat one of task 1 analysis (1.4-1.5, 1.7-1.9) on FS2
- 2.2: Per user data volume
- 2.3: Flow sampling
- 2.4: Conclusions

# Task 3: Analysing active measurements

- Data collection started by *Basic Measurements* assignment.
- If you have two weeks of data, that is enough (you can stop measurements).
- Data sets **AS1.x**, **x=**
  - name servers with DNS (**d1**, **d2**, **d3** ) and ICMP (**n1**, **n2**, **n3**)
  - research servers (**r1**, **r2**, **r3**)
  - iperf servers (**i1**, **i2**)

# Data analysis, part 3

- Latency data: distributions, statistics (3.1) and time series (3.2)
- Throughput: distributions, statistics (3.3) and time series (3.4)
- Conclusions

# Final conclusions

- How was your own traffic (Task 1) different from the data provided (Task 2)? What kind of differences you can identify? What could be a reason for that?
- Comparing RTT latency about TCP connections (3.10), were active latency measurements around the same magnitude or was another much larger than the other?
- Discuss how data protection needs to be taken into account if you as a network provider employee were doing similar measurements as if it is perfomed by network provider
- How you rate complexity of different tasks? Was some tasks more difficult or laborious than others? Did data volume cause any issues with your analysis?

# Ask Questions in Zulip and in sessions