Traffic Identification in High-Speed Computer Networks based on Machine Learning and Hybrid System Architecture

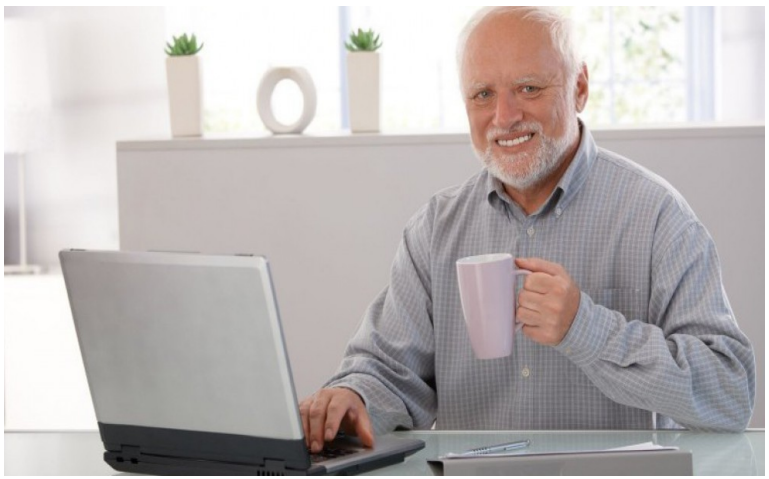
Student: Nachyn A.O.
Advisors: Pavlovskiy E.N., Khazankin G.R.

Novosibirsk State University, Department of Mechanics and Mathematics

December 6, 2017

Introduction

Traffic identification[1] - automated process which categorises computer network traffic according to various parameters into a number of traffic classes.



[1] https://en.wikipedia.org/wiki/Traffic_classification

General methods

- **Port-based** (earliest): HTTP - 80, SSL - 443. However, there are a lot of new protocols that don't follow this rule).
- **Signature-based** (since 2002): Efficient, but time-consuming. When protocol specification changes or a new protocol produces, must start again for finding valuable signatures.
- **Statistical Features** and **Machine Learning** (recently): Features of traffic transmission such as the packet size, time, IP. Models like Naive Bayes, Random Forest, Neural Network. Time-consuming training, but real-time or near inference.

General stages

- Data collection
- Preprocessing
- Analysis
- Inference

Existing works

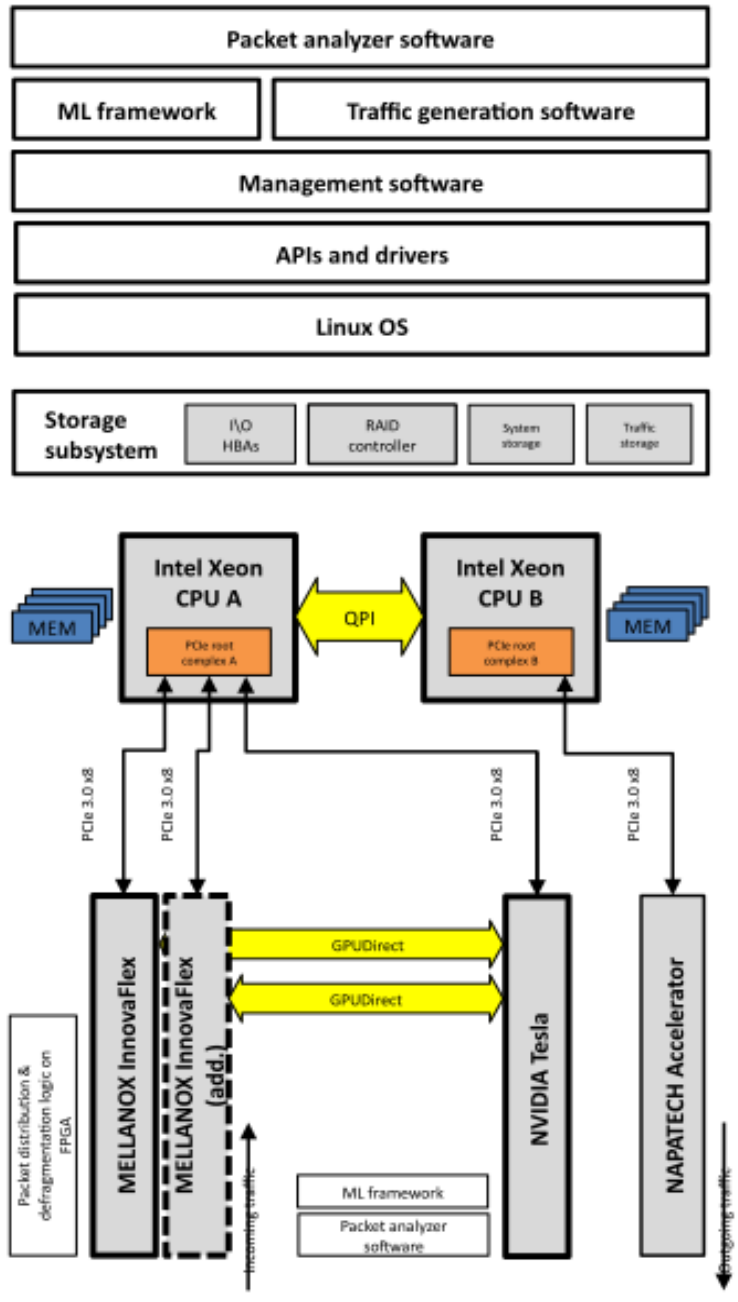
- The Random Forest based Detection of Shadowsock's Traffic. 2017. Ziye Deng, Zihan Liu, Zhouguo Chen, Yubin Guo.
- Traffic Analysis with Deep Learning. 2017. Se Eun Oh, Saikrishna Sunkam, Nicholas Hopper.
- Traffic Flow Prediction With Big Data: A Deep Learning Approach. 2014. Yisheng Lv, Yanjie Duan, Wenwen Kang.
- Skype multimedia application traffic analysis on home Unifi network. 2017. Murizah Kassim, Siti Fatimah Ramle, Ruhani Ab. Rahman.
- And much more at IEEE Xplore.

Architecture

- Bandwidths $\geq$ 80 Gbps
- Hybrid system architecture (CPU + GPU)

Advantages:

- GPUDirect - allow to bypass CPU memory
- Well scalable



Extracted features

- MAC
- OS Name
- Screen resolution
- Datetime
- Session ID
- IP
- Destination IP
- Language encoding for the browser
- And much more

Important:
The captured data should be sampled on time
Need for a suitable preprocessing

- Collect time-series data for single user
- Train the model
- Evaluate (try to identify deviant behavior)

Analyzing

Encrypted part

Statistical methods

Open part

Time series analysis with deep learning: RNN, CNN

Thanks for attention!