The Future of Network Flow Monitoring

Prague Embedded Systems Workshop (PESW 2019)

Friday 28th June, 2019

Petr Velan

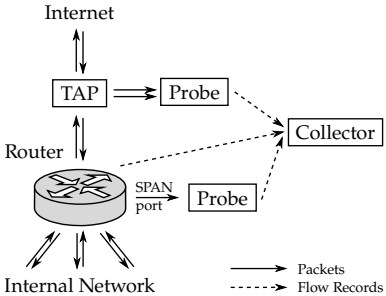
MUNI ICS

CSIRT-MU

Flow Monitoring Introduction

Flow monitoring is widely used for:

- Accounting
- Security (IDS, forensics)
- Data retention
- Network diagnostics



A flow record example:

```
Flow start   Duration Proto     Src IP Addr:Port          Dst IP Addr:Port   Flags   Packets  Bytes
09:41:21.763   0.101  TCP     172.16.96.48:15094 ->  209.85.135.147:80      .AP.SF        4    715
09:41:21.893   0.031  TCP   209.85.135.147:80    ->    172.16.96.48:15094  .AP.SF        4   1594
```


CSIRT-MU

The Past

History of Flow Monitoring

IETF Internet Accounting Working Group

First mention of a flow export in RFC 1272 published in **1991**

- The goal was to provide background information on internet accounting
- How to deploy it, how to collect and process data

The common belief was that monitoring is intrusive

- It was generally frown upon

Lack of interest led to conclusion of the WG in **1993**

- Negative attitude towards monitoring persists even now (RFC 7258)

History of Flow Monitoring

IETF Realtime Traffic Flow Measurement WG

A method for Internet traffic flow profiling based on packet aggregation

- Presented by Claffy et al. in **1995**

Renewed interest in flow monitoring

- Establishment of the IETF RTFM WG (1996 - 2000)

Published RFCs covering flow measurement framework

- Even bidirectional flow export

Again, lack of interest of vendors → no standards emerged

- The WG was concluded, its goals finished.

History of Flow Monitoring

Cisco and NetFlow

Information about packet flows stored in routers and switches

- Similar flow information as proposed by RTFM

The main goals is packet switching/routing, not monitoring

- The configuration and features of the monitoring process are limited

NetFlow was patented in **1996**

- General public was using NetFlow v5 available from circa **2002**

Official specification for NetFlow v5 was never released

- Inconsistencies occurred as some elements were reused for different purposes

NetFlow v9 superseded v5 and is used even now

History of Flow Monitoring

Other Vendors

Everybody had to have their own protocol

- Very similar, NetFlow remained the most well known and used

Juniper: **JFlow**

- Current version is JFlow v9

Alcatel-Lucent (now Nokia): **CFlow**

- Versions v9 and v10 interoperable with NetFlow v9 and IPFIX respectively

Ericsson: **RFlow**

- Uses NetFlow v5 format

Many others: Huawei (**NetStream**), Citrix (**AppFlow**), ...

History of Flow Monitoring

Lo and Behold: IPFIX

2001: flow monitoring is clearly an item now

- No standard protocol exists (NetFlow v5 not even public yet)

IETF IP Flow Information eXport WG

- Lot of goals on the charter, but the primary was to create flow export protocol

WG specified requirements and let vendors submit their proposals

- Cisco NetFlow v9 codified in RFC 3955 to compete

NetFlow v9 was the most advanced protocol → selected as a base for IPFIX

- IPFIX sometimes called NetFlow v10

IPFIX WG concluded in **2014**

- Published 29 RFCs, some work continues beyond the WG

History of Flow Monitoring

Security and Flow Monitoring

Cisco proposed to used flows for anomaly detection and traffic analysis in **2005**

- Used mostly for accounting and network management until then

The quality of flows had to improve

- Sampling has negative impact $\rightarrow$ dedicated probes

L7 information in flows (Flexible NetFlow Technology by Cisco in 2006)

- Analysis of HTTP, TLS, SSH, DNS, SMTP, ...

Cisco Joy (2016)

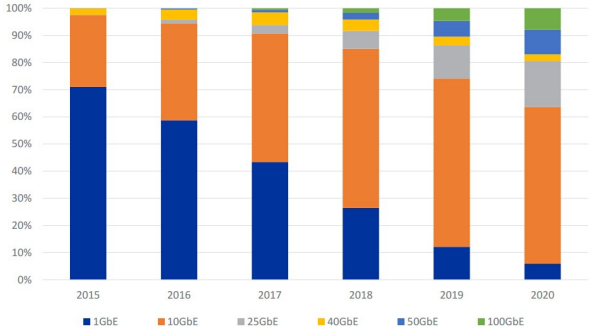
- Application data, statistical data beyond simple counters

The Present

Current State of Flow Monitoring

Growing Speed of Networks

10 G, 25 G, 40 G and 100 G: Seeing Broad Adoption in Data Center



http://techblog.comsoc.org/tag/25-100g-ethernet/

CSIRT-MU

Current State of Flow Monitoring

Growing Speed of Networks

100G+ network probes (L2-L4) using HW accelerated NICs

- Always with custom kernel drivers and userspace libraries
- FPGA vs ASIC
- Basic acceleration (RSS, timestamps) provided by commodity cards

L7 monitoring up to 10G

- Processing payloads requires a lot more performance

Parallelisation of the monitoring

- Utilisation of multicore CPUs

Current State of Flow Monitoring

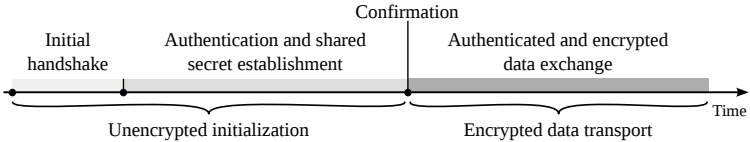
Growing Amount of Encrypted Traffic

Wide adoption of encryption:

- TLS used to encapsulate everything (HTTP/2 implementations require TLS)
- WireGuard VPN

Some information remains disclosed even for encrypted traffic:

- Initialisation of the encrypted connection is usually unencrypted
- TLS up to version 1.3 discloses certificates
- SNI still available, but propositions are being made to encrypt it



CSIRT-MU

Current State of Flow Monitoring

Encrypted Traffic Classification

Identification of encrypted protocols is not often possible.

- Unencrypted payload does not provide enough information
- Machine learning and statistical methods can be used

Current problems of machine learning on network flow data:

- Not enough features to work with
- Labelled data are needed for semi-supervised and supervised ML
- Training on a static data sets

The Future

The Future of Flow Monitoring

Monitoring beyond 100G

Distributed architecture

- Divide the traffic to multiple devices and process separately

Limited set of features

- Collecting only basic statistics
- L7 is encrypted anyway

Further hardware acceleration

- New chips come with more memory and performance
- Implement most of the monitoring process in dedicated HW

The Future of Flow Monitoring

Machine Learning I.

Need for features

- Per-packet information is needed (size, timestamp)
- Flows must be extended to collect these features
- Accuracy vs amount of data

Performance of ML

- Training is costly
- HW acceleration chips for ML (Huawei)
- Compute directly on packets?

Accuracy of ML

- Network traffic properties are changing
- Accuracy decreases over time $\rightarrow$ need to periodically retrain models
- Training on a stream of data

The Future of Flow Monitoring

Machine Learning II.

Ground truth

- Labelled data are needed for semi-supervised and supervised ML
- Manually created data sets vs manually annotated real network traffic
- Continuous retraining needs continuous ground truth

Getting the ground truth

- Data from other sources (server logs, DNS logs, IDS, . . .)
- Combine the data with flows $\rightarrow$ labelled dataset
- Allows to continuously retrain

The Future of Flow Monitoring

Quality of Flow Data

Flow data is used for anomaly and attack detection

- Does quality of flow data matter?
- Impact of data loss, imprecise timestamps
- Especially for machine learning
- Balance quality and performance

THANK YOU FOR YOUR ATTENTION!

https://csirt.muni.cz/

@csirtmu

Petr Velan

velan@ics.muni.cz

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION, YOUTH AND SPORTS

MUNI ICS

CSIRT-MU