

A Case for Packet Sampling



Fraunhofer
Institute for Open
Communication Systems

Tanja Zseby, zseby@fokus.fhg.de

Competence Center for Autonomic Networking Technologies

Motivation: FloCon 2005

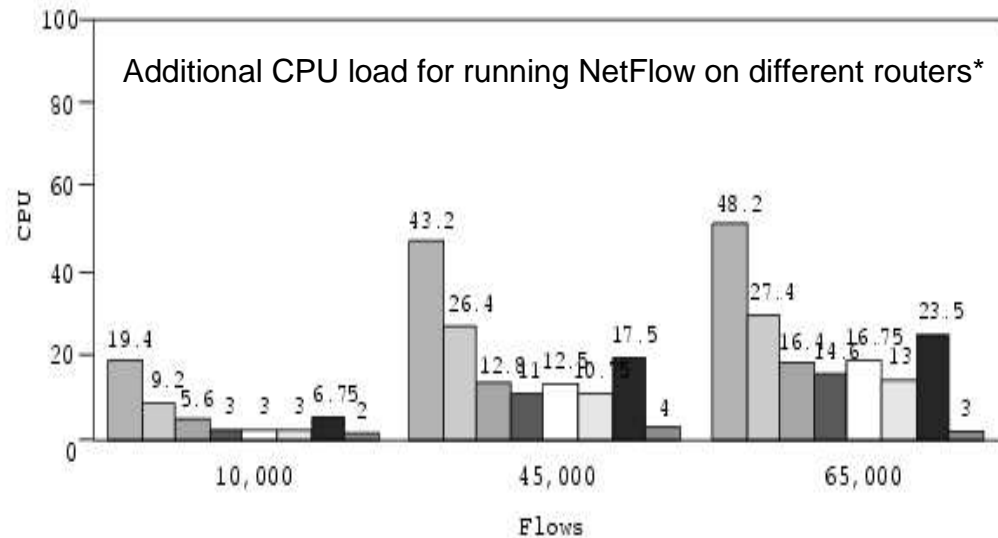
FloCon05 participants:

“We don’t believe in Sampling”

- Happy to use flow data
- Very skeptical to packet sampling

The Problem: Limited Resources

- Full packet capture at each node not feasible
 - Increasing data rates
 - Hardware costs
 - Privacy concerns
- Resources are limited
 - Storage
 - Processing
 - Transmission

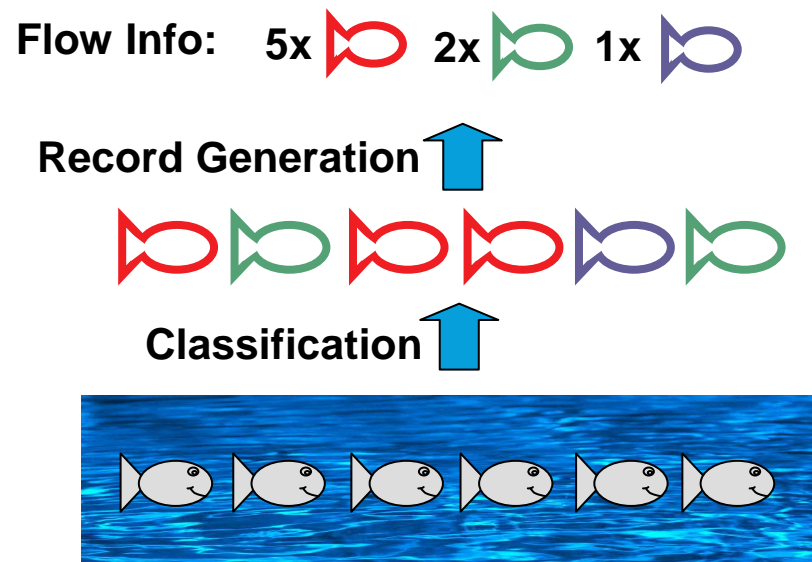


We cannot measure everything

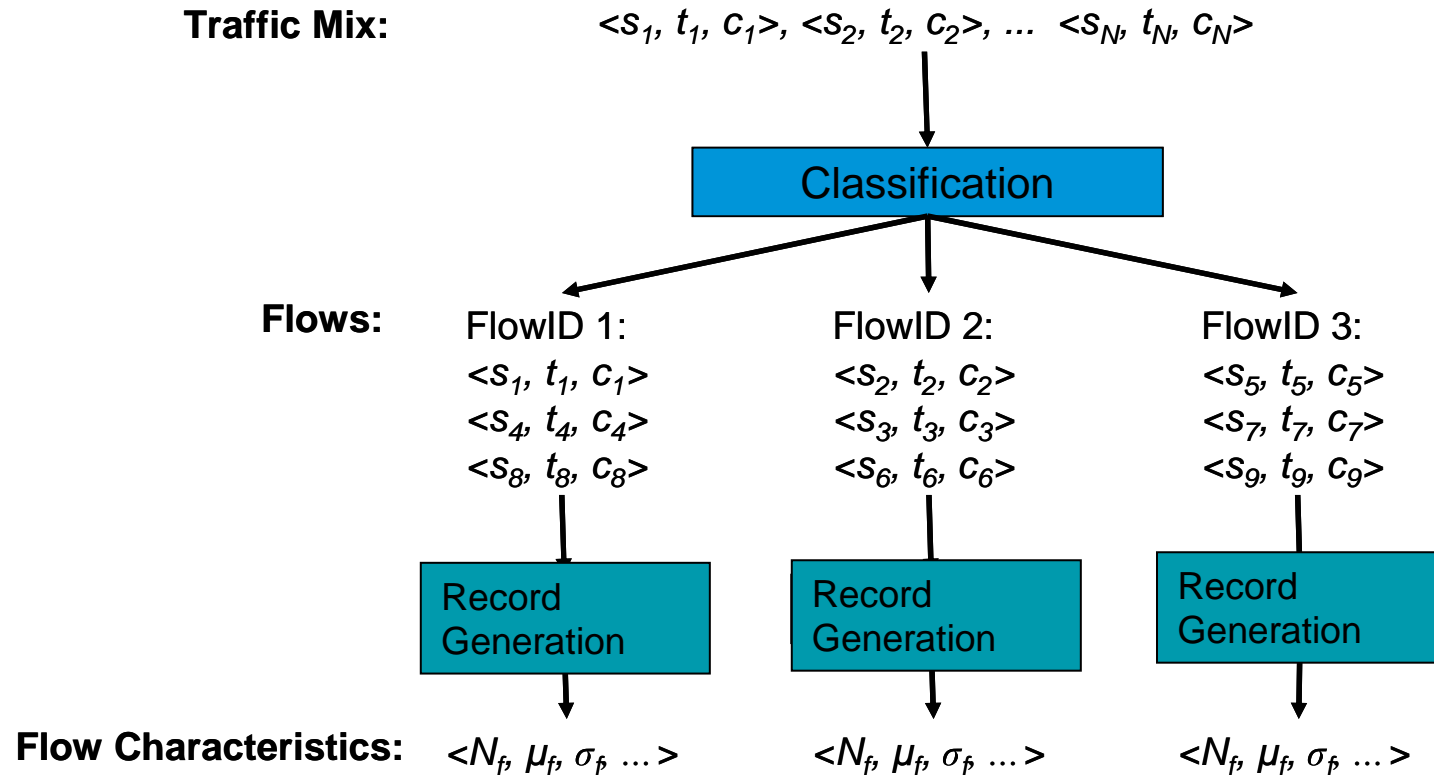
*source: NetFlow Performance Analysis, Cisco white paper
http://www.cisco.com/warp/public/cc/pd/iosw/prodlit/ntfo_wpa.jpg

Solution1: Flow Data

- Grouping of packets into flows (classification)
- Reporting of flow information only
- Disadvantages:
 - Per-packet information is lost
 - Information and effort depends on flow definition



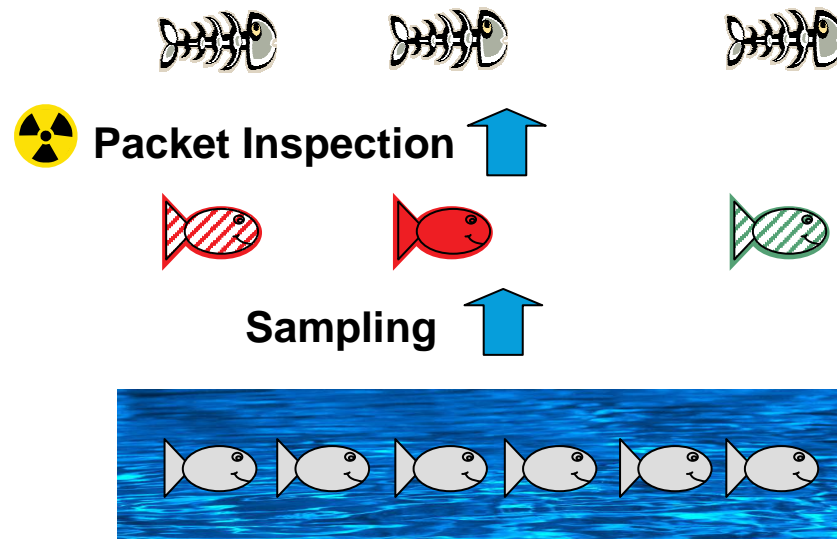
Flow Data Generation



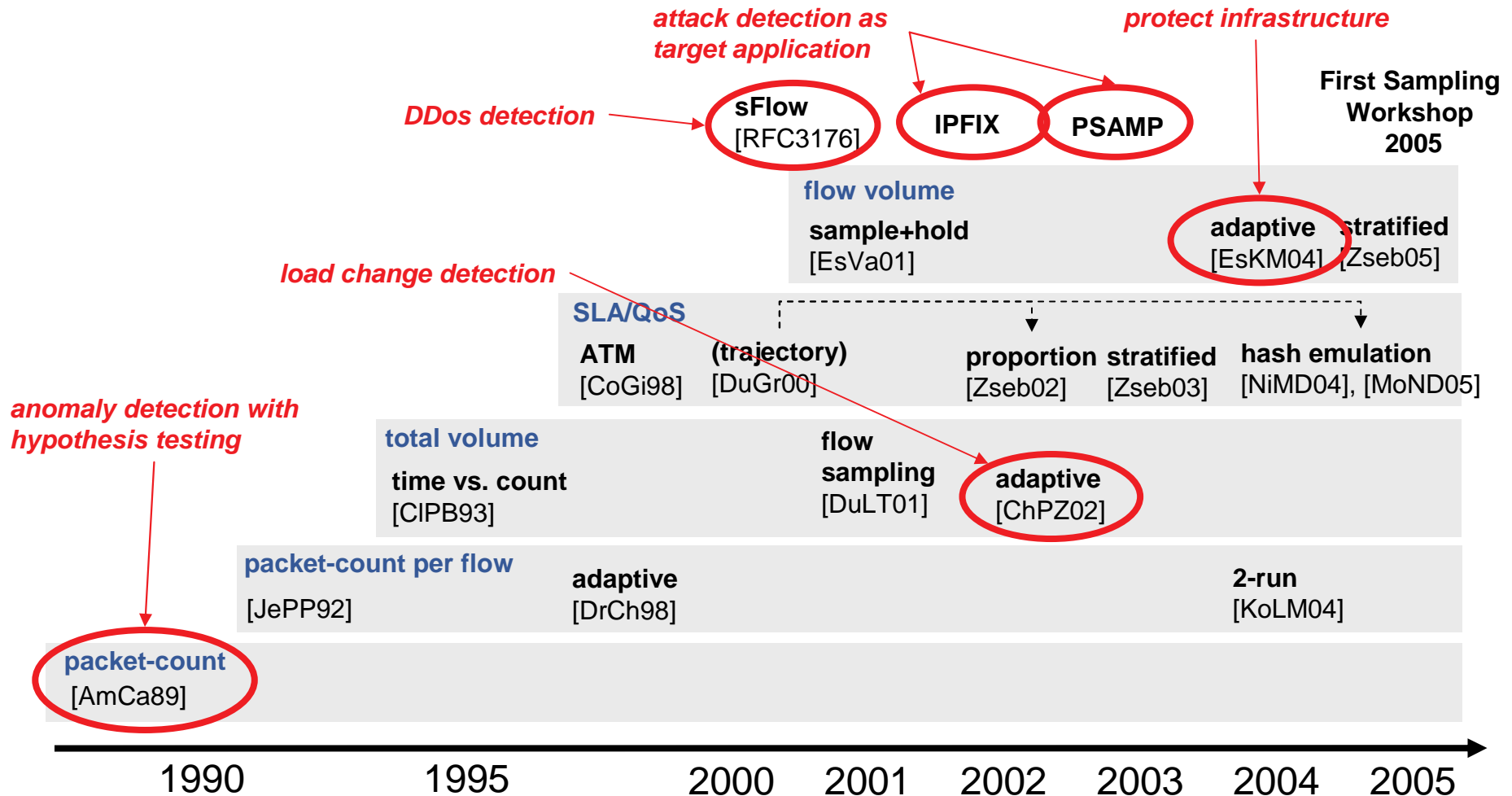
- Information about packets is discarded
- Available information depends on
 - Flow definition
 - Flow characteristics that are reported

Solution2: Packet Sampling

- Random Selection of some packets
 - Report parts or full packet information
 - Estimation of metrics based on sample
- Provides different viewpoint
 - Packet data can reveal further information
 - Sampled data sufficient for some metrics
- Helps to protect measurement infrastructure during attack



Sampling: State of Art



Packet Sampling

Real metric substituted by estimate

→ Accuracy statement is essential

Accuracy depends on

- Sampling scheme
- Estimation method
- Position of sampling process in measurement sequence
- Population characteristics (e.g. variance of metric of interest)

A Simple Example

Goal: Estimation of packet proportions (e.g. TCP-SYN packets in a flow)

Real proportion: $P = \frac{M}{N}$ Estimate: $\hat{P} = \frac{m}{n}$

Estimation Accuracy (random n-of-N): $\sigma_{\hat{P}} = \sqrt{\frac{P \cdot (1-P)}{n}} \cdot \sqrt{\frac{N-n}{N-1}}$

Confidence Limits: $Prob\left(\hat{P} - z_c \cdot \sigma_{\hat{P}} \leq P \leq \hat{P} + z_c \cdot \sigma_{\hat{P}}\right) = 1 - \alpha$

Example: - Measurement interval with N=10,000 packets
- Random packet selection 1% (n=100)

$\hat{P} = 0.9 \rightarrow \sigma_{\hat{P}} = 0.03 \rightarrow 0.8226 \leq P \leq 0.977$, with 99% confidence

$\hat{P} = 0.1 \rightarrow$ same accuracy

$\hat{P} = 0.5$ (worst case) $\rightarrow \sigma_{\hat{P}} = 0.05 \rightarrow 0.371 \leq P \leq 0.629$, with 99% confidence

Works with other packet properties, too!

Advise

- Don't restrict your analysis to flow data
 - Include further viewpoints
 - Use sampling in addition or as alternative to flow data
- Trust the power of statistics
 - It's a mature and well established field
 - full range of proven techniques
- Use sampling where applicable
 - Applicability depends on traffic profile, metric of interest, accuracy demand
 - Sampled data sufficient to detect large events (high volumes, high packet counts)
 - May be sufficient to estimate #pkts with specific properties (e.g. SYN, VoIP packets, small packets, packets with same content, etc.)
 - Others → depends on scenario
 - Difficulties with rare events (stealth attacks, slow port scans)
 - Not suitable to re-assemble connections (but filtering may be)

***Thank you for your
attention!***



Fraunhofer
Institute for Open
Communication Systems