The CINBAD update

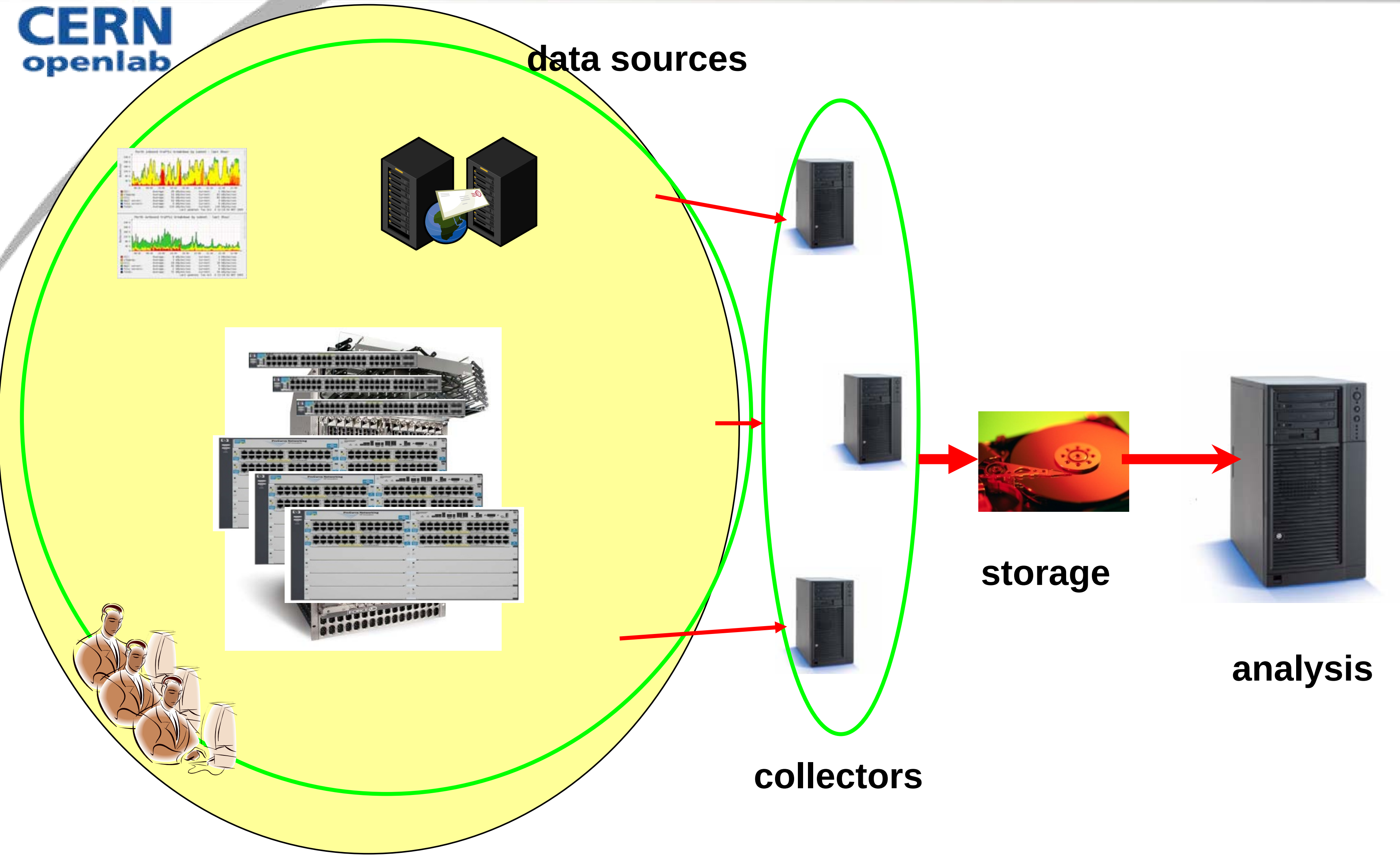
29th January 2008

Ryszard Erazm Jurga
Milosz Marian Hulboj

Packet Sampling Studies

- Packet Sampling Studies
  - Motivation
  - Results and Conclusions
  - Improvements and Further Directions
- Conclusions

Where are we?



data sources

collectors

storage

analysis

Why Packet Sampling Studies?

- Gain some insight about packet sampling in context of the CINBAD project
(both sFlow & Netflow use sampling)

- Review of various sampling methods and their applications

- Estimate network parameters from the sampled data

- Is sampling directly suitable for all the tasks?
    - traffic monitoring for billing, accounting, SLA, etc.
    - network anomaly detections, viruses, worms, etc.

Results of Packet Sampling Studies

- Over 100 technical papers read and analysed
- Thorough Technical Report available on the CINBAD website

- Collaboration between CERN and HP in the 90s
  - Peter Phaal (InMon – sFlow inventor, was in HP)
  - Bjørn Blindheim (technical student at CERN in the 90s)
- Technical Report “A security auditor based on TCP transaction records”
- Many thanks to Ben Segal for sharing with us his memories about the project

- Had been successfully used in many applications:
  - Typical usage: accounting and billing
  - Deployed at Amsterdam Internet Exchange Point
  - Many hardware manufacturers support sampling (i.e. ProCurve in 3400, 3500, 5400 series)

Less Explored Packet Sampling Areas (I)

- Network Anomaly Detection
  - Very few publications on sample based analysis
  - Most of the approaches require full data or special hardware support (deep packet inspection)
- Data aggregation
  - For large high speed networks even sampling can generate terabytes of data per day
  - Raw data is almost useless, we need to build some aggregates
  - Simple statistics per interface/device/network are usually not sufficient
  - There is no agreement on what data should be stored
  - We need more dynamic representations – i.e. data flows
  - Building data flows could be a challenge if we have only partial data

Less Explored Packet Sampling Areas (II)

- Adaptive sampling

  Facts:

  - Accuracy of estimates depends on the number of samples
  - Fixed level of error (invariant of the conditions) is desired
  - Each network device has certain sampling limits

  However:

  - Network state and traffic are dynamic
  - Most of the publications deal with fixed rate sampling
  - Anomaly detection is much different from typical sampling applications
  - We need the best possible accuracy
  - Dynamic adaptation of sampling rate would be a good solution

Need for Packet Sampling Improvements

- Improve the accuracy of estimates to get more accurate anomaly detection
  - Analysis of sampling sources at CERN
    - sFlow tested on CERN network devices (ProCurve and others)
    - synthetic tests with traffic generators
    - tests on production network
    - some potential issues are being discussed
  - Simulating sampling methods with real network data:
    - Systematic sampling, sFlow
    - Adaptive sampling and prediction techniques
  - Estimate traffic parameters and compare sampling methods
    - Mean
    - Sampling variance
    - Hurst parameter
    - Φ coefficient
    - Mean square error
    - …

Impact of Packet Sampling on Anomaly Detection

- Examining the packet sampling influence on Intrusion Detection Systems (Snort)
  - some anomalies only require a single sample in order to provide 100% accuracy of detection
- Traces with known attacks will be used for analysis
- Snort will be fed with full trace and sampled variants
  - examine what is sampling impact on detection ratio
  - determine what sampling approach is the best
- Does good sampling from the 'conventional' point of view give good result in anomaly detection?
  - If not, then look for improvements

Need for data aggregation

- Aggregates should:
  - provide information about dynamic data flows,
  - be unaffected by partial data,
    (i.e. can not depend on TCP connection state)
  - be efficient for storage,
  - be scalable and easy to combine into bigger sets,
  - **be useful for further analysis**

- Thus we want to build various aggregates:
  - From full traces (captured on CERN's network)
  - From sampled traces (with different sampling parameters)

- Evaluate the accuracy of aggregates using different metrics

New Data Sources (I)

- Packet sampling data is not enough!
  - Data is partial
  - It cannot provide 100% accuracy

- More data to understand flow of data in the network
  - External sources provide useful information and time triggers
  - Which problem is accompanied with given traffic pattern?
  - Correlation between various data sources

- More data sources = More information

  = Better accuracy and less false alarms

New Data Sources (II)

- Central Antivirus Service at CERN
  - On-line information from antivirus programs installed on Windows machines
  - For each host information about antivirus actions is logged:
    - Virus name
    - Virus type
    - Action taken
    - Date of action
    - etc
- CERN network monitoring tool
  - Provides plenty of events and alerts about the network devices and the network state itself
  - ATLAS Experiment is using some data from this tool
  - We need to understand what we can actually get out of all this data

Conclusions

- Packet sampling studies were great tour of the sampling landscape and the strengths, weaknesses and opportunities for further research

- Our work will build on established research and will not duplicate topics already investigated

- First results from analysis of packet sampling methods for anomaly detection are expected in the following weeks